ON THE CURSE OF MEMORY IN RECURRENT NEURAL NETWORKS: APPROXIMATION AND OPTIMIZATION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study 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. Mathematically, the latter can be understood as a sequence of linear functionals. We prove a universal approximation theorem of such linear functionals, and characterize the approximation rate. Moreover, we perform a fine-grained dynamical analysis of training linear RNNs by gradient descent. A unifying theme uncovered is the non-trivial effect of memory, a notion that can be made precise in our framework, on approximation and optimization: when there is long term memory in the target, it takes a large number of neurons to approximate it. Moreover, the training process will suffer from slow downs. In particular, both of these effects become exponentially more pronounced with memory - a phenomenon we call the "curse of memory". These analyses represent a basic step towards a concrete mathematical understanding of new phenomenon that may arise in learning temporal relationships using recurrent architectures.

1 INTRODUCTION

Recurrent neural networks (RNN) (Rumelhart et al., 1986) are among the most frequently employed methods to build machine learning models on temporal data. Despite its ubiquitous application (Baldi et al., 1999; Graves & Schmidhuber, 2009; Graves, 2013; Graves et al., 2013; Graves & Jaitly, 2014; Gregor et al., 2015), some fundamental theoretical questions remain to be answered. These come in several flavors. First, one may pose the *approximation* problem, which asks what kind of temporal 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 observed 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 the static feed-forward setting, the key distinguishing feature here is the presence of temporal dynamics in terms of both the recurrent architectures in the model and the dynamical structures in 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 linear setting. For this reason, we will focus our analysis on linear RNNs, i.e. those with linear activations. Further, we will employ a continuous-time analysis initially studied in the context of feed-forward architectures (E, 2017; Haber & Ruthotto, 2017; Li et al., 2017; Li & Hao, 2018) and recently in recurrent settings (Ceni et al., 2019; Chang et al., 2019; Lim, 2020; Sherstinsky, 2018; Niu et al., 2019; Herrera et al., 2020; Rubanova et al., 2019) and idealize the RNN as a continuous-time dynamical system. This allows us to phrase the problems under investigation in convenient analytical settings that accentuates the effect of dynamics. In this case, the RNNs serve

to approximate relationships represented by sequences of linear functions/functionals. On first look the setting appears to be simple, but we show that it yields representative results that underlie 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.

Our main results are:

1. We give a systematic analysis of the approximation of linear functionals by continuous-time linear RNNs, including a precise characterization of the approximation rates in terms of regularity and memory of the target functional.
2. We give 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 memory.

These results together paint a comprehensive picture of the interaction of learning and dynamics, and makes concrete the heuristic observations that the presence of long term memory affects RNN learning in a negative manner (Bengio et al., 1994; Hochreiter et al., 2001). In particular, mirroring the classical curse of dimensionality (Bellman, 1957), we introduce the concept of the *curse of memory* that captures the new phenomena that arises from learning temporal relationships: when there is long term memory in the data, one requires an exponentially large number of neurons for approximation, and the learning dynamics suffers from exponential slow downs. These results form a basic step towards a mathematical understanding of the recurrent structure and its effects on learning from temporal data.

2 Related Work

We will discuss related work on RNNs on three fronts concerning the central results in this paper, namely approximation theory, optimization analysis and the role of memory in learning. A number of universal approximation results for RNNs have been obtained in discrete (Schäfer & Zimmermann, 2006; 2007; Matthews, 1993) and continuous time (Funahashi & Nakamura, 1993; Nakamura & Nakagawa, 2009; Li et al., 2005; Chow & Xiao-Dong Li, 2000). Most of these focus on the case where the target relationship is generated from a hidden dynamical system in the form of difference of differential equations. The formulation of functional approximation here is more general, albeit our results are currently limited to the linear setting. Nevertheless, this is already sufficient to reveal new phenomena involving the interaction of learning and dynamics. This will be especially apparent when we discuss approximation rates and optimization dynamics. We also note that the functional/operator approximation using neural networks has been explored in (Chen & Chen, 1993; Tianping Chen & 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). The main difference here is that we explicit study the effect of memory in target functionals on learning using recurrent structures.

On the optimization side, there are a number of recent results concerning the training of RNNs using gradient methods, and they are mostly positive in the sense that trainability is proved under specific settings. These include recovering linear dynamics (Hardt et al., 2018) or training in over-parameterized settings (Allen-Zhu et al., 2019). Here, our result concerns the general setting of learning linear functionals that need not come from some underlying linear differential/difference equations, and is also away from the over-parameterized regime. In our case, we discover on the contrary that training can become very difficult even in the linear case, and this can be understood in a quantitative way, in relation to long term memory in the target functionals.

This points to the practical literature in relation to memory and learning. The fine-grain dynamical analysis puts 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 & Rolnick, 2018; Hanin, 2018) that are typically restricted to initialization settings with more precise characterizations in the dynamical regime during training.

3 PROBLEM FORMULATION

The basic problem of supervised learning on time series data is to learn a mapping from an input temporal sequence to an output sequence. Formally, one can think of the output at each time as being produced from the input via an unknown function that depends on the entire input sequence, or at least up to the time at which the prediction is made. In the discrete-time case, one can write the data generation process

$$y_k = H_k(x_0, \dots, x_k), \qquad k = 0, 1, \dots \tag{1}$$

where $\{H_k : k = 0, 1, \dots\}$ is a sequence of ground truth functions of increasing input dimension accounting for temporal evolution. The goal of supervised learning is to learn an approximation of the sequence of functions $\{H_k : k = 0, \dots, K\}$ given observation data.

Recurrent neural networks (RNN) (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

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

Here, $\{h_k\}$ are the *hidden/latent* states and its evolution is governed by a recursive application of a feed-forward layer with activation σ. We omit the bias term here and only consider a linear readout or output layer. For each time step k, the mapping $\{x_0, \dots, x_k\} \mapsto \hat{y}_k$ 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. To understand the working principles of RNNs, we need to characterize how $\{\hat{H}_k\}$ approximates $\{H_k\}$.

(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. This allows us to employ a large variety of continuum analysis tools to gain insights to the learning problem. Let us now introduce this framework.

Continuous-time formulation. Consider a sequence of inputs indexed by a real-valued variable $t \in \mathbb{R}$ instead of a discrete variable k considered previously. Concretely, we consider the 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. Here d is the dimension of each point in the time series. We write $\boldsymbol{x} \in \mathcal{X}$ to mean $\boldsymbol{x} := \{x_t \in \mathbb{R}^d : t \in \mathbb{R}\}$ and equip $\mathcal{X}$ with the supremum norm $\|\boldsymbol{x}\|_{\mathcal{X}} := \sup_{t\in\mathbb{R}} \|x_t\|_\infty$. For the space of outputs we will take a scalar time series, i.e. the space of bounded continuous functions from $\mathbb{R}$ to $\mathbb{R}$:

$$\mathcal{Y} = C_b(\mathbb{R}, \mathbb{R}). \tag{4}$$

Vector-valued outputs can be handled by considering each output separately, and will not be explicitly treated in the following analyses. In continuous time, the target relationship to be learned is

$$y_t = H_t(\boldsymbol{x}), \qquad t \in \mathbb{R} \tag{5}$$

where for each $t \in \mathbb{R}$, H_t is a functional $H_t : \mathcal{X} \to \mathbb{R}$. Correspondingly, we define a continuous version of (2) as a hypothesis space to model continuous-time functionals

$$\frac{d}{dt} h_t = \sigma(Wh_t + Ux_t), \qquad \hat{y}_t = c^\top h_t. \tag{6}$$

The dynamics then naturally defines a hypothesis space of sequences of functionals $\{\hat{H}_t(\boldsymbol{x}) = \hat{y}_t : t \in \mathbb{R}\}$, which can be used to approximate the target functionals $\{H_t\}$ via adjusting (c, W, U).

3.1 LINEAR RNN IN CONTINUOUS TIME

In this paper we mainly investigate the approximation and optimization property of linear RNN, which already reveals the essential effect of dynamics on RNNs. The linear RNN obeys (6) with σ the identity map. 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. In this case, (6) has the following solution

$$\hat{y}_t = \int_0^\infty c^\top e^{Ws} U x_{t-s} ds. \tag{7}$$

Since we will investigate uniform approximations over large time intervals, we will consider stable RNNs, where $W \in \mathcal{W}_m$ with

$$\mathcal{W}_m = \{W \in \mathbb{R}^{m\times m} : \text{eigenvalues of } W \text{ have negative real parts}\}. \tag{8}$$

Owing to the representation of solutions in (7), the linear RNN defines a family of functionals

$$\begin{aligned} \hat{\mathcal{H}} &:= \cup_{m\geq 1} \hat{\mathcal{H}}_m \\ \hat{\mathcal{H}}_m &:= \left\{ \{\hat{H}_t(\boldsymbol{x}), t \in \mathbb{R}\} : \hat{H}_t(\boldsymbol{x}) = \int_0^\infty c^\top e^{Ws} U x_{t-s} ds, W \in \mathcal{W}_m, U \in \mathbb{R}^{m\times d}, c \in \mathbb{R}^m \right\} \end{aligned} \tag{9}$$

Here, m is the width of the network and controls the complexity of the hypothesis space. 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.

Definition 3.1 *Let $\{H_t : t \in \mathbb{R}\}$ be a sequence of functionals.*

1. *H_t is* causal *if it does not depend on future values of the input: for every pair of $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}$ such that $x_s = x'_s$ for all $s \leq t$, we have $H_t(\boldsymbol{x}) = H_t(\boldsymbol{x'})$.*
2. *H_t is* continuous *and* linear *if for any $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}$ and $\lambda, \lambda' \in \mathbb{R}$ we have*
$$H_t(\lambda \boldsymbol{x} + \lambda' \boldsymbol{x'}) = \lambda H_t(\boldsymbol{x}) + \lambda' H_t(\boldsymbol{x'}) \tag{10}$$
and moreover that $\sup_{\boldsymbol{x}\in\mathcal{X}, \|\boldsymbol{x}\|_{\mathcal{X}}\leq 1} H_t(\boldsymbol{x}) < \infty$, in which case we can define the induced norm as $\|H_t\| := \sup_{\boldsymbol{x}\in\mathcal{X}, \|\boldsymbol{x}\|_{\mathcal{X}}\leq 1} |H_t(\boldsymbol{x})|$.
3. *H_t is* regular *if for any sequence $\{\boldsymbol{x}(n) \in \mathcal{X}, : n \geq 0\}$ such that $x(n)_t \to 0$ for Lebesgue almost every $t \in \mathbb{R}$, then $\lim_{n\to\infty} H_t(\boldsymbol{x}(n)) = 0$.*
4. *$\{H_t : t \in \mathbb{R}\}$ is* time-homogeneous *if for every $t, \tau \in \mathbb{R}$, we have $H_t(\boldsymbol{x}) = H_{t+\tau}(\boldsymbol{x}(\tau))$, where $x(\tau)_s = x_{s-\tau}$ for all $s \in \mathbb{R}$, i.e. $\boldsymbol{x}(\tau)$ is $\boldsymbol{x}$ whose time index is shifted to the right by τ.*

Linear, continuous and causal functionals are common definitions. 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. an infinitely thin spike input. 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 coordination, the output value remains the same. Given these definitions, the following observation can be verified directly and its proof is immediate and hence omitted.

Proposition 3.1 *Let $\{\hat{H}_t : t \in \mathbb{R}\}$ be a sequence of functionals in the RNN hypothesis space $\hat{\mathcal{H}}$ (see (9)). Then for each $t \in \mathbb{R}$, $\hat{H}_t$ is a causal, continuous, linear and regular functional. Moreover, the sequence of functionals $\{\hat{H}_t : t \in \mathbb{R}\}$ is time-homogeneous.*

4 Approximation Theory

The most basic approximation problem for RNN 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 of the form

$$H_t(\boldsymbol{x}) = y_t \qquad \text{where} \qquad \dot{h}_t = f(h_t, x_t), \qquad y_t = g(h_t) \tag{11}$$

for any linear or nonlinear functions f, g. This differs from previous work on approximation theory of RNNs, where it is assumed that the sequence of target functionals are indeed generated from such a system. In that case, the approximation problem reduces to that of the functions f, g, and the obtained results often resemble those in feed-forward 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 more flexible and natural, since in applications it is often not clear how one can describe the data generation process. Moreover, notice that in the linear case, if the target functionals $\{H_t\}$ are generated from a linear ODE system, then the approximation question is trivial: as long as h_t's dimension in the approximating RNN is greater than or equal to that which generates the data, we must have perfect approximation. However, we will see that in the more general case here, this question becomes much more interesting, even in the linear regime. In fact, we 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.

Our first main result is a converse of Proposition 3.1 in the form of an universal approximation theorem for certain classes of linear functionals. The proof is found in Appendix A.

Theorem 4.1 (Universal approximation of linear functionals) *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*

$$\sup_{t\in\mathbb{R}} \|H_t - \hat{H}_t\| \equiv \sup_{t\in\mathbb{R}} \sup_{\|\boldsymbol{x}\|_{\mathcal{X}}\leq 1} |H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| \leq \epsilon. \tag{12}$$

The key ingredient in the proof is the classical Riesz-Markov-Kakutani representation theorem, which allows us to pass from the approximation of functionals to that of functions. In particular, we can prove that if $\{H_t\}$ be a family of continuous, linear, regular, causal and time homogeneous functionals then there exists a unique integrable function $\rho : [0, \infty) \to \mathbb{R}^d$ such that

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top \rho(s) ds, \qquad t \in \mathbb{R}, \quad \boldsymbol{x} \in \mathcal{X}. \tag{13}$$

Comparing such representation with the solution (7) of the continuous RNN, we find that the approximation property of the linear RNNs is closely related to how well ρ can be approximated by the exponential sums, as detailed in Appendix A. Due to this representation, we will hereafter refer to H_t and $\rho(t)$ interchangeably.

Approximation Rates and the Curse of Memory in Approximation. 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 linear RNNs depends on the former's smoothness and memory properties. We note that this is a much less explored area in the approximation theory of RNNs.

To characterize smoothness and memory of linear 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 $e_{i,t} = e_i 1_{\{t\geq 0\}}$. Then, smoothness and memory is characterized by the regularity and decay rate of the maps $t \mapsto H_t(\boldsymbol{e}_i)$, $i = 1, \ldots, d$ respectively. Our second main result shows that these two properties are intimately tied with the approximation rate. The proof is found in Appendix B.

Theorem 4.2 (Approximation rate of linear RNN) *Assuming the conditions as in Theorem 4.1. Consider the output of constant signal $y_i(t) := H_t(\boldsymbol{e}_i)$, $i = 1, \ldots, d$. Suppose there exist constants $\alpha \in \mathbb{Z}^+, \beta, \gamma \in \mathbb{R}^+$ such that for $i = 1, \ldots, d$, $y_i(t) \in C^{(\alpha+1)}(\mathbb{R})$ and for $k = 1, \ldots, \alpha + 1$,*

$$e^{\beta t} y_i^{(k)}(t) = o(1) \text{ as } t \to +\infty, \qquad \text{and} \qquad \sup_{t\geq 0} \beta^{-k} |e^{\beta t} y_i^{(k)}(t)| \leq \gamma. \tag{14}$$

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

$$\sup_{t\in\mathbb{R}} \|H_t - \hat{H}_t\| \equiv \sup_{t\in\mathbb{R}} \sup_{\|\boldsymbol{x}\|_{\mathcal{X}}\leq 1} |H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| \leq \frac{C(\alpha)\gamma d}{\beta m^{\alpha}}. \tag{15}$$

For approximation of non-linear 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 4.2, the bound scales linearly with d. 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.

The key observation is that the rate result requires exponential decay of derivatives of $H_t(\boldsymbol{e}_i)$, but the density result (Theorem 4.1) makes no such assumption. The natural question is thus, what happens when no exponential decay is present? We assume $d = 1$ and consider an example in which the target functional's representation satisfies $\rho(t) \in C^{(1)}(\mathbb{R})$ and $\rho(t) \sim t^{-(1+\omega)}$ as $t \to +\infty$. Here $\omega > 0$ indicates the decay rate of the memory effects in our target functional family. 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 $\beta \in \mathbb{R}^+$ making (14) 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 $T \leq t \leq T+1$. With the auxiliary linear functional $\tilde{H}_t(\boldsymbol{x}) := \int_0^t x_{t-s}\tilde{\rho}(s)ds$, we can have an error estimate (with technical details provided in Appendix B)

$$\sup_{t\in\mathbb{R}} \|H_t - \hat{H}_t\| \leq \sup_{t\in\mathbb{R}} \|H_t - \tilde{H}_t\| + \sup_{t\in\mathbb{R}} \|\tilde{H}_t - \hat{H}_t\| \leq C\left(T^{-\omega} + \frac{\omega}{m}T^{1-\omega}\right). \tag{16}$$

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

$$m = \mathcal{O}\left(\omega T^{1-\omega}/\epsilon\right) = \mathcal{O}\left(\omega\epsilon^{-1/\omega}\right). \tag{17}$$

This estimate gives us a quantitative relationship between the degree of freedom needed and the decay speed. When ω is small, i.e., the system has long memory, the size of the RNN 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.

5 Fine-grained Dynamical Analysis of Optimizations

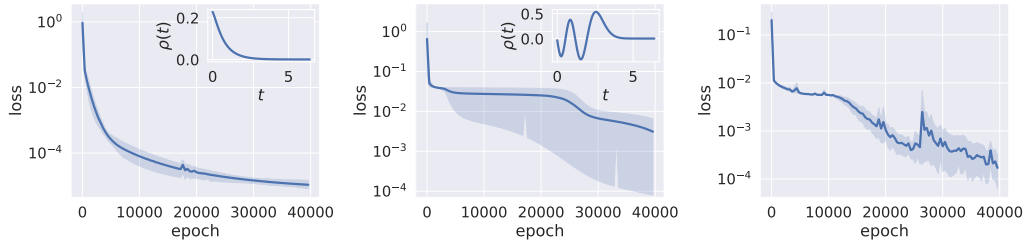
According to previous sections, memory plays an important role in determining the approximation rates. The results therein only depends on the model architecture, and does not concern the actual training dynamics. In this section, we perform a fine-grained analysis on the latter, which again reveals an interesting interaction between memory and learning.

The loss function (for training) is defined as

$$\mathbb{E}_{\boldsymbol{x}} J_m(\boldsymbol{x}; c, W, U) := \mathbb{E}_{\boldsymbol{x}} |\hat{H}_T(\boldsymbol{x}) - H_T(\boldsymbol{x})|^2 = \mathbb{E}_{\boldsymbol{x}} \left| \int_0^T [c^\top e^{Wt} U - \rho(t)^\top] x_{T-t} dt \right|^2, \tag{18}$$

where the input time series $\boldsymbol{x}$ is assumed to be finitely cut off, i.e. $x_t = 0$ for any $t \leq 0$ almost surely, without loss of generality. Training the RNN amounts to optimizing $\mathbb{E}_{\boldsymbol{x}} J_m$ with respect to the parameters (c, W, U). The most commonly applied method is (stochastic) gradient descent (GD), which updates them in the steepest descent direction.

We first show that the training dynamics of $\mathbb{E}_{\boldsymbol{x}} J_m$ exhibits very interesting behaviors depending on the form of target functionals. Take $d = 1$ and consider learning different target functionals with $\boldsymbol{x}$



(a) Exponential sum target (b) Airy function target (c) Lorenz 96 dynamics target

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. Observe that learning complex functionals (airy, lorenz) suffers from slow-downs in the form of long plateaus.

being the white noise, we first investigate two choices for ρ: a simple decaying exponential sum and a scaled airy function. Observe from Figure 1 that training proceeds efficiently for the exponential sum case. However, in the second airy function case, there are interesting "plateauing" behaviors of the loss function, where the loss decrement slows down significantly after some time in training. Also, the plateau is sustained for a long time before an eventual reduction is observed. As a further demonstration of that this behavior may be generic, we also consider a nonlinear forced dynamical system, the Lorenz 96 system (Lorenz, 1996), where the similar plateauing behavior is observed even for a non-linear RNN model trained with the Adam optimizer (Kingma & Ba, 2015). All experimental details are found in Appendix C.3.1. The results in Figure 1 hints at the fact that there are certainly some functionals that are much harder to learn than others, and it is the purpose of the remaining analyses to understand precisely when and why such difficulties occur. In particular, we will again relate this to memory effects in the target functional, which shows yet another facet of the *curse of memory* when it comes to optimization.

Dynamical Analysis. To make analysis amenable, we will make a sequence of simplifications to the training loss function (18), by assuming that $\boldsymbol{x}$ is the white noise, $d = 1$, $T \to \infty$, and the recurrent kernel W is diagonal. This allows us to write (see Appendix C.1 for full details) the optimization problem as

$$\min_{a\in\mathbb{R}^m, w\in\mathbb{R}_+^m} J_m(a,w) := \int_0^\infty \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t)\right)^2 dt. \tag{19}$$

We will subsequently see that these simplifications do not lose the key features of the training dynamics, i.e. plateauing behaviors. We start with some informal discussion on a probable reason behind the plateauing behavior. A straightforward computation shows, for $k = 1, 2, \ldots, m$,

$$\frac{\partial J_m}{\partial w_k}(a,w) = 2a_k \int_0^\infty (-t)e^{-w_k t}\left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t)\right) dt. \tag{20}$$

A similar expression holds for $\frac{\partial J_m}{\partial a_k}$. Write the (simplified) model as $\hat{\rho}_m(t;a,w) := \sum_{i=1}^m a_i e^{-w_i t}$. Observe that plateauing under GD dynamics occurs if the gradient ∇J_m is small but the loss J_m is large. A sufficient condition is the residual $\hat{\rho}_m(t;a,w) - \rho(t)$ is large *only for* large t (meaning the exponential term multiplied to the residual is small). That is, the learned functional differs from the target functional only at large times. This again relates to the long term memory (in the target).

Based on this observation, We build this memory effect explicitly into the target functional by considering ρ of the form

$$\rho(t) := \bar{\rho}(t) + \rho_{0,\omega}(t), \tag{21}$$

where $\bar{\rho}$ is a function which can be approximated well by $\hat{\rho}_m$, e.g. $\bar{\rho}(t)$ is the exponential sum $\sum_{j=1}^{m^*} a_j^* e^{-w_j^* t}$ with $m^* < m$ (where $w_j^* > 0$). While $\rho_{0,\omega}(t) := \rho_0(t - 1/\omega)$ ($\omega > 0$) controls the memory of (the target) ρ, with ρ_0 being any bounded template function in $L^2(\mathbb{R}) \cap C^2(\mathbb{R})$

with sub-Gaussian tails. As $\omega \to 0^+$, the support of $\rho_{0,\omega}$ shifts towards large times, modelling the dominance of long term memory in ρ. In this case, if the initialization is such that $\hat{\rho}_m \approx \bar{\rho}$, the sufficient condition informally discussed above is satisfied as $\omega \to 0^+$, which will lead to the plateauing behavior.

A simple example of such a target functional is $\rho(t) = a^* e^{-w^* t} + c_0 e^{-\frac{(t-1/\omega)^2}{2\sigma^2}}$, where $w^* > 0$ and c_0, σ are fixed constants. This corresponds to the simple case of $m^* = 1$ and ρ_0 being the Gaussian density. Observe that as $\omega \to 0^+$, the memory of this sequence of functionals (corresponding to ρ) increases. It can be numerically verified that this simple target functional gives rise to the plateauing behavior, which gets worse as memory increases (see Figure 2 in Appendix C.3.2 for details).

Our main result on training dynamics quantifies this behavior theoretically for general functionals possessing the decomposition (21). We now state the informal version, and the full technical details are given in Appendix C.2.

Theorem 5.1 (Informal) *Consider the gradient flow*

$$a'(\tau) = -\nabla_a J_m(a(\tau), w(\tau)), \qquad w'(\tau) = -\nabla_w J_m(a(\tau), w(\tau)), \tag{22}$$

with the initial value $\hat{\rho}_m(t; a(0), w(0)) \approx \bar{\rho}$. Then under technical conditions, the dynamical system (22) becomes trapped in a plateau with a timescale $\mathcal{O}\left(e^{c_{w^*}/\omega}\omega^2 \ln(1/\omega)\right)$ as $\omega \to 0^+$. Here $\mathcal{O}(\cdot)$ hides universal constants independent of ω and $c_{w^*} > 0$ is a global constant only depending on w^*.

The Curse of Memory in Optimization. The timescale proved in Theorem 5.1 is verified numerically in Appendix C.3.3, where we also show that the analytical setting here is representative of general cases, where plateauing occurs even in non-linear RNNs trained with accelerated optimizers, as long as the target functional has the memory structure imposed in (21). Theorem 5.1 reveals another aspect of the curse of memory, this time in optimization. When $\omega \to 0^+$, the influence of target functional H_t does not decay, much like the case considered in the curse of memory in approximation. However, different from the approximation case where an exponentially large number of hidden states is required to achieve approximation, here in optimization the adverse effect of the memory comes with the exponentially slow down of gradient flow dynamics. While this is theoretically proved under some sensible but restrictive conditions, we show numerically in Appendix C.3.3 that this is representative of the general case.

In the literature, a number of results have been obtained pertaining to the analysis of training dynamics of RNN. A positive result for training by GD is established in (Hardt et al., 2018), but this is in the setting of identifying linear systems, i.e. the target functional comes from a linear dynamical system, hence it must possess good decay properties provided it is stable. On the other hand, convergence can also be ensured if the RNN is sufficiently over-parameterized (large m) (Allen-Zhu et al., 2019). However, in reality both of these settings may not be sufficient. Here we provide an alternative analysis of a setting that is representative of the difficulties one may encounter in practical RNN training. In particular, the curse of memory that we established is consistent with the difficulty in RNN training often observed in applications, where heuristic attributions to "memory" are often alluded to (Hu et al., 2018; Campos et al., 2017; Talathi & Vartak, 2015; Li et al., 2018). The analysis here makes the connection between memory and optimization difficulty precise, and forms a basis for principled development of means to overcome such difficulties in applications.

6 Conclusion

In this paper, we analyzed the basic approximation and optimization aspects of using RNN to learn input-output relationships involving temporal sequences in the linear, continuous-time setting. In particular, we coined the concept *curse of memory*, and revealed two of its facets. That is, when the target relationship has the long term memory, both approximation and optimization become exceedingly difficult. These analyses make concrete heuristic observations of the adverse effect of memory on learning RNNs. Moreover, it quantifies the interaction between the structure of the model (RNN functionals) and the structure of the data (target functionals), where the latter is a much less studied topic. More broadly, this approach may be a basic starting point for understanding partially observed time series data in general, including gated RNN variants (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) and other methods 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.

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*, 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.

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.

Vladimir I Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

Dietrich Braess. Approximation by exponential sums. In *Nonlinear Approximation Theory*, pp. 168–180. Springer, 1986.

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. *arXiv preprint arXiv:1708.06834*, 2017.

Andrea Ceni, P. Ashwin, and L. Livi. Interpreting recurrent neural networks behaviour via excitable network attractors. *Cognitive Computation*, 12:330–356, 2019.

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*, 2019.

Tianping Chen and Hong Chen. Approximations of continuous functionals by neural networks with application to dynamical systems. *IEEE Transactions on Neural Networks*, 4:910–918, 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.

Weinan E. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017. ISSN 2194-6701.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801 – 806, 1993. ISSN 0893-6080.

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*, pp. 1764–1772, 2014.

Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pp. 545–552, 2009.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. IEEE, 2013.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. volume 37 of *Proceedings of Machine Learning Research*, pp. 1462–1471, Lille, France, 07–09 Jul 2015. PMLR.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):14004, 2017. ISSN 0266-5611.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems 31*, pp. 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*, pp. 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.

Calypso Herrera, Florian Krach, and Josef Teichmann. Theoretical guarantees for learning conditional expectation using controlled ode-rnn. *ArXiv*, abs/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 S. C. Kremer and J. F. Kolen (eds.), *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

Yuhuang Hu, Adrian E. G. Huber, Jithendar Anumula, and Shih-Chii Liu. Overcoming the vanishing gradient problem in plain recurrent networks. *CoRR*, abs/1801.06105, 2018.

Dunham Jackson. *The theory of approximation*, volume 11. American Mathematical Soc., 1930.

David W Kammler. Approximation with sums of exponentials in Lp[0, ∞). *Journal of Approximation Theory*, 16(4):384–408, April 1976. ISSN 0021-9045. doi: 10.1016/0021-9045(76)90071-X.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Qianxiao Li and Shuji Hao. An Optimal Control Approach to Deep Learning and Applications to Discrete-Weight Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2985–2994, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

Qianxiao Li, Long Chen, Cheng Tai, and Weinan E. Maximum principle based algorithms for deep learning. *The Journal of Machine Learning Research*, 18(1):5998–6026, 2017. ISSN 1532-4435.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 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, 10 2005.

Soon Hoe Lim. Understanding recurrent neural networks using nonequilibrium response theory. *ArXiv*, abs/2006.11052, 2020.

Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

Lu Lu, Pengzhan Jin, and G. Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *ArXiv*, abs/1910.03193, 2019.

Michael B. Matthews. Approximating nonlinear fading-memory operators using neural network models. *Circuits, Systems and Signal Processing*, 12:279–307, 1993.

Ch H Müntz. Über den approximationssatz von weierstrass. In *Mathematische Abhandlungen Hermann Amandus Schwarz*, pp. 303–312. Springer, 1914.

Yuichi Nakamura and Masahiro Nakagawa. Approximation capability of continuous time recurrent neural networks for non-autonomous dynamical systems. In *Artificial Neural Networks – ICANN 2009*, pp. 593–602, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

Murphy Yuezhen Niu, Lior Horesh, and Isaac Chuang. Recurrent neural networks in the eye of differential equations. *ArXiv*, abs/1904.12933, 2019.

Aaron 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. volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Yulia Rubanova, Ricky T. Q. Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 5320–5330. Curran Associates, Inc., 2019.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. In Stefanos D. Kollias, Andreas Stafylopatis, Włodzisław Duch, and Erkki Oja (eds.), *Artificial Neural Networks – ICANN 2006*, pp. 632–640, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-38627-8.

Anton Maximilian Schäfer and Hans-Georg Zimmermann. Recurrent neural networks are universal approximators. *International journal of neural systems*, 17(04):253–263, 2007.

Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *ArXiv*, abs/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*, pp. 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.

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, July 1995. ISSN 1941-0093. doi: 10.1109/72.392253.

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*, pp. 5998–6008, 2017.

A UNIVERSAL APPROXIMATION THEOREM OF LINEAR FUNCTIONALS BY LINEAR RNNS

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 A.1 (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*

$$H(\boldsymbol{x}) = \int_{\mathbb{R}} x_s^\top d\mu(s) = \sum_{i=1}^{d} \int_{\mathbb{R}} x_{s,i} d\mu_i(s). \tag{23}$$

Moreover, we have

$$\|H\| := \sup_{\|\boldsymbol{x}\|_{\mathcal{X}} \leq 1} |H(\boldsymbol{x})| = \|\mu\|_1(\mathbb{R}) := \sum_i |\mu_i|(\mathbb{R}). \tag{24}$$

Proof A.1 *Well-known, see e.g. Bogachev (2007), CH 7.10.4.*

We will use the representation theorem to prove Theorem 4.1. First, We prove some lemmas.

Lemma A.1 *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.*

$$\|\rho\|_{L^1([0,\infty))} := \sum_{i=1}^{d} \int_0^\infty |\rho_i(s)| ds < \infty \tag{25}$$

and

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top \rho(s) ds, \qquad t \in \mathbb{R}. \tag{26}$$

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 A.2 *By the Riesz-Markov-Kakutani representation theorem (theorem A.1), for each t there is a unique regular signed Borel measure μ_t such that*

$$H_t(\boldsymbol{x}) = \int_{\mathbb{R}} x_s^\top d\mu_t(s), \tag{27}$$

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 $\boldsymbol{x}$ and thus

$$H_t(\boldsymbol{x}) = \int_{-\infty}^t x_s^\top d\mu_t(s). \tag{28}$$

Now, by time homogeneity we have

$$\int_{-\infty}^t x_s^\top d\mu_t(s) = H_t(\boldsymbol{x}) = H_{t+\tau}(\boldsymbol{x}^{(\tau)}) = \int_{-\infty}^{t+\tau} x_{s-\tau}^\top d\mu_{t+\tau}(s). \tag{29}$$

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

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top d\mu(s). \tag{30}$$

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

$$\begin{aligned} |H_{t+\delta}(\boldsymbol{x}) - H_t(\boldsymbol{x})| &= \left| \int_0^\infty (x_{t+\delta-s} - x_{t-s})^\top d\mu(s) \right| \\ &\leq \sum_i \int_0^\infty \|x_{t+\delta-s} - x_{t-s}\|_\infty d|\mu_i|(s), \end{aligned} \tag{31}$$

which converges to 0 as $\delta \to 0$ by 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 $n \geq 0$ set $K_n \subset E, K'_n \subset E'$ where K_n, K'_n are closed and $\mu_i(E \setminus K_n) \leq 1/n$, $\mu_i(E' \setminus K'_n) \leq 1/n$. For a fixed $i \in \{1, \ldots, d\}$, define $\boldsymbol{x}^{(n)}$ to be such that $x^{(n)}_{t-s,j} = 0$ for all $j \neq i$ and all s. For $j = i$, we set $x^{(n)}_{t-s,i} = 1$ 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, $x^{(n)}_{t-s} \to 0$ for λ-a.e. s, thus by dominated convergence theorem

$$0 = \lim_{n\to\infty} H_t(\boldsymbol{x}^{(n)}) = \mu_i(E). \tag{32}$$

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

$$\int_A d\mu_i(s) = \int_A \rho_i(s)ds, \tag{33}$$

for $i = 1, \ldots, d$. Hence, we have

$$H_t(\boldsymbol{x}) = \int_0^\infty x^\top_{t-s}\rho(s)ds \tag{34}$$

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

Lemma A.2 *Let $\rho : [0, \infty) \to \mathbb{R}$ 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*

$$\left\|\rho - p(e^{-\cdot})\right\|_{L^1([0,\infty))} = \int_0^\infty |\rho(t) - p(e^{-t})|dt \leq \epsilon. \tag{35}$$

Proof A.3 *The approach here is similar to that of the approximation of functions using exponential sums (Kammler, 1976; Braess, 1986) Fix $\epsilon > 0$. Define*

$$R(u) = \begin{cases} \frac{1}{u}\rho(-\log u), & u \in (0, 1], \\ 0, & u = 0. \end{cases} \tag{36}$$

Then, we can check that

$$\|R\|_{L^1([0,1])} = \|\rho\|_{L^1([0,\infty))} < \infty. \tag{37}$$

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

$$\|R - \tilde{R}\|_{L^1([0,1])} \leq \epsilon/2. \tag{38}$$

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

$$\|q - \tilde{R}\|_{L^1([0,1])} \leq \epsilon/2, \tag{39}$$

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

$$\begin{aligned} \left\|\rho - p(e^{-\cdot})\right\|_{L^1([0,\infty))} &= \int_0^1 |R(u) - p(u)/u|du \\ &\leq \int_0^1 |R(u) - \tilde{R}(u)|du + \int_0^1 |\tilde{R}(u) - p(u)/u|du \leq \epsilon. \end{aligned} \tag{40}$$

We are now ready to present the proof of Theorem 4.1.

Proof A.4 (Proof of Theorem 4.1) *By (7), for each $\{\hat{H}_t\} \in \hat{\mathcal{H}}$ we can write*

$$\hat{H}_t(\boldsymbol{x}) = \int_0^\infty x^\top_{t-s}(U^\top[e^{Ws}]^\top c)ds. \tag{41}$$

By Lemma A.1, we can write

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top \rho(s) ds, \tag{42}$$

where ρ is integrable. Thus, we can apply Lemma A.2 to conclude that there exists polynomials $p_i, i = 1, \ldots, d$ with $p_i(0) = 0$ such that

$$\sum_i \|\rho_i - p_i(e^{-\cdot})\|_{L^1([0,\infty))} \leq \epsilon. \tag{43}$$

Notice that we can write each $p_i(u) = \sum_{j=1}^m \alpha_{ij} u^j$ for some m equaling the maximal order of $\{p_i\}$. Taking $W = \mathrm{diag}(-1, \ldots, -m)$, $c = (1, \ldots, 1)$ and $U_{ij} = \alpha_{ji}$, we have

$$(U^\top [e^{Ws}]^\top c)_i = p_i(e^{-s}), \qquad i = 1, \ldots, d. \tag{44}$$

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

$$\begin{aligned} |H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| = & \left| \int_0^\infty x_{t-s}^\top \rho(s) ds - \int_0^\infty x_{t-s}^\top p(e^{-s}) ds \right| \\ \leq & \sum_i \int_0^\infty |x_{t-s,i}| \left| \rho_i(s) - p_i(e^{-s}) \right| ds \leq \sum_i \|\rho_i - p_i(e^{-\cdot})\|_{L^1([0,\infty))} \\ \leq & \epsilon. \end{aligned} \tag{45}$$

B Approximation Rates and the Curse of Memory in Approximation

We first give the proof of Theorem 4.2.

Proof B.1 (Proof of Theorem 4.2) *We fix $i \in \{1, \ldots, d\}$ below until the last part of the proof. By Lemma A.1, there exists $\rho_i(t) \in C^\alpha[0, \infty)$ such that*

$$y_i(t) = H_t(\boldsymbol{e}_i) = \int_0^t \rho_i(r) dr, \quad t \geq 0. \tag{46}$$

By the assumption,

$$\rho_i^{(k)}(t) = o(e^{-\beta t}) \text{ as } t \to \infty, \quad k = 0, \ldots, \alpha. \tag{47}$$

Consider the transform

$$q_i(s) = \begin{cases} 0 & s = 0, \\ \frac{\rho_i\left(\frac{-(\alpha+1)\log s}{\beta}\right)}{s} & s \in (0, 1]. \end{cases} \tag{48}$$

For $k = 0, \ldots, \alpha$, one can prove by induction that

$$q_i^{(k)}(s) = \sum_{j=0}^k c(j,k) \left(-\frac{\alpha+1}{\beta}\right)^j \frac{\rho_i^{(j)}\left(\frac{-(\alpha+1)\log s}{\beta}\right)}{s^{j+1}}, \tag{49}$$

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

$$\left| q_i^{(k)}(e^{-\frac{\beta}{\alpha+1}t}) \right| = \left| \sum_{j=0}^k c(j,k) \left(-\frac{\alpha+1}{\beta}\right)^i \frac{\rho_i^{(j)}(t)}{e^{-\frac{(j+1)\beta}{\alpha+1}t}} \right| \leq \sum_{j=0}^k \left| c(j,k)|(\alpha+1)^j \right| \gamma \leq C(\alpha)\gamma, \tag{50}$$

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

$$\lim_{s \to 0^+} \frac{\rho_i^{(j)}\left(\frac{-(\alpha+1)\log s}{\beta}\right)}{s^{j+1}} = \lim_{t \to \infty} \frac{\rho_i^{(j)}(t)}{e^{-\frac{(j+1)\beta}{\alpha+1}t}} = \lim_{t \to \infty} \frac{\rho_i^{(j)}(t)}{e^{-\beta t}} e^{-\frac{(\alpha-j)\beta}{\alpha+1}t} = 0, \tag{51}$$

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 $m = 1, 2, \ldots$, there exists a polynomial $Q_{i,m}$ of degree $m-1$ such that

$$\|q_i - Q_{i,m}\|_{L^\infty([0,1])} \leq \frac{C(\alpha)\gamma}{m^\alpha}. \tag{52}$$

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

$$Q_{i,m}(s) = \sum_{j=0}^{m-1} \alpha_{i,j} s^j, \tag{53}$$

and define

$$\phi_{i,m}(t) = e^{-\frac{\beta}{\alpha+1}t} Q_{i,m}(e^{-\frac{\beta}{\alpha+1}t}). \tag{54}$$

Then we have

$$\phi_{i,m}(t) = c^\top e^{Wt} u_i, \tag{55}$$

where

$$c = (1, 1, \ldots, 1), \tag{56}$$

$$W = \begin{bmatrix} -\frac{\beta}{\alpha+1} & & & \\ & -\frac{2\beta}{\alpha+1} & & \\ & & \ddots & \\ & & & -\frac{m\beta}{\alpha+1} \end{bmatrix}, \tag{57}$$

$$u_i = (\alpha_{i,0}, \alpha_{i,1}, \ldots, \alpha_{i,m-1}). \tag{58}$$

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

$$\begin{aligned} \|\rho_i - \phi_{i,m}\|_{L^1([0,\infty))} &= \int_0^\infty |\rho_i(t) - \phi_m(t)| dt \\ &= \int_0^1 \left| \rho_i\left(\frac{-(\alpha+1)\log s}{\beta}\right) - sQ_{i,m}(s) \right| \frac{\alpha+1}{\beta s} ds \\ &= \frac{\alpha+1}{\beta} \int_0^1 |q_i(s) - Q_{i,m}(s)| ds \\ &\leq \frac{C(\alpha)\gamma}{\beta m^\alpha}. \end{aligned} \tag{59}$$

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

$$c^\top e^{Wt} U = (\phi_{1,m}(t), \ldots, \phi_{d,m}(t)). \tag{60}$$

Recall the dynamical system (6)

$$\frac{d}{dt} h_t = \sigma(Wh_t + Ux_t), \qquad \hat{y}_t = c^\top h_t, \tag{61}$$

which is together determined by the parameters c, W, U. Similar to the argument in the proof of Theorem 4.1, for any $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_\infty \leq 1$ and t, we have

$$|H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| \leq \sum_i \|\rho_i - \phi_{i,m}\|_{L^1([0,\infty))} \leq \frac{C(\alpha)\gamma d}{\beta m^\alpha}. \tag{62}$$

The curse of memory in approximation. Now we explain more technical details of why Theorem 4.2 implies the curse of memory in approximation, as pointed out in the main text. We assume $d = 1$ and consider an example

$$H_t(\boldsymbol{x}) \coloneqq \int_0^t x_{t-s} \rho(s) ds, \ t \geq 0, \tag{63}$$

in which the density satisfies $\rho(t) \in C^{(1)}(\mathbb{R})$ and

$$\rho(t) \sim t^{-(1+\omega)} \text{ as } t \to +\infty. \tag{64}$$

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 the following condition (eq. (14) in the main text) true:

$$e^{\beta t} y_i^{(k)}(t) = o(1) \text{ as } t \to +\infty, \qquad \text{and} \qquad \sup_{t \geq 0} \beta^{-k} |e^{\beta t} y_i^{(k)}(t)| \leq \gamma, \tag{65}$$

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 $\tilde{H}_t(\boldsymbol{x}) := \int_0^t x_{t-s} \tilde{\rho}(s) ds$, we have the truncation error estimate

$$|H_t(\boldsymbol{x}) - \tilde{H}_t(\boldsymbol{x})| \leq \|\boldsymbol{x}\|_{\mathcal{X}} \left(\int_T^\infty |\rho(s)| ds \right) \sim \|\boldsymbol{x}\|_{\mathcal{X}} T^{-\omega}. \tag{66}$$

Now Theorem 4.2 (eq. (62)) is applicable to the truncated $\{\tilde{H}_t\}$ (with $\alpha = 1$), and we have for $\forall \beta > 0$, there is a linear RNN (U, W, c) such that the associated functionals $\{\hat{H}_t\} \in \hat{\mathcal{H}}_m$ satisfy

$$\sup_{t \in \mathbb{R}} \|\tilde{H}_t - \hat{H}_t\| \leq \frac{C\gamma}{\beta m} := \frac{C}{\beta m} \sup_{t \geq 0} \frac{|e^{\beta t} y^{(1)}(t)|}{\beta} = \frac{C\omega}{m} \frac{e^{\beta T}}{\beta^2 T^{\omega+1}}. \tag{67}$$

It is straightforward to verify that when $\beta = 2/T$, the right-hand side of (67) achieves the minimum, which gives us

$$\sup_{t \in \mathbb{R}} \|\tilde{H}_t - \hat{H}_t\| \leq \frac{C\omega}{m} T^{1-\omega}. \tag{68}$$

Combining (66) and (68) gives us

$$\sup_{t \in \mathbb{R}} \|H_t - \hat{H}_t\| \leq C \left(T^{-\omega} + \frac{\omega}{m} T^{1-\omega} \right). \tag{69}$$

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

$$m = \mathcal{O}\left(\frac{\omega T^{1-\omega}}{\epsilon} \right) = \mathcal{O}\left(\omega \epsilon^{-\frac{1}{\omega}} \right). \tag{70}$$

This estimate gives us a quantitative relationship between the degree of freedom needed and the decay speed. When ω is small, i.e., the system has long memory, the size of the RNN required grows exponentially.

C Optimization Properties (Fine-grained Analysis)

C.1 Simplifications on the Loss

Recall the loss function

$$\mathbb{E}_{\boldsymbol{x}} J_m(\boldsymbol{x}; c, W, U) = \mathbb{E}_{\boldsymbol{x}} \left| \int_0^T [c^\top e^{Wt} U - \rho(t)^\top] x_{T-t} dt \right|^2, \tag{71}$$

The simplifications on $\mathbb{E}_{\boldsymbol{x}} J_m$ are listed as follows.

1. Take the inputs $\boldsymbol{x}$ to be the white noise, so that

$$x_{T-t} dt \overset{\text{in distribution}}{=} dB_t, \tag{72}$$

where B_t is the standard d-dimensional Wiener process. As a consequence, simplifying (71) via Itô's isometry gives

$$J_m(c, W, U) := \mathbb{E}_{\boldsymbol{x}} J_m(\boldsymbol{x}; c, W, U) = \int_0^T \left\| c^\top e^{Wt} U - \rho(t)^\top \right\|_2^2 dt. \tag{73}$$

2. Note that from the previous approximation results (Theorem 4.2), it is clear that the spatial dimension plays little role in the overall learning process (since each spatial dimension can be handled separately). Therefore, we focus on the temporal dimension and consider $d = 1$ in (73). Moreover, to study the training dynamics on large time horizons, we may take $T \to \infty$. Then we get
$$J_m(c, W, b) \coloneqq \int_0^\infty \left(c^\top e^{Wt} b - \rho(t)\right)^2 dt, \tag{74}$$
where $\rho(t)$ becomes a scalar-valued target and b is the sole column of U before. This corresponds to the so-called single-input-single-output (SISO) system.
3. Due to the difficulty of directly analyzing $\nabla_W e^{Wt}$ and $\nabla_W^2 e^{Wt}$, a further simplified ansatz will be considered. We suppose that W is a diagonal matrix with negative entries (to guarantee the stability of the model) so that $W = -\mathrm{diag}(w)$ with $w > 0$ element-wisely. Hence we can combine $a = b \circ c$ ($\circ$ denotes the Hadamard product) and rewrite the model as
$$\hat{\rho}(t; c, W, b) \coloneqq c^\top e^{Wt} b = \sum_{i=1}^m a_i e^{-w_i t} = \hat{\rho}_m(t; a, w). \tag{75}$$
Then (74) becomes
$$J_m(a, w) \coloneqq \int_0^\infty \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t)\right)^2 dt, \tag{76}$$
where $a \in \mathbb{R}^m$, $w \in \mathbb{R}_+^m$, and the target $\rho \in L^2([0, \infty)) \cap C^2([0, \infty))$ in principle.
4. As a further simplification, we will use a continuous-time idealization of the gradient descent dynamics by considering the gradient flow with respect to $J_m(a, w)$. That is, consider the dynamical system
$$\begin{cases} a'(\tau) = -\nabla_a J_m(a(\tau), w(\tau)), \\ w'(\tau) = -\nabla_w J_m(a(\tau), w(\tau)), \end{cases} \tag{77}$$
with some initial values $a(0) \in \mathbb{R}^m$, $w(0) \in \mathbb{R}_+^m$.

As we will show later, the loss function (76) is able to serve as the starting point in the fine-grained dynamical analysis, since it still preserves the plateauing behavior observed in the optimization process.

C.2 Concrete Analysis and the Curse of Memory in Optimization

The procedure of our analysis is as follows:

1. We consider using the model of exponential sums $\hat{\rho}_m(t; a, w) = \sum_{i=1}^m a_i e^{-w_i t}$, to learn the target with memory (21), i.e. $\rho(t) = \bar{\rho}(t) + \rho_{0,\omega}(t)$ with $\bar{\rho}$ being the well-approximated part (e.g. another exponential sum) and $\rho_{0,\omega} = \rho_0(t - 1/\omega)$ ($\omega > 0$) controlling the memory effect (e.g. the Gaussian density);
2. We prove that J_m has a large value but small gradients near the well-approximated part of the target (i.e. when $\hat{\rho}_m \approx \bar{\rho}$);
3. We prove that when $\hat{\rho}_m \approx \bar{\rho}$, the Hessian $\nabla^2 J_m$ is positive semi-definite as $\omega = 0$, but for $0 < \omega \ll 1$, $\nabla^2 J_m$ has $\mathcal{O}(1)$ positive eigenvalues and $o(1)$ negative eigenvalues. This is the most important part in our analysis;
4. Based on these results, we perform a local linearization analysis on the gradient flow (77) with the initialization $\hat{\rho}_m \approx \bar{\rho}$, from which the timescale of the plateauing behavior is derived.

(1) Target with memory

We consider the following form of the target with memory
$$\rho(t) = \bar{\rho}(t) + \rho_{0,\omega}(t) = \sum_{j=1}^{m^*} a_j^* e^{-w_j^* t} + \rho_0(t - 1/\omega). \tag{78}$$

Here, for the well-approximated part $\bar{\rho}$, $a_j^* \neq 0$, $w_j^* > 0$ and $w_i^* \neq w_j^*$ for any $i \neq j$, $i, j = 1, \cdots, m^*$, and $m^* < m$ (to ensure that the model can perfectly represent $\bar{\rho}(t)$). For the memory term $\rho_{0,\omega}$, $\omega > 0$, and ρ_0 is a fixed template function which satisfies the following assumptions.

Assumptions: (i) $\rho_0 \in L^2(\mathbb{R}) \cap C^2(\mathbb{R})$; (ii) $\|\rho_0\|_{L^2[0,\infty)} > 0$; (iii) ρ_0 is bounded on $\mathbb{R}$, i.e. $\sup_{t\in\mathbb{R}} |\rho_0(t)| < +\infty$; (iv) $\lim_{t\to-\infty} \rho_0(t) = 0$.

Remark C.1 *One can easily check a large set of functions satisfies the above assumptions, e.g. the sub-Gaussian function*

$$|\rho_0(t)| \leq c_0 e^{-c_1 t^2}, \quad |t| \geq t_0 \tag{79}$$

for some fixed positive constants c_0, c_1, t_0 [1].

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

Lemma C.1 *For any $w > 0$, $n \in \mathbb{N}$, let*

$$\Delta_{n,\omega}(w) := \int_0^\infty t^n e^{-wt} \rho_{0,\omega}(t) dt. \tag{80}$$

We have $\lim_{\omega\to 0^+} \Delta_{n,\omega}(w) = 0$. In particular, if ρ_0 is sub-Gaussian, we further have

$$\Delta_{n,\omega}(w) = \mathcal{O}\left(\omega^{-n} e^{-w/\omega}\right), \quad \omega \to 0^+, \tag{81}$$

where $\mathcal{O}(\cdot)$ hides universal constants only related to n, w.

Proof C.1 (Proof of Lemma C.1) *Since ρ_0 is bounded on $\mathbb{R}$, and $t^n e^{-wt} \in L^1([0,\infty))$, by Lebesgue's Dominant Convergence Theorem,*

$$\lim_{\omega\to 0^+} \Delta_{n,\omega}(w) = \int_0^\infty t^n e^{-wt} \cdot \lim_{\omega\to 0^+} \rho_{0,\omega}(t) dt \overset{\text{(iv)}}{=} 0, \quad \forall w > 0,\ \forall n \in \mathbb{N}, \tag{82}$$

where the last equality is due to Assumption (iv). That is, for any $t \geq 0$,

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

One can then estimate $\Delta_{n,\omega}(w)$ under the sub-Gaussian condition (79). Suppose $0 < \omega < 1/t_0$, we have

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

Let $M_0 := \sup_{t\in\mathbb{R}} |\rho_0(t)| < +\infty$, then

$$\begin{aligned}
I_1 &\leq M_0 \int_{1/\omega - t_0}^{1/\omega + t_0} t^n e^{-wt} dt \leq M_0 e^{-w(1/\omega - t_0)} \int_{1/\omega - t_0}^{1/\omega + t_0} t^n dt \\
&= M_0 e^{wt_0} \cdot e^{-w/\omega} \cdot \frac{(1 + \omega t_0)^{n+1} - (1 - \omega t_0)^{n+1}}{(n+1)\omega^{n+1}} \\
&= M_0 e^{wt_0} \omega^{-n} e^{-w/\omega} (2t_0 + \mathcal{O}(\omega)), \quad \omega \to 0^+,
\end{aligned}$$

[1] Obviously, if ρ_0 is proportional to the Gaussian density $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}}$, or ρ_0 is any continuous function with compact support on $[0,\infty)$ (denoted by $C_c([0,\infty))$, ρ_0 is sub-Gaussian. For the corresponding $\rho_{0,\omega}$, $1/\omega$ denotes respectively the mean of the normal distribution and start point of the support interval.

where $\mathcal{O}(\cdot)$ hides universal constants only related to n (with t_0 fixed).

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

where

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

for any $0<\omega<2c_1/w$. Here the last inequality utilizes the Mellin Transform of absolute moments of the Gaussian density (see Proposition 1 of Li et al. (2005)). The argument is similar for I_3, which in fact gives the same bound with I_2. Therefore, for any $w>0$, $n\in\mathbb{N}$, taking $0<\omega\ll 1$ gives $|\Delta_{n,\omega}(w)|\le C_{n,w}\omega^{-n}e^{-w/\omega}$, where $C_{n,w}>0$ is some fixed constant only depending on n,w (with $\sigma>0$ fixed). The proof is completed.

(2) Loss and gradient

The main idea to show plateauing behavior is to analyze the local dynamics of the gradient flow (77) when $\hat{\rho}_m\approx\bar{\rho}$. To do this, we need to analyze the derivatives of J_m at (a^*,w^*) (with degeneracy since $m>m^*$). As we will show later, the loss J_m remains lower bounded, but $\|\nabla J_m\|_2$ converges to 0 as $\omega\to 0^+$. This shows that the loss will saturate at a high value.

Definition C.1 (Weights degeneracy) *For any partition $\mathcal{P}$: $\{1,\cdots,m\}=\cup_{j=1}^{m^*}\mathcal{I}_j$, with $\mathcal{I}_{j_1}\cap\mathcal{I}_{j_2}=\varnothing$ for any $j_1\ne j_2$, $j_1,j_2=1,\cdots,m^*$, define the affine space (with respect to $\mathcal{P}$)*

$$\mathcal{M}_{\mathcal{P}}^*\coloneqq\left\{(\hat{a},\hat{w})\in\mathbb{R}^m\otimes\mathbb{R}_+^m:\hat{w}_i=w_j^*\text{ for any }i\in\mathcal{I}_j,\ \sum_{i\in\mathcal{I}_j}\hat{a}_i=a_j^*,\quad j=1,\cdots,m^*\right\}.$$

Define the collection of all such affine spaces $\mathcal{M}^*\coloneqq\bigcup_{\mathcal{P}}\mathcal{M}_{\mathcal{P}}^*$. Intuitively, $\mathcal{M}^*$ is the set of equivalent points to (a^*,w^*) for the purpose of approximation via exponential sums.

Remark C.2 *It is straightforward to check that the dimension of each $\mathcal{M}_{\mathcal{P}}^*$ 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 (see details in the proof of Theorem D.1).*

Proposition C.1 *For any $(\hat{a},\hat{w})\in\mathcal{M}_{\mathcal{P}}^*$, $J_m(\hat{a},\hat{w})\ge\|\rho_0\|_{L^2[0,\infty)}^2>0$. That is, the loss value is lower bounded uniformly in ω.*

Proof C.2 (Proof of Proposition C.1) *By Definition C.1 and Assumption (ii) on ρ_0, we have*

$$\begin{aligned}J_m(\hat{a},\hat{w}) &= \left\|\sum_{j=1}^{m^*}\sum_{i\in\mathcal{I}_j}\hat{a}_i e^{-\hat{w}_i t}-\sum_{j=1}^{m^*}a_j^* e^{-w_j^* t}-\rho_{0,\omega}(t)\right\|_{L^2[0,\infty)}^2 \\ &= \left\|\sum_{j=1}^{m^*}a_j^* e^{-w_j^* t}-\sum_{j=1}^{m^*}a_j^* e^{-w_j^* t}-\rho_{0,\omega}(t)\right\|_{L^2[0,\infty)}^2 \\ &= \|\rho_{0,\omega}\|_{L^2[0,\infty)}^2\ge\|\rho_0\|_{L^2[0,\infty)}^2>0.\end{aligned}$$

Proposition C.2 *For any bounded subset $\mathcal{M}_0^* \subset \mathcal{M}_{\mathcal{P}}^*$ and any $(\hat{a}, \hat{w}) \in \mathcal{M}_0^*$, we have*

$$\|\nabla J_m(\hat{a}, \hat{w})\|_2 = o(1) \qquad \text{as} \qquad \omega \to 0^+. \tag{83}$$

In particular, if ρ_0 satisfies the sub-Gaussian condition (79), we further have

$$\|\nabla J_m(\hat{a}, \hat{w})\|_2 = \mathcal{O}\left(\omega^{-1} e^{-c_{w^*}/\omega}\right) \qquad \text{as} \qquad \omega \to 0^+.$$

Here, $o(\cdot)$ and $\mathcal{O}(\cdot)$ hide universal constants related only to m, w^* and the diameter of $\mathcal{M}_0^*$, and $c_{w^*} > 0$ is a global constant only depending on w^*.

Proof C.3 (Proof of Proposition C.2) *A straightforward computation shows, for $k = 1, \cdots, m$,*

$$\frac{\partial J_m}{\partial a_k}(a, w) = 2\left[\sum_{i=1}^{m} \frac{a_i}{w_k + w_i} - \sum_{j=1}^{m^*} \frac{a_j^*}{w_k + w_j^*}\right] - 2\Delta_{0,\omega}(w_k), \tag{84}$$

$$\frac{\partial J_m}{\partial w_k}(a, w) = -2a_k\left[\sum_{i=1}^{m} \frac{a_i}{(w_k + w_i)^2} - \sum_{j=1}^{m^*} \frac{a_j^*}{(w_k + w_j^*)^2}\right] + 2a_k\Delta_{1,\omega}(w_k). \tag{85}$$

Notice that for any $n = 1, 2, \cdots$,

$$\begin{aligned}\sum_{i=1}^{m} \frac{\hat{a}_i}{(\hat{w}_k + \hat{w}_i)^n} - \sum_{j=1}^{m^*} \frac{a_j^*}{(\hat{w}_k + w_j^*)^n} &= \sum_{j=1}^{m^*} \sum_{i \in \mathcal{I}_j} \frac{\hat{a}_i}{(\hat{w}_k + \hat{w}_i)^n} - \sum_{j=1}^{m^*} \frac{a_j^*}{(\hat{w}_k + w_j^*)^n} \\ &= \sum_{j=1}^{m^*} \frac{\sum_{i \in \mathcal{I}_j} \hat{a}_i}{(\hat{w}_k + w_j^*)^n} - \sum_{j=1}^{m^*} \frac{a_j^*}{(\hat{w}_k + w_j^*)^n} = 0,\end{aligned} \tag{86}$$

we get

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

Combining with Lemma C.1 gives the desired conclusion.

(3) Hessian

The previous analysis justifies the trapping in the plateau. Now, we show that the dynamics will eventually escape this plateau by showing that, while the Hessian $\nabla^2 J_m$ (at $\hat{\rho}_m \approx \bar{\rho}$) becomes positive semi-definite as $\omega \to 0^+$ with $\mathcal{O}(1)$ positive eigenvalues, at sufficiently small but non-zero ω, $\nabla^2 J_m$ contains small negative eigenvalues, which results in the eventual escape. Two propositions are given in this regard, which shows respectively the minimal eigenvalue of $\nabla^2 J_m$ is small and negative.

Proposition C.3 *For any bounded subset $\mathcal{M}_0^* \subset \mathcal{M}_{\mathcal{P}}^*$, for any $(\hat{a}, \hat{w}) \in \mathcal{M}_0^*$, denote the eigenvalues of $\nabla^2 J_m(\hat{a}, \hat{w})$ by $\lambda_1 \geq \cdots \geq \lambda_{2m}$, we have*

$$\lambda_1 > 0, \cdots, \lambda_{m'} > 0, \lambda_{m'+1} = o(1), \cdots, \lambda_{2m} = o(1), \qquad \text{as} \qquad \omega \to 0^+,$$

with $m' \leq 2m^*$. In particular, if ρ_0 satisfies the sub-Gaussian condition (79), we can further quantify the scale $o(1)$ by

$$\mathcal{O}\left(\omega^{-2} e^{-c_{w^*}/\omega}\right), \qquad \text{as } \omega \to 0^+.$$

Here, $o(\cdot)$ and $\mathcal{O}(\cdot)$ hide universal constants only related to w^* and the diameter of $\mathcal{M}_0^*$, and $c_{w^*} > 0$ is a global constant only depending on w^*.

Proof C.4 (Proof of Proposition C.3) *A straightforward computation shows, for $k, l = 1, \cdots, m$,*

$$\frac{\partial^2 J_m}{\partial a_k \partial a_l}(a, w) = \frac{2}{w_k + w_l}, \tag{87}$$

$$\frac{\partial^2 J_m}{\partial a_k \partial w_l}(a, w) = \frac{-2a_l}{(w_k + w_l)^2}, \quad k \neq l, \tag{88}$$

$$\frac{\partial^2 J_m}{\partial a_k \partial w_k}(a, w) = -2\left[\sum_{i=1}^{m} \frac{a_i}{(w_k + w_i)^2} - \sum_{j=1}^{m^*} \frac{a_j^*}{(w_k + w_j^*)^2}\right] - \frac{a_k}{2w_k^2} + 2\Delta_{1,\omega}(w_k), \tag{89}$$

$$\frac{\partial^2 J_m}{\partial w_k \partial w_l}(a, w) = \frac{4a_k a_l}{(w_k + w_l)^3}, \quad k \neq l, \tag{90}$$

$$\frac{\partial^2 J_m}{\partial w_k \partial w_k}(a, w) = 4a_k\left[\sum_{i=1}^{m} \frac{a_i}{(w_k + w_i)^3} - \sum_{j=1}^{m^*} \frac{a_j^*}{(w_k + w_j^*)^3}\right] + \frac{a_k^2}{2w_k^3} - 2a_k\Delta_{2,\omega}(w_k). \tag{91}$$

By (86), we have

$$\frac{\partial^2 J_m}{\partial a_k \partial a_l}(\hat{a}, \hat{w}) = \frac{2}{\hat{w}_k + \hat{w}_l},$$

$$\frac{\partial^2 J_m}{\partial a_k \partial w_l}(\hat{a}, \hat{w}) = \frac{-2\hat{a}_l}{(\hat{w}_k + \hat{w}_l)^2} \quad (k \neq l), \qquad \frac{\partial^2 J_m}{\partial a_k \partial w_k}(\hat{a}, \hat{w}) = -\frac{\hat{a}_k}{2\hat{w}_k^2} + 2\Delta_{1,\omega}(\hat{w}_k),$$

$$\frac{\partial^2 J_m}{\partial w_k \partial w_l}(\hat{a}, \hat{w}) = \frac{4\hat{a}_k\hat{a}_l}{(\hat{w}_k + \hat{w}_l)^3} \quad (k \neq l), \qquad \frac{\partial^2 J_m}{\partial w_k \partial w_k}(\hat{a}, \hat{w}) = \frac{\hat{a}_k^2}{2\hat{w}_k^3} - 2\hat{a}_k\Delta_{2,\omega}(\hat{w}_k).$$

Let

$$\tilde{J}_m(a, w) := \left\| \sum_{i=1}^{m} a_i e^{-w_i t} - \sum_{j=1}^{m^*} a_j^* e^{-w_j^* t} \right\|_{L^2[0,\infty)}^2,$$

and

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

with $\Delta_{n,\omega}(w)$ ($n = 1, 2$) performed element-wisely and $\mathrm{Diag}(x)$ denoting the diagonal matrix with the diagonal vector being x. Then we have

$$\nabla^2 J_m(a, w) = \nabla^2 \tilde{J}_m(a, w) + 2\mathcal{E}_\omega(a, w). \tag{92}$$

We analyze $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$ and $\mathcal{E}_\omega(\hat{a}, \hat{w})$ respectively.

a) $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$: $\tilde{J}_m(\hat{a}, \hat{w}) = 0$ implies $\nabla \tilde{J}_m(\hat{a}, \hat{w}) = 0$ and $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$ is positive semi-definite. We further show that $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$ has multiple zero eigenvalues. In fact, since

$$\frac{\partial^2 \tilde{J}_m}{\partial a_k \partial a_l}(\hat{a}, \hat{w}) = \frac{2}{\hat{w}_k + \hat{w}_l},$$

$$\frac{\partial^2 \tilde{J}_m}{\partial a_k \partial w_l}(\hat{a}, \hat{w}) = \frac{-2\hat{a}_l}{(\hat{w}_k + \hat{w}_l)^2} \quad (k \neq l), \qquad \frac{\partial^2 \tilde{J}_m}{\partial a_k \partial w_k}(\hat{a}, \hat{w}) = -\frac{\hat{a}_k}{2\hat{w}_k^2},$$

$$\frac{\partial^2 \tilde{J}_m}{\partial w_k \partial w_l}(\hat{a}, \hat{w}) = \frac{4\hat{a}_k\hat{a}_l}{(\hat{w}_k + \hat{w}_l)^3} \quad (k \neq l), \qquad \frac{\partial^2 \tilde{J}_m}{\partial w_k \partial w_k}(\hat{a}, \hat{w}) = \frac{\hat{a}_k^2}{2\hat{w}_k^3},$$

it is straightforward to verify that $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})_{i,:} = \nabla^2 \tilde{J}_m(\hat{a}, \hat{w})_{j,:}$ and $\hat{a}_j \cdot \nabla^2 \tilde{J}_m(\hat{a}, \hat{w})_{m+i,:} = \hat{a}_i \cdot \nabla^2 \tilde{J}_m(\hat{a}, \hat{w})_{m+j,:}$ for any $i, j \in \mathcal{I}_p$, $p = 1, \cdots, m^*$, where $A_{i,:}$ denotes the i-th row of matrix A. That is to say, there are at most $2m^*$ different rows in $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w}) \in \mathbb{R}^{2m\times 2m}$, hence $\mathrm{rank}(\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})) \leq 2m^*$. Therefore, the number of zero eigenvalues of $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w}) \geq$

$\dim\left\{x \in \mathbb{R}^{2m} : \nabla^2 \tilde{J}_m(\hat{a}, \hat{w}) \cdot x = 0\right\} = 2m - \operatorname{rank}(\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})) \geq 2(m - m^*)$. *Since* $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$ *is positive semi-definite, all the nonzero eigenvalues must be positive.*

b) $\mathcal{E}_\omega(\hat{a}, \hat{w})$*: Let*

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

By Gershgorin's Circle Theorem, for any μ *being the eigenvalue of* $\mathcal{E}_\omega(\hat{a}, \hat{w})$*,* $\mu \in \bigcup_{k=1}^m (G_k^{(1)} \cup G_k^{(2)})$*. This gives* $|\mu| \leq |\hat{a}_k||\Delta_{2,\omega}(\hat{w}_k)| + |\Delta_{1,\omega}(\hat{w}_k)|$*,* $k = 1, \cdots, m$*.*

Combining with Lemma C.1 and Weyl's Theorem completes the proof.

Proposition C.4 *Suppose* $m > 2m^*$*,* [2] *and* $\Delta_{1,\omega}(w_j^*) \neq 0$ *for any* $\omega > 0, j = 1, \cdots, m^*$*. Then for any* $(\hat{a}, \hat{w}) \in \mathcal{M}_{\mathcal{P}}^*$*,* $\nabla^2 J_m(\hat{a}, \hat{w})$ *has negative eigenvalues.*

Proof C.5 (Proof of Proposition C.4) *By (92),* $\nabla^2 J_m(\hat{a}, \hat{w}) = \nabla^2 \tilde{J}_m(\hat{a}, \hat{w}) + 2\mathcal{E}_\omega(\hat{a}, \hat{w})$*. Denote the eigenvalues of* $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$ *and* $\mathcal{E}_\omega(\hat{a}, \hat{w})$ *by* $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_{2m}$ *and* $\mu_1 \geq \cdots \geq \mu_{2m}$ *respectively.*

a) $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$*: According to the proof of Proposition C.3,* $\nabla^2 \tilde{J}_m(\hat{a}, \hat{w})$ *has at least* $2(m - m^*)$ *zero eigenvalues, i.e.* $\tilde{\lambda}_1 > 0, \cdots, \tilde{\lambda}_{m'} > 0$ *and* $\tilde{\lambda}_{m'+1} = \cdots = \tilde{\lambda}_{2m} = 0$ *with* $m' \leq 2m^*$*.*

b) $\mathcal{E}_\omega(\hat{a}, \hat{w})$*:* $\Delta_{1,\omega}(w_j^*) \neq 0$ *implies that* $\operatorname{Diag}(\Delta_{1,\omega}(\hat{w}))$ *is invertible. By the congruent transformation in the form of*

$$\begin{bmatrix} I & \boldsymbol{O} \\ -\frac{1}{2} D_2 D_1^{-1} & I \end{bmatrix} \begin{bmatrix} \boldsymbol{O} & D_1 \\ D_1 & D_2 \end{bmatrix} \begin{bmatrix} I & -\frac{1}{2} D_1^{-1} D_2 \\ \boldsymbol{O} & I \end{bmatrix} = \begin{bmatrix} \boldsymbol{O} & D_1 \\ D_1 & \boldsymbol{O} \end{bmatrix},$$

in which D_1, D_2 *are diagonal matrices and* D_1 *is invertible, we know*

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

have the same index of inertia. One can verify that the eigenvalues of $\tilde{\mathcal{E}}_\omega(\hat{a}, \hat{w})$ *is* $\pm\Delta_{1,\omega}(\hat{w}_i)$ *for* $i = 1, \cdots, m$*. This gives* $\mu_1 > 0, \cdots, \mu_m > 0$ *and* $\mu_{m+1} < 0, \cdots, \mu_{2m} < 0$*.*

By Weyl's Theorem, the minimal eigenvalue of $\nabla^2 J_m(\hat{a}, \hat{w}) \leq \tilde{\lambda}_j + \mu_k$ *for* $j + k \leq 2m + 1$*. Let* $j = m > 2m^* \geq m'$ *and* $k = m + 1$*, we have* $\tilde{\lambda}_j = 0$ *and* $\mu_k < 0$*, which completes the proof.*

Remark C.3 *A simple case below shows that the assumption* $m > 2m^*$ *is not necessary. Let* $m = 2$*,* $m^* = 1$*. We can directly perform the congruent transformation on* $\nabla^2 J_2(\hat{a}, \hat{w})$*:*

$$\nabla^2 J_2(\hat{a}, \hat{w}) = \begin{bmatrix} \frac{1}{w^*} & \frac{1}{w^*} & \frac{-\hat{a}_1}{2w^{*2}} + 2\Delta_{1,\omega}(w^*) & \frac{\hat{a}_1 - a^*}{2w^{*2}} \\ \frac{1}{w^*} & \frac{1}{w^*} & \frac{-\hat{a}_1}{2w^{*2}} & \frac{\hat{a}_1 - a^*}{2w^{*2}} + 2\Delta_{1,\omega}(w^*) \\ \frac{-\hat{a}_1}{2w^{*2}} + 2\Delta_{1,\omega}(w^*) & \frac{-\hat{a}_1}{2w^{*2}} & \frac{\hat{a}_1^2}{2w^{*3}} - 2\hat{a}_1 \Delta_{2,\omega}(w^*) & \frac{\hat{a}_1(a^* - \hat{a}_1)}{2w^{*3}} \\ \frac{\hat{a}_1 - a^*}{2w^{*2}} & \frac{\hat{a}_1 - a^*}{2w^{*2}} + 2\Delta_{1,\omega}(w^*) & \frac{\hat{a}_1(a^* - \hat{a}_1)}{2w^{*3}} & \frac{(a^* - \hat{a}_1)^2}{2w^{*3}} - 2(a^* - \hat{a}_1)\Delta_{2,\omega}(w^*) \end{bmatrix}$$
$$\to \begin{bmatrix} \frac{1}{w^*} & 0 & 0 & 0 \\ 0 & 0 & -2\Delta_{1,\omega}(w^*) & 2\Delta_{1,\omega}(w^*) \\ 0 & -2\Delta_{1,\omega}(w^*) & \frac{\hat{a}_1^2}{4w^{*3}} + 2(\frac{\Delta_{1,\omega}(w^*)}{w^*} - \Delta_{2,\omega}(w^*))\hat{a}_1 - 4(\Delta_{2,\omega}(w^*))^2 w^* & \frac{\hat{a}_1(a^* - \hat{a}_1)}{4w^{*3}} + \frac{\Delta_{1,\omega}(w^*)}{w^*}(a^* - \hat{a}_1) \\ 0 & 2\Delta_{1,\omega}(w^*) & \frac{\hat{a}_1(a^* - \hat{a}_1)}{4w^{*3}} + \frac{\Delta_{1,\omega}(w^*)}{w^*}(a^* - \hat{a}_1) & \frac{(a^* - \hat{a}_1)^2}{4w^{*3}} - 2(a^* - \hat{a}_1)\Delta_{2,\omega}(w^*) \end{bmatrix}.$$

Since

$$\det\left(\begin{bmatrix} 0 & -2\Delta_{1,\omega}(w^*) & 2\Delta_{1,\omega}(w^*) \\ -2\Delta_{1,\omega}(w^*) & \frac{\hat{a}_1^2}{4w^{*3}} + 2(\frac{\Delta_{1,\omega}(w^*)}{w^*} - \Delta_{2,\omega}(w^*))\hat{a}_1 - 4(\Delta_{2,\omega}(w^*))^2 w^* & \frac{\hat{a}_1(a^* - \hat{a}_1)}{4w^{*3}} + \frac{\Delta_{1,\omega}(w^*)}{w^*}(a^* - \hat{a}_1) \\ 2\Delta_{1,\omega}(w^*) & \frac{\hat{a}_1(a^* - \hat{a}_1)}{4w^{*3}} + \frac{\Delta_{1,\omega}(w^*)}{w^*}(a^* - \hat{a}_1) & \frac{(a^* - \hat{a}_1)^2}{4w^{*3}} - 2(a^* - \hat{a}_1)\Delta_{2,\omega}(w^*) \end{bmatrix}\right)$$
$$= -4(\Delta_{1,\omega}(w^*))^2 \left(\frac{a^{*2}}{4w^{*3}} + o(1)\right) < 0, \qquad \omega \to 0^+,$$

[2] $m > 2m^*$ is not necessary for the conclusion to hold. See Remark C.3.

where $o(\cdot)$ hides polynomial dependence on $\Delta_{i,\omega}(w^*)$ ($i = 1, 2$) with the coefficients being polynomials of $\hat{a}$. Then we obtain that $\nabla^2 J_2(\hat{a}, \hat{w})$ has negative eigenvalues for any $(\hat{a}, \hat{w}) \in \mathcal{M}_0^*$ and $0 < \omega \ll 1$.

Remark C.4 *Under the assumption of Proposition C.4 and together with Proposition C.3, we obtain that $\nabla^2 J_m(\hat{a}, \hat{w})$ has small and negative eigenvalues for any $(\hat{a}, \hat{w}) \in \mathcal{M}_0^*$, where $\mathcal{M}_0^*$ (bounded) is any bounded subset of $\mathcal{M}_{\mathcal{P}}^*$.*

(4) 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 training dynamics can be trapped in a plateau with an *exponentially* large timescale. That is to say, *the curse of memory* does occur, and this time in optimization dynamics instead of approximation rates.

Theorem C.1 (Formal) *Let $\theta \coloneqq (a, w)$ and $\theta(\tau) \coloneqq (a(\tau), w(\tau))$. Consider the gradient flow*

$$\theta'(\tau) = -\nabla J_m(\theta(\tau)), \quad \theta(0) = (\hat{a}, \hat{w}). \tag{93}$$

Suppose $\Delta_{n,\omega}(w) = \mathcal{O}\left(\omega^{-n} e^{-w/\omega}\right)$ as $\omega \to 0^+$ for $n = 0, 1, 2$, [3] *and the assumptions of Proposition C.4 holds. Then for any bounded subset $\mathcal{M}_0^* \subset \mathcal{M}_{\mathcal{P}}^*$ and any $(\hat{a}, \hat{w}) \in \mathcal{M}_0^*$, the dynamical system (93) becomes trapped in the plateau with a timescale $\mathcal{O}\left(e^{c_{w^*}/\omega} \omega^2 \ln(1/\omega)\right)$ as $\omega \to 0^+$. Here $\mathcal{O}(\cdot)$ hides universal constants only related to m, w^* and the diameter of $\mathcal{M}_0^*$, and $c_{w^*} > 0$ is a global constant only depending on w^*.*

Proof C.6 (Proof of Theorem C.1) *Consider the asymptotic expansion with the form*

$$\theta(\tau) = \theta_0(\tau) + \sum_{i=1}^{\infty} \delta^i \theta_i(\tau) = \theta_0(\tau) + \delta\theta_1(\tau) + \delta^2\theta_2(\tau) + o(\delta^2), \tag{94}$$

for some $\delta \in (0, 1)$ (with $\delta \ll 1$) and $\theta_i(\tau) = \mathcal{O}(1)$ ($i = 0, 1, \cdots$). For consistency, we have $\theta_0(0) = \theta(0)$ and $\theta_i(0) = 0$ for $i = 1, 2, \cdots$. Let τ_0 be the hitting time

$$\tau_0 \coloneqq \inf\left\{\tau \geq 0 : \|\theta(\tau) - \theta(0)\|_2 > \delta\right\}. \tag{95}$$

Then by continuity and the definition of infimum, $\tau_0 > 0$ and $\|\theta(\tau) - \theta(0)\|_2 \leq \delta$ for any $\tau \in [0, \tau_0]$. It is our goal now to quantify the scale of τ_0.

Let $g_0 \coloneqq \nabla J_m(\theta(0))$ and $H_0 \coloneqq \nabla^2 J_m(\theta(0))$. The local linearization on (93) shows

$$\theta'(\tau) = -g_0 - H_0(\theta(\tau) - \theta(0)) + \mathcal{O}(\delta^2), \quad \tau \in [0, \tau_0].$$

Combining with (94), we have

$$\begin{aligned}
&\theta_0'(\tau) = -g_0 - H_0(\theta_0(\tau) - \theta(0)), \quad \theta_0(0) = \theta(0), && \text{at } \mathcal{O}(1) \text{ scale},\\
&\theta_1'(\tau) = -H_0\theta_1(\tau), \quad \theta_1(0) = 0, && \text{at } \mathcal{O}(\delta) \text{ scale},\\
&\theta_2'(\tau) = -H_0\theta_2(\tau) + \mathcal{O}(1), \quad \theta_2(0) = 0, && \text{at } \mathcal{O}(\delta^2) \text{ scale}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\theta_0(\tau) &= \theta(0) - \left(\int_0^\tau e^{-H_0 s} ds\right) g_0,\\
\theta_1(\tau) &= e^{-H_0\tau}\theta_1(0) = 0,
\end{aligned}$$

which gives

$$\theta(\tau) = \theta(0) - \left(\int_0^\tau e^{-H_0 s} ds\right) g_0 + \mathcal{O}(\delta^2), \quad \tau \in [0, \tau_0]. \tag{96}$$

[3] The assumption is reasonable according to Lemma C.1. That is, it can be sufficiently guaranteed by a sub-Gaussian ρ_0.

To achieve a parameter separation gap δ_0, i.e. $\|\theta(\tau)-\theta(0)\|_2 \le \delta_0$ with $\delta_0 = \mathcal{O}(\delta)$ and $\delta_0 \le \delta$, we need to take

$$\left\| \left(\int_0^\tau e^{-H_0 s} ds \right) g_0 \right\|_2 = \mathcal{O}(\delta). \tag{97}$$

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

$$\begin{aligned}\left\| \left(\int_0^\tau e^{-H_0 s} ds \right) g_0 \right\|_2 &= \left\| P^\top \left(\int_0^\tau e^{-\Lambda s} ds \right) P g_0 \right\|_2 \le \left\| \int_0^\tau e^{-\Lambda s} ds \right\|_2 \|g_0\|_2 \\ &= \|g_0\|_2 \max_{1\le i\le 2m} \frac{1}{|\lambda_i|} |e^{-\lambda_i \tau} - 1|.\end{aligned}$$

One can easily verify that $h_\tau(\lambda) := \frac{1}{|\lambda|}|e^{-\lambda\tau}-1|$, $\tau \ge 0$ monotonically decreases on $\lambda \in \mathbb{R}$ for any $\tau \ge 0$. [4] *Hence*

$$\left\| \left(\int_0^\tau e^{-H_0 s} ds \right) g_0 \right\|_2 \le \|g_0\|_2 \frac{1}{|\lambda_{2m}|} |e^{-\lambda_{2m}\tau} - 1|. \tag{98}$$

Let the right hand side of (98) equal to $\mathcal{O}(\delta)$, we can obtain a lower bound for the smallest τ such that (97) holds. According to Proposition C.2, Proposition C.3 and Proposition C.4, we have $\|g_0\|_2 = \mathcal{O}\left(\omega^{-1} e^{-c_{w^*}/\omega}\right)$ and $0 > \lambda_{2m} = \mathcal{O}\left(\omega^{-2} e^{-c_{w^*}/\omega}\right)$ as $\omega \to 0^+$, where $\mathcal{O}(\cdot)$ hides universal constants only related to m, w^* and the diameter of $\mathcal{M}_0^*$. For any $0 < \omega, \delta \ll 1$ with $\omega = o(\delta)$, solving RHS of (98) $= \mathcal{O}(\delta)$ gives

$$\tau \ge \mathcal{O}(e^{c_{w^*}/\omega} \cdot \omega^2 \ln(\delta/\omega)).$$

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

$$\begin{aligned}J_m(\theta(\tau)) - J_m(\theta(0)) &= g_0^\top (\theta(\tau)-\theta(0)) + (\theta(\tau)-\theta(0))^\top H_0 (\theta(\tau)-\theta(0)) + o(\delta^2) \\ &\ge -\|g_0\|_2 \|\theta(\tau)-\theta(0)\|_2 + \lambda_{2m}\|\theta(\tau)-\theta(0)\|_2^2 + o(\delta^2) \\ &= \mathcal{O}\left(\omega^{-1} e^{-c_{w^*}/\omega}\right)\mathcal{O}(\delta) + \mathcal{O}\left(\omega^{-2} e^{-c_{w^*}/\omega}\right)\mathcal{O}(\delta^2) + o(\delta^2) \\ &= o(\delta^2), \quad \omega \to 0^+.\end{aligned}$$

The proof is completed.

Remark C.5 *The estimate above is a lower bound on the escape time, so it does not appear to preclude the situation that the plateau will last forever. However, observe that in the proof above, suppose $\tau_0 = +\infty$ in (95), i.e. the hypothetical situation where the parameters are trapped forever, and write*

$$\tilde{g}_0 = P g_0 = (\tilde{g}_{0,1}, \cdots, \tilde{g}_{0,2m}),$$

we have

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

for any j such that $\lambda_j < 0$. If there is some $\tilde{g}_{0,j} \ne 0$, from (96) we get

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

which is a contradiction. That is to say, the parameter separation has to achieve a gap $\delta' > \delta$ within a finite time. This also implies a finite time of plateauing for any $\omega > 0$, even if it is exponentially large.

[4] With the convention that $h_\tau(0) = \tau$ for any $\tau > 0$ and $h_0(\lambda) \equiv 0$ for any $\lambda \in \mathbb{R}$.

Remark C.6 *As is shown, the exponential timescale estimate holds regardless of the specific partition $\mathcal{P}$. That is to say, the number of qualified affine spaces for initialization (since $\theta(0) = (\hat{a}, \hat{w}) \in \mathcal{M}_0^* \subset \mathcal{M}_{\mathcal{P}}^*$) is equal to that of partitions, i.e. $m^*! \begin{Bmatrix} m \\ m^* \end{Bmatrix}$, the cardinality of $\mathcal{M}^*$ (see details in the proof of Theorem D.1). In addition, initializations sufficiently near these affine spaces are also qualified by continuity.*

Remark C.7 *The dynamical analysis above is generically performed locally. However, motivated by the idea of weights degeneracy (see Definition C.1), we can further utilize similar methods to give a global landscape analysis on the loss function J_m. The results therein show that plateauing behaviors are likely to occur all over the landscape. See details in Appendix D.*

C.3 Numerical Experiments

C.3.1 Motivating Tests

In this section we give details of Figure 1.

(1) Learning linear functionals using linear RNNs (with GD optimizer)

The target functional is $H_T(\boldsymbol{x}) = \int_0^T \rho(t) x_{T-t} dt$ with $\boldsymbol{x}$ being the 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^* are standard normal random vectors of m^* dimensions, 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 $\mathrm{Ai}(t)$ is the Airy function of the first kind, given by the improper integral

$$\mathrm{Ai}(t) = \frac{1}{\pi} \lim_{\xi \to \infty} \int_0^\xi \cos\left(\frac{u^3}{3} + tu\right) du. \tag{99}$$

Note that in the first example, the memory of target functional decays quickly. However, for the second example, the effective rate of decay is controlled by the parameter t_0. For $t \leq t_0$, the airy function is oscillatory. Hence for large t_0, a large amount of memory is present in the target. In Figure 1 we set $m^* = 8$ for exponential sums, and $t_0 = 3$, $s_0 = 2.25$ for airy functions.

In Figure 1 (a) and (b), we plot the dynamics of gradient descent on training the linear RNNs (discretized using Euler method, hence equivalent to residual RNNs). We observe an efficient training process for the exponential sum case, while "plateauing" behaviors in the airy function case. This causes a severe slow down of training. In addition, we also find that the plateauing effect gets worse as t_0 (or s_0) is increased, which corresponds to more complex airy functions in the sense of more memory effects. That is, the long term memory adversely affects the optimization process via gradient descent.

(2) Learning nonlinear functionals using nonlinear RNNs (with Adam optimizer)

To show that the plateauing behavior may be generic, we also consider a nonlinear forced dynamical system target, the Lorenz 96 system (Lorenz, 1996):

$$\begin{aligned} \dot{y} &= -y + x + \sum_{k=1}^{K} z_k / K, \\ \dot{z}_k &= 2[z_{k+1}(z_{k-1} - z_{k+2}) - z_k + y], \quad k = 1, \cdots, K, \end{aligned} \tag{100}$$

with cyclic indices $z_{k+K} = z_k$, and x is an external stochastic noise. 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 (but not a linear functional). We use a standard nonlinear RNN (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 this system with the presence of memory also exhibits the interesting plateauing phenomenon.

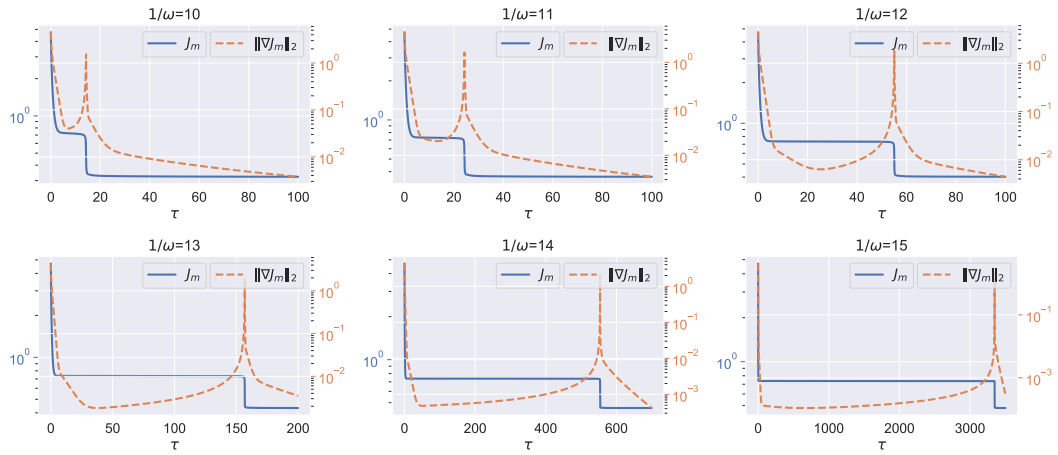


Figure 2: The training dynamics of the target functional defined by (101) using the model (102), where we set $m = 2$. Observe that the plateauing time increases rapidly as the memory $1/\omega$ becomes longer.

In all cases, the trained RNN model has a hidden dimension of 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.

C.3.2 Long-term Memory Contributes to Plateaus

Recall the simple example of target with memory

$$\rho(t) = a^* e^{-w^* t} + c_0 e^{-\frac{(t-1/\omega)^2}{2\sigma^2}}, \qquad w^* > 0. \tag{101}$$

We aim to learn ρ with the model of exponential sum

$$\hat{\rho}_m(t; a, w) = \sum_{i=1}^{m} a_i e^{-w_i t}, \qquad a \in \mathbb{R}^m,\ w \in \mathbb{R}_+^m. \tag{102}$$

That is, the simplified linear RNN model with the diagonal recurrent kernel. The optimization is performed by the gradient flow (77), i.e. a continuous-time idealization of the gradient descent dynamics. The ODE (77) is numerically solved by the Adams-Bashforth-Moulton method. The results are illustrated in Figure 2. It is shown that the plateauing time increases rapidly as the memory $1/\omega$ becomes longer.

C.3.3 Numerical Verification

(1) The timescale estimate

We first numerically verify the timescale proved in Theorem 5.1 (or Theorem C.1). That is, the time of plateauing (and also parameter separation $\|\theta(\tau) - \theta(0)\|_2$) is 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.

(2) General case

To facilitate mathematical tractability, the analysis so far is done on the restrictive cases of the diagonal recurrent kernel W with negative entries, linear activations and the gradient flow training dynamics. However, we show here that the plateauing behavior - which we now understood as a generic feature of long-term memory of the target functional and its interaction with the optimization dynamics - is present even when for general cases, and hence our simplified analytical setting is representative of the general situation.

In Figure 4, we take the target functional as defined in (101), but apply more general models to learn it, including using full (non-diagonal) recurrent kernels of the RNN with no restrictions on

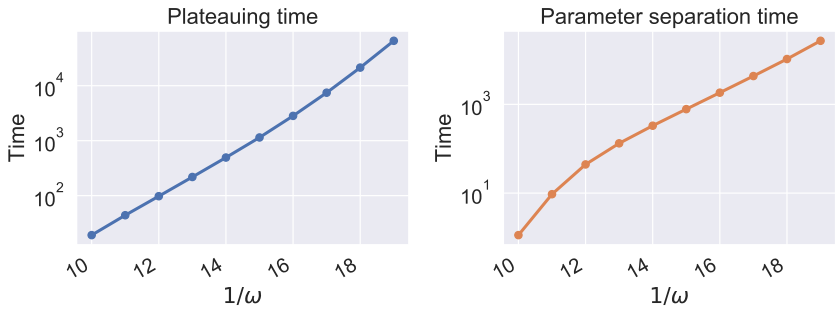


Figure 3: The timescale of plateauing and parameter separation. The logarithm of time of plateauing and parameter separation is almost linear to the memory $1/\omega$. Here the model and target are both selected the same as Figure 2, but with a larger width $m = 10$ in order to see the corresponding time data under large memories.

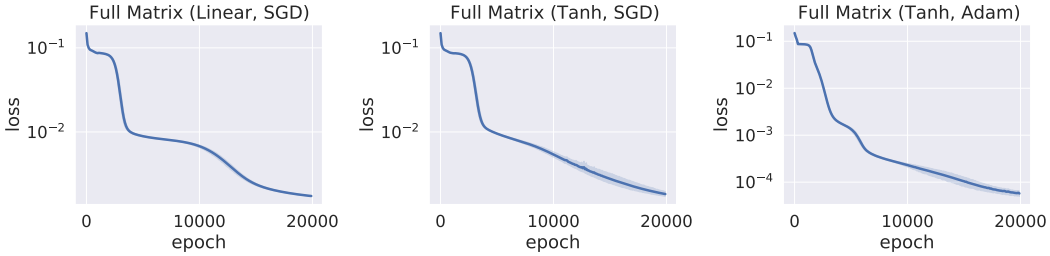


Figure 4: The plateauing behavior in general cases, with non-diagonal recurrent kernels, non-linear activations, and Adam (momentum-based) optimizers. Plateaus occur in all cases and momentum generally improves the situation but does not resolve the difficulty. Here we use the target functional the same as Figure 2 with $1/\omega = 20$. The time horizon is chosen as $T = 32$, and 128 input samples are generated from a standard Wiener process. Learning rate for GD is 1.0 and Adam is 0.001. For each experiment, 10 initial conditions are sampled and trained.

entries, using non-linear ($\tanh$) activations and using the Adam optimizer. Furthermore, we do not take the Itô isometry simplification and instead use actual input sample paths of finite time horizons, just as one would do in RNNs in practice. We observe that the plateauing behavior is present in all cases. Moreover, in the last case of Adam (which can be thought of as a momentum-based method), the plateauing behavior is somewhat alleviated, although the separation of timescales is still present. This is consistent with our supplemental analysis in Appendix E, where we show that the momentum-based methods will speed up training based on our dynamical and landscape analysis of plateauing.

D Landscape Analysis

As mentioned in Remark C.7, we can perform a global landscape analysis on the loss function based on the idea of weights degeneracy, which arises from Definition C.1. Recall that the loss function reads

$$\min_{a\in\mathbb{R}^m, w\in\mathbb{R}_+^m} J_m(a,w) := \int_0^\infty \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t)\right)^2 dt. \tag{103}$$

The main results of the appendix are summarized as follows.

- In Theorem D.1, we prove that the loss function has infinitely many critical points, which form a factorial number of affine spaces;

- In Theorem D.2, we prove that such (critical) affine spaces are much more than global minimizers provided the target being an exponential sum; [5]
- In Theorem D.3, we prove that on such (critical) affine spaces, the Hessian is singular in the sense of processing multiple zero eigenvalues;
- In Proposition D.1, 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 Appendix C.2, 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 Theorem 5.1 or Theorem C.1) in the following aspects.

- It is shown that the weights degeneracy is quite common in the whole landscape of the loss function. Unfortunately, the 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 from (see Theorem 5.1 or Theorem C.1).

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

D.1 Symmetry Analysis on the Landscape

This subsection consists of two parts: in Appendix D.1.1, we give main results provided the existence of weights degeneracy; in Appendix D.1.2, we give sufficient conditions to guarantee the existence. Since the key observation to utilize weights degeneracy is to notice the permutation symmetry of coordinates of gradients, we also called it "symmetry analysis".

D.1.1 Generic Theory

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

Definition D.1 *(coincided critical solutions and affine spaces) Let $d \in \mathbb{N}_+$ and $1 \leq d \leq m$. We say (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 D.2 *$(\hat{a}, \hat{w}) \in \mathbb{R}^m \otimes \mathbb{R}_+^m$ is called the non-degenerate global minimizer of J_m, if and only if*

$$J_m(\hat{a}, \hat{w}) = \inf_{a \in \mathbb{R}^m, w \in \mathbb{R}_+^m} J_m(a, w), \tag{104}$$

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

$$\hat{a}_i \neq 0, \ \hat{w}_i \neq \hat{w}_j \text{ for } i \neq j, \quad i, j = 1, 2, \cdots, m. \tag{105}$$

[5] 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 D.3 in Appendix D.1.1 for details.

For convenience, we also define an index set

$$\mathcal{N} \coloneqq \{n \in \mathbb{N}_+ : \ J_m \text{ has non} - \text{degenerate global minimizers for any } m \leq n\}, \tag{106}$$

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 D.1 *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 D.1 *We aim to show that there exists $\hat{a} \neq 0$ and $\hat{w} > 0$, such that*

$$J_1(\hat{a}, \hat{w}) = \inf_{a \in \mathbb{R}, w \in \mathbb{R}_+} J_1(a, w). \tag{107}$$

The basic idea is to limit the unbounded domain $a \in \mathbb{R}, w \in \mathbb{R}_+$ 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 $a(w) \coloneqq 2w\mathcal{L}[\rho](w)$. Write $h(w) \coloneqq J_1(a(w), w)$, 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 D.1 *If the target 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 D.1, and thus $\mathcal{N} \neq \varnothing$. In fact, $\mathcal{L}[\rho](w) = \sum_{j=1}^{m^*} \frac{a_j^*}{w + w_j^*}$ implies that $\mathcal{L}[\rho](w) = O(1)$ when $w \to 0^+$, and $\mathcal{L}[\rho](w) = O(1/w)$ when $w \to +\infty$.*

Theorem D.1 *Assume that $\mathcal{N} \neq \varnothing$ with $\mathcal{N}$ defined as (106). Let $M \coloneqq \sup \mathcal{N}$. Then for any $m \in \mathbb{N}_+$, $1 \leq d \leq \min\{m, M\}$, there exists at least $d! \begin{Bmatrix} m \\ d \end{Bmatrix}$ d-coincided critical affine spaces of J_m,* [6] *where $\begin{Bmatrix} m \\ d \end{Bmatrix} \in \mathbb{N}_+$ is called the Stirling number of the second kind.*

Proof D.2 *(i) Existence. The key observation is the permutation symmetry of ∇J_m: by (20), 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 $j_1 \neq j_2$, $j_1, j_2 = 1, \cdots, d$,

[6] The affine spaces degenerate to distinct points when $d = m$. For sufficient conditions to guarantee $M > 1$ (to avoid vacuous results), see Theorem D.6 and Remark D.7 in Appendix D.1.2.

define the affine space

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

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

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

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

$$\frac{\partial J_m}{\partial a_k}(a,w) = \frac{\partial J_d}{\partial b_s}(b,v), \quad b_s \frac{\partial J_m}{\partial w_k}(a,w) = a_k \frac{\partial J_d}{\partial v_s}(b,v), \quad k \in \mathcal{I}_s, \ s = 1, 2, \cdots, d. \tag{108}$$

[7] *Since* $d \le \min\{m, M\}$*,* $d \in \mathcal{N}$*. In fact, for any* $k \in \mathbb{N}_+$*, if* $k \notin \mathcal{N}$*, there exists* $i \le k$ *such that* $J_{(i)}$ *has no non-degenerate global minimizers, we have* $j \notin \mathcal{N}$ *for any* $j \ge i$*, hence* $M \le i - 1 \le k - 1$*. Hence* $M = \infty$ *implies* $\mathcal{N} = \mathbb{N}_+$ *and* $M < \infty$ *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* $(\hat{b}, \hat{v}) \in \mathbb{R}^d \otimes \mathbb{R}^d_+$ *such that*

$$J_d(\hat{b}, \hat{v}) = \inf_{b \in \mathbb{R}^d, v \in \mathbb{R}^d_+} J_d(b,v), \tag{109}$$

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

$$\hat{b}_i \neq 0, \ \hat{v}_i \neq \hat{v}_j \text{ for any } i \neq j, \quad i, j = 1, 2, \cdots, d. \tag{110}$$

By (109), we get $\nabla J_d(\hat{b}, \hat{v}) = 0$*. Combining with (108) and (110), 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*

[7] 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 some trivial degenerate cases (e.g. $a_k = 0$ or $b_s = 0$).

set of m labelled objects into d nonempty 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 D.1, Remark D.1 and Theorem D.1 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 being an exponential sum).

Theorem D.2 *Fix any $m \in \mathbb{N}_+$ relatively large. Consider the loss J_m associated with the target being 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 $m \in \mathcal{N}$ [8] with $\mathcal{N}$ defined in (106). 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.*

The following simple lemma will be used in the proof of Theorem D.2.

Lemma D.2 *For any $m \in \mathbb{N}_+$, $\lambda = (\lambda_1, \cdots, \lambda_m)$, $\lambda_i \neq \lambda_j$ for any $i \neq j$, $i, j = 1, \cdots, m$, the series of functions $\{e^{\lambda_i t}\}_{i=1}^m$ is linear independent on any interval $I \subset \mathbb{R}$.*

Proof D.3 *The aim is to show*

$$\sum_{i=1}^m c_i e^{\lambda_i t} = 0,\ t \in I \Rightarrow c_i = 0,\ i = 1, \cdots, m. \tag{111}$$

(111) holds trivially for $m = 1$. Assume that (111) holds for $m - 1$, then for any interval $I \subset \mathbb{R}$,

$$\begin{aligned} \sum_{i=1}^m c_i e^{\lambda_i t} = 0,\ t \in I &\Rightarrow \sum_{i=1}^{m-1} c_i e^{(\lambda_i - \lambda_m)t} + c_m = 0,\ t \in I \\ &\Rightarrow \sum_{i=1}^{m-1} c_i (\lambda_i - \lambda_m) e^{(\lambda_i - \lambda_m)t} = 0,\ t \in I. \end{aligned} \tag{112}$$

By induction, we get $c_i(\lambda_i - \lambda_m) = 0$ for $i = 1, \cdots, m-1$. Since λ_i's are distinct, we have $c_i = 0$ for $i = 1, \cdots, m-1$. Combining with (112), we get $c_m = 0$, which completes the proof.

Proof D.4 (Proof of Theorem D.2) *(i) Global minimizers. Since the target 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

$$\sum_{i=1}^m a_i e^{-w_i t} - \sum_{j=1}^m a_j^* e^{-w_j^* t} = 0, \quad t \geq 0. \tag{113}$$

It is easy to see that for any $j = 1, \cdots, m$, there exists $i(j)$ such that $w_{i(j)} = w_j^*$. Otherwise, if $w_i \neq w_j^*$, $i = 1, \cdots, m$, by (111) or Lemma D.2, we have $a_j^* = 0$, which is a contradiction. Notice that $w_i^* \neq w_j^*$ for any $i \neq j$, different w_j^*'s will correspond to different w_i's, hence the correspondence is one-to-one. Therefore, let $w_i = w_{j(i)}^*$, (113) 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 D.2, 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.

[8] Although the assumption $m \in \mathcal{N}$ seems strong, we will provide sufficient conditions to guarantee its validity in Appendix D.1.2. See an complement in Theorem D.7.

(ii) Coincided critical affine spaces. Obviously $\mathcal{N} \neq \varnothing$, and $M = \sup \mathcal{N} \geq m$. According to Theorem D.1, 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

$$\sum_{d=1}^{m-1} d! \left\{ {m \atop d} \right\}. \tag{114}$$

(iii) Comparison. To give a bound between (114) 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 D.2 *We only take the last two terms of (114) for a lower bound, which is obviously rather loose. In principle, a $\mathrm{Poly}(m)$ 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 see that even if only counting the number of affine spaces, there are still much less global minimizers (provided the width m relatively large).*

Remark D.3 *When the target ρ is not an exponential sum, it is straightforward to see Theorem D.2 still holds if there is a 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 D.3 *Fix any $m, d \in \mathbb{N}_+$, $1 \leq d \leq m$. On the d-coincided critical affine spaces (induced by non-degenerate global minimizers of J_d[9]) of J_m, $\nabla^2 J_m$ is with rank at most $m+d$, and hence has at least $m-d$ zero eigenvalues.*

Proof D.5 *A straightforward computation shows that, for $k, l = 1, 2, \cdots, m$,*

$$\frac{\partial^2 J_m}{\partial a_k \partial a_l}(a, w) = \frac{2}{w_k + w_l}, \tag{115}$$

$$\frac{\partial^2 J_m}{\partial a_k \partial w_l}(a, w) = \frac{-2a_l}{(w_k + w_l)^2}, \quad k \neq l, \tag{116}$$

$$\frac{\partial^2 J_m}{\partial a_k \partial w_k}(a, w) = \frac{-a_k}{2w_k^2} + 2\int_0^\infty (-t)e^{-w_k t} \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t) \right) dt. \tag{117}$$

[9] That is, the affine space $\mathcal{M}_{\mathcal{P},(\hat{b},\hat{v}),(m,d)}$. See details in the proof of Theorem D.1

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 D.1. Since $(\hat{b},\hat{v})$ is the non-degenerate global minimizer of J_d, we have

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

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

$$\frac{\partial^2 J_m}{\partial a_k\partial w_k}(\hat{a},\hat{w})=\frac{-\hat{a}_k}{2\hat{w}_k^2}. \tag{118}$$

Now we show that, for any $i,j\in\mathcal{I}_s$, $i\neq j$, $s=1,2,\cdots,d$, $\nabla^2 J_m(\hat{a},\hat{w})_{i,:}=\nabla^2 J_m(\hat{a},\hat{w})_{j,:}$. In fact, for any $k=1,\cdots,m$, let $k\in\mathcal{I}_{s'}$, then by (115),

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

For $k\neq i$ and $k\neq j$, (116) gives

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

By (118), 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 D.4 *The bound in Theorem D.3 is not sharp, since the estimate on $\operatorname{rank}\big(\nabla^2 J_m\big)$ here is loose as only rows with the same elements are considered. In practice (numerical tests), it is often observed 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 D.5 *Theorem D.3 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.*

D.1.2 Sufficient Conditions

There is still a gap when connecting Theorem D.1 and Theorem D.2. 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 target ρ within a smaller function space, the so-called completely monotonic functions.

Definition D.3 *$F\in C[0,\infty]\cap C^\infty(0,\infty)$ 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 D.6 *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

$$V_d(\mathbb{R}_+) := \left\{\hat{\rho} \in C^d[0,\infty) : [(D+w_1)\cdots(D+w_d)]\hat{\rho} = 0 \text{ for some } w_1, \cdots w_d \in \mathbb{R}_+\right\} \quad (119)$$

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 $d \in \mathbb{N}_+$.

Kammler (1979) proves the following theorem.

Theorem D.4 *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.*

$$\|\hat{\rho}^0 - \rho\|_{L^2[0,\infty)} = \inf_{\hat{\rho}\in V_d(\mathbb{R}_+)} \|\hat{\rho} - \rho\|_{L^2[0,\infty)}. \quad (120)$$

When $\rho \notin V_d(\mathbb{R}_+)$*, any such best approximation admits a non-degenerate form*

$$\hat{\rho}^0(t) = \sum_{j=1}^{d} \hat{b}_j e^{-\hat{v}_j t}, \quad 0 < \hat{v}_1 < \cdots < \hat{v}_d,\ \hat{b}_j > 0,\ j = 1, 2, \cdots, d, \quad (121)$$

and satisfies the generalized Aigrain-Williams equations

$$\mathcal{L}[\hat{\rho}^0](\hat{v}_j) = \mathcal{L}[\rho](\hat{v}_j), \quad j = 1, 2, \cdots, d, \quad (122)$$

$$\frac{d}{ds}\mathcal{L}[\hat{\rho}^0](s)\Big|_{s=\hat{v}_j} = \frac{d}{ds}\mathcal{L}[\rho](s)\Big|_{s=\hat{v}_j}, \quad j = 1, 2, \cdots, d. \quad (123)$$

Note that (120) and (121) are pretty similar to Definition D.2, except for a different choice of hypothesis function space. Now we show a connection between these two problems.

Theorem D.5 *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* $(\hat{b}, \hat{v}) \in \mathbb{R}^d \otimes \mathbb{R}_+^d$*.*

Proof D.6 *According to Theorem D.4, there exists a non-degenerate best approximation* $\hat{\rho}^0$ *to* ρ *from* $V_d(\mathbb{R}_+)$*, i.e.*

$$\|\hat{\rho}^0 - \rho\|_{L^2[0,\infty)} = \inf_{\hat{\rho}\in V_d(\mathbb{R}_+)} \|\hat{\rho} - \rho\|_{L^2[0,\infty)}, \quad (124)$$

$$\hat{\rho}^0(t) = \sum_{j=1}^{d} \hat{b}_j e^{-\hat{v}_j t}, \quad 0 < \hat{v}_1 < \cdots < \hat{v}_d,\ \hat{b}_j > 0,\ j = 1, 2, \cdots, d. \quad (125)$$

We aim to prove $J_d(\hat{b}, \hat{v}) = \inf_{b\in\mathbb{R}^d, v\in\mathbb{R}_+^d} J_d(b, v)$*. Define the following subsets of exponential sums*

$$\mathcal{V}_d(\mathbb{R}_+) := \left\{\hat{\rho} : \hat{\rho}(t) = \sum_{i=1}^{d} a_i e^{-w_i t}, a_i \in \mathbb{R}, w_i > 0\right\},$$

$$\mathcal{V}_{d,k}(\mathbb{R}_+) := \left\{\hat{\rho} \in \mathcal{V}_d(\mathbb{R}_+) : w = (w_i) \text{ has } k \text{ different components}\right\}, \quad 1 \le k \le d,$$

then we have $\inf_{b\in\mathbb{R}^d, v\in\mathbb{R}_+^d} J_d(b,v) = \inf_{\hat{\rho}\in\mathcal{V}_d(\mathbb{R}_+)} \|\hat{\rho}-\rho\|^2_{L^2[0,\infty)}$. *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 (124), 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 $\inf_{\hat{\rho}\in\mathcal{V}_d(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)} = \inf_{\hat{\rho}\in\mathcal{V}_{d,d}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)}$. *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* $\tilde{a} := (a_1,\cdots,a_k,0)$, $\quad \tilde{w} := (w_1,\cdots,w_k,1+\max_{1\le i\le k} w_i)$, *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 D.1 and Theorem D.5 immediately gives the following result.

Theorem D.6 *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 D.7 *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* $D_0=\infty$.[10] *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 D.5,* $J_{(k)}$ *has non-degenerate global minimizers for any* $k\le m'$, *i.e.* $m'\in\mathcal{N}$. *According to Theorem D.1, for any* $d\in\mathbb{N}_+$, $1\le d\le m' = \min\{m,m'\}\le\min\{m,M\}$, *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 D.7 *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 (114).*

An complement for Theorem D.2 is as follows.

[10] 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 $l\ge d$.

Theorem D.7 *Fix any $m \in \mathbb{N}_+$ relatively large. Consider the loss J_m associated with the target being 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 D.8 *By Theorem D.2, we only 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 D.5, $J_{(k)}$ 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^*).*

D.1.3 A Low-dimensional Example

To further understand the structure of coincided critical affine spaces, we focus on a specific low-dimensional example here. That is

$$\min_{a \in \mathbb{R}^2, w \in \mathbb{R}_+^2} J_2(a, w) = \left\| \sum_{i=1}^{2} a_i e^{-w_i t} - \rho(t) \right\|_{L^2[0,\infty)}^2,$$

with the target 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 D.1, Remark D.1 and Theorem D.1, 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 [11], 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 D.1 *Suppose $m = m^* = 2$, and $0 < w_1^* < w_2^*$. Let $I_1 := [0, \hat{a}]$ and $I_2 := (-\infty, 0) \cup (\hat{a}, +\infty)$ [12]. 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 D.9 *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*

$$\mathcal{A}_{(\hat{a},\hat{w})}(a_1) = \begin{bmatrix} \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{-a_1}{2\hat{w}^2} & \frac{-a_1}{2\hat{w}^2} & \frac{a_1^2}{2\hat{w}^3} + 4c(\hat{w})a_1 & \frac{a_1 a_2}{2\hat{w}^3} \\ \frac{-a_2}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} & \frac{a_1 a_2}{2\hat{w}^3} & \frac{a_2^2}{2\hat{w}^3} + 4c(\hat{w})a_2 \end{bmatrix}.$$

[11] Here we omit the corresponding partition $\mathcal{P}$ since it is unique.

[12] 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 conclusions hold.

Considering the congruent transformation of $\mathcal{A}_{(\hat{a},\hat{w})}(a_1)$, which does not affect the index of inertia:

$$\mathcal{A}_{(\hat{a},\hat{w})}(a_1) = \begin{bmatrix} \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{-a_1}{2\hat{w}^2} & \frac{-a_1}{2\hat{w}^2} & \frac{a_1^2}{2\hat{w}^3} + 4c(\hat{w})a_1 & \frac{a_1 a_2}{2\hat{w}^3} \\ \frac{-a_2}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} & \frac{a_1 a_2}{2\hat{w}^3} & \frac{a_2^2}{2\hat{w}^3} + 4c(\hat{w})a_2 \end{bmatrix}$$
$$\to \begin{bmatrix} \frac{1}{\hat{w}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{a_1^2}{4\hat{w}^3} + 4c(\hat{w})a_1 & \frac{a_1 a_2}{4\hat{w}^3} \\ 0 & 0 & \frac{a_1 a_2}{4\hat{w}^3} & \frac{a_2^2}{4\hat{w}^3} + 4c(\hat{w})a_2 \end{bmatrix},$$

we see 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

$$\det(\mathcal{A}'_{(\hat{a},\hat{w})}(a_1)) = a_1(\hat{a} - a_1) \cdot 4c(\hat{w}) \left(\frac{\hat{a}}{4\hat{w}^3} + 4c(\hat{w}) \right) \tag{126}$$
$$= \frac{1}{\hat{a}^2 \hat{w}^3} a_1(\hat{a} - a_1) \cdot \hat{a}c(\hat{w}) \cdot \left(\hat{a}^2 + 16\hat{w}^3 \hat{a}c(\hat{w}) \right). \tag{127}$$

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

$$\hat{a} = 2\hat{w} \sum_{j=1}^{m^*} \frac{a_j^*}{\hat{w} + w_j^*} = 4\hat{w}^2 \sum_{j=1}^{m^*} \frac{a_j^*}{(\hat{w} + w_j^*)^2}, \tag{128}$$

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

$$\begin{aligned} 8\hat{w}^3\hat{a}c(\hat{w}) &= \hat{a}^2 - 8\hat{a} \sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^3} \\ &= 16 \left[\sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^2} \right]^2 - 16 \sum_{j=1}^{m^*} \frac{a_j^*}{1 + v_j} \cdot \sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^3} \\ &= \frac{-16 a_1^* a_2^* (v_1 - v_2)^2}{(1 + v_1)^3 (1 + v_2)^3} \\ &> 0, \end{aligned} \tag{129}$$

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 (129),* $\hat{a}c(\hat{w}) < 0$.

$$\begin{aligned}
\hat{a}^2 + 16\hat{w}^3\hat{a}c(\hat{w}) &= 3\hat{a}^2 - 16\hat{a}\sum_{j=1}^{m^*}\frac{a_j^*}{(1+v_j)^3} \\
&= 16\left\{3\left[\sum_{j=1}^{m^*}\frac{a_j^*}{(1+v_j)^2}\right]^2 - 2\sum_{j=1}^{m^*}\frac{a_j^*}{1+v_j}\cdot\sum_{j=1}^{m^*}\frac{a_j^*}{(1+v_j)^3}\right\} \\
&= \frac{a_1^{*2}}{u_1^4u_2^4}\left[u_2^4 + c^2u_1^4 + 6cu_1^2u_2^2 - 2cu_1^3u_2 - 2cu_1u_2^3\right] \quad (c = a_2^*/a_1^*, u_j := 1+v_j > 1) \\
&= \frac{a_1^{*2}}{u_2^4}\left(s^4 + c^2 + 6cs^2 - 2cs - 2cs^3\right) \quad (s = u_2/u_1 > 1) \\
&= \frac{a_1^{*2}}{u_2^4}\left[c^2 - 2s(s^2-3s+1)c + s^4\right].
\end{aligned}$$

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 (126),* $\det(\mathcal{A}'_{(\hat{a},\hat{w})}(a_1))$ *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 (126), the proof is completed.*

Remark D.8 *From Proposition D.1, we see 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 of large magnitude.*

E MOMENTUM HELPS TRAINING: QUADRATIC EXAMPLES

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 (from plateaus) 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-in-time analogs of two optimization algorithms: gradient decent (GD) and momentum (heavy ball) method.

(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 $\delta = O(\epsilon)$. Thus $\|\nabla f(x_0)\|_2 = O(\epsilon)$, and

$$\begin{cases} x_1'(\tau) = -x_1(\tau), & x_1(0) = \delta \\ x_2'(\tau) = \epsilon x_2(\tau), & x_2(0) = 1 \end{cases} \Rightarrow \begin{cases} x_1(\tau) = \delta e^{-\tau} \\ x_2(\tau) = e^{\epsilon\tau} \end{cases}$$
$$\Rightarrow f(x(\tau)) = (\delta^2 e^{-2\tau} - \epsilon e^{2\epsilon\tau})/2 =: \ell_1(\tau).$$

It is easy to show that there are different timescales of $\ell_1(\tau)$. In fact, when $\tau = O(1/\epsilon)$, $\ell_1(\tau) = O(\epsilon^2)e^{-|O(1/\epsilon)|} - \epsilon e^{|O(1)|} = O(\epsilon)$. However, when τ continuous to increase, say

$$\tau \geq \frac{1}{2\epsilon}\ln\frac{\delta_0}{\epsilon} =: \tau_1^\epsilon, \tag{130}$$

where $\delta_0 > 0$ denotes the gap satisfying $\epsilon = o(\delta_0)$, we get $\ell_1(\tau) \leq \ell_1(\tau_1^\epsilon) = O(\epsilon^2) - \epsilon e^{2\epsilon\cdot\frac{1}{2\epsilon}\ln\frac{\delta_0}{\epsilon}}/2 = O(\epsilon^2) - \delta_0/2 < -\delta_0/4$ for any $\tau \geq \tau_1^\epsilon$.

(2) Momentum

The momentum algorithm has the update rule

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \rho(x_k - x_{k-1}), \tag{131}$$

where $\rho \in \mathbb{R}$, $\eta > 0$ is the learning rate. The continuous-in-time analog can be derived as (similar to the arguments in Su et al. (2014))

$$\begin{aligned} 0 &= \rho \frac{x_{k+1} - 2x_k + x_{k-1}}{\eta} + \frac{(1-\rho)}{\sqrt{\eta}} \frac{x_{k+1} - x_k}{\sqrt{\eta}} + \nabla f(x_k) \\ &\approx \rho x''(t) + \frac{(1-\rho)}{\sqrt{\eta}} x'(t) + \nabla f(x(t)), \end{aligned}$$

with $x_k \coloneqq x(k\sqrt{\eta})$ and the step size $\sqrt{\eta}$ of the simple finite differences [13]. 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 $\eta = 1$ [14] and $\rho = 1$ [15]. Plugging the expression of f, we can solve the ODE

$$\begin{aligned} x''(t) + \nabla f(x(t)) = 0 &\Leftrightarrow \begin{cases} x_1''(\tau) + x_1(\tau) = 0, & x_1(0) = \delta, \quad x_1'(0) = -\delta \\ x_2''(\tau) - \epsilon x_2(\tau) = 0, & x_2(0) = 1, \quad x_2'(0) = \epsilon \end{cases} \\ &\Rightarrow \begin{cases} x_1(\tau) = \delta(\cos\tau - \sin\tau) \\ x_2(\tau) = \frac{1+\sqrt{\epsilon}}{2} e^{\sqrt{\epsilon}\tau} + \frac{1-\sqrt{\epsilon}}{2} e^{-\sqrt{\epsilon}\tau} \end{cases} \\ &\Rightarrow f(x(\tau)) = \frac{1}{2}\left[\delta^2(\cos\tau - \sin\tau)^2 - \epsilon\left(\frac{1+\sqrt{\epsilon}}{2} e^{\sqrt{\epsilon}\tau} + \frac{1-\sqrt{\epsilon}}{2} e^{-\sqrt{\epsilon}\tau}\right)^2\right]. \end{aligned}$$

Write $\ell_2(\tau) \coloneqq f(x(\tau))$. It is not hard to show that there are still different timescales of $\ell_2(\tau)$. In fact, when $\tau = O(1/\sqrt{\epsilon})$, $\ell_2(\tau) = O(\epsilon^2)|O(1)| - \epsilon|O(1)|(e^{|O(1)|} + e^{-|O(1)|})^2 = O(\epsilon)$. However, when τ continuous to increase, say

$$\tau \geq \frac{1}{2\sqrt{\epsilon}} \ln \frac{4\delta_0}{\epsilon} =: \tau_2^\epsilon, \tag{132}$$

we get $\ell_2(\tau_2^\epsilon) = O(\epsilon^2) - \epsilon(\frac{1+\sqrt{\epsilon}}{2})^2 e^{2\sqrt{\epsilon}\tau}/2 + O(\epsilon) = O(\epsilon) - \delta_0 < -\delta_0/2$, hence $\ell_2(\tau) \leq O(\epsilon) - \delta_0 < -\delta_0/2$ for any $\tau \geq \tau_2^\epsilon$.

Combining **(1)** and **(2)** gives that

- For both training dynamics, there are different timescales in the loss function. That is to say, relatively long time is needed to escape from the plateaus;
- Compare (130) and (132), we get different timescales of escaping: $O\left(1/\epsilon \cdot \ln(1/\epsilon)\right)$ for GD and $O\left(1/\sqrt{\epsilon} \cdot \ln(1/\epsilon)\right)$ for momentum. Just like the convex case, where momentum improves the convergence rate by weakening the dependence on condition number, we see momentum can also help to escape rather flat saddles.

[13] It is easy to check that the error of discretization is of order $O(\sqrt{\eta})$.

[14] In the continuous-in-time analog of GD, i.e. the gradient flow, the step size is taken as 1.

[15] As is seen later, $\rho = 1$ not only simplifies the analysis, but also helps to obtain the best acceleration.