2023-08-28

Kernel limit of recurrent neural networks trained on ergodic data sequences

Kernel Limit of Recurrent Neural Networks Trained on Ergodic Data Sequences

Samuel Chun-Hei Lam[*], Justin Sirignano[†], and Konstantinos Spiliopoulos[‡§¶]

August 29, 2023

Abstract

Mathematical methods are developed to characterize the asymptotics of recurrent neural networks (RNN) as the number of hidden units, data samples in the sequence, hidden state updates, and training steps simultaneously grow to infinity. In the case of an RNN with a simplified weight matrix, we prove the convergence of the RNN to the solution of an infinite-dimensional ODE coupled with the fixed point of a random algebraic equation. The analysis requires addressing several challenges which are unique to RNNs. In typical mean-field applications (e.g., feedforward neural networks), discrete updates are of magnitude $\mathcal{O}(\frac{1}{N})$ and the number of updates is $\mathcal{O}(N)$. Therefore, the system can be represented as an Euler approximation of an appropriate ODE/PDE, which it will converge to as $N \to \infty$. However, the RNN hidden layer updates are $\mathcal{O}(1)$. Therefore, RNNs cannot be represented as a discretization of an ODE/PDE and standard mean-field techniques cannot be applied. Instead, we develop a fixed point analysis for the evolution of the RNN memory states, with convergence estimates in terms of the number of update steps and the number of hidden units. The RNN hidden layer is studied as a function in a Sobolev space, whose evolution is governed by the data sequence (a Markov chain), the parameter updates, and its dependence on the RNN hidden layer at the previous time step. Due to the strong correlation between updates, a Poisson equation must be used to bound the fluctuations of the RNN around its limit equation. These mathematical methods give rise to the neural tangent kernel (NTK) limits for RNNs trained on data sequences as the number of data samples and size of the neural network grow to infinity.

1 Introduction

Recurrent neural networks (RNN) are widely used to model sequential data. Examples include natural language processing (NLP) and speech recognition [1, 2]. The key architectural feature of an RNN is a hidden layer which is updated at each time step of the sequence. This hidden layer – sometimes referred to as a "memory layer" – is a nonlinear representation of the history of the data sequence. Using its hidden layer, the RNN can – in principle – learn functions which map the path of a sequence (of arbitrary length) to fixed-dimensional vector predictions. The RNN's hidden layer therefore provides a parsimonious, nonlinear representation of the data in the sequence up until the current time. An RNN is trained by minimizing an appropriate loss function over a high-dimensional set of parameters using a gradient-descent-type algorithm.

The mathematical theory for RNNs is limited. In this article, we study the asymptotics of a single-layer RNN as the number of hidden units, training steps, and data samples in the sequence tend to infinity. In the case of an RNN with a simplified weight matrix, we prove the convergence of the RNN to the solution of an infinite-dimensional ODE coupled with the fixed point of a random algebraic equation.

[*]Mathematical Institute, University of Oxford, E-mail: Samuel.Lam@math.ox.ac.uk.
[†]Mathematical Institute, University of Oxford, E-mail: Justin.Sirignano@maths.ox.ac.uk.
[‡]Department of Mathematics & Statistics, Boston University, E-mail: kspiliop@math.bu.edu.
[§]This article is part of the project "DMS-EPSRC: Asymptotic Analysis of Online Training Algorithms in Machine Learning: Recurrent, Graphical, and Deep Neural Networks" (NSF DMS-2311500).
[¶]Konstantinos Spiliopoulos was also partially supported by NSF DMS-2107856 and Simons Foundation Award 672441.

For feedforward neural networks with i.i.d. data samples, limits have been proven as the number of hidden units, training steps, and data samples tend to infinity. The dynamics of the output of the trained network converges to either an ODE (the Neural Tangent Kernel NTK limit) [3] or a (random) PDE (the mean-field limit) [4, 5, 6, 7] depending upon the normalization used for the neural network output. For the NTK case, the equation for the limit neural network can be studied to prove global convergence of the neural network to the global minimizer of the objective function. Proving limit results for RNNs is substantially more challenging. The data sequence is not i.i.d., which complicates the analysis of the evolution of the trained neural network. Furthermore, the RNNs cannot be studied using standard mean-field or weak convergence analysis (e.g., as is true for feedforward neural networks). We explain in more detail below.

Consider a classic recurrent neural network (the standard Elman network [8]) with one hidden layer that takes in the input sequence $X = (X_k)_{k\geq 0}$ and outputs a prediction $(\hat{Y}_k^N)_{k\geq 0}$ for the target data $(Y_k^N)_{k\geq 0}$. The RNN predictions are given by the model outputs $\hat{Y}_k^N = g_k^N(X;\theta)$. The RNN depends on the parameters $\theta = (C, W, B)$, which must be trained on data. In particular, for all $k \geq 0$, the RNN hidden layer S_k^N and predictions $\hat{Y}_k^N$ are updated as:

$$S_{k+1}^{i,N}(X;\theta) := \sigma\left((W^i)^\top X_k + \frac{1}{N^{\beta_1}}\sum_{j=1}^N B^{ij} S_k^{j,N}(X;\theta)\right), \quad S_0^{i,N}(X;\theta) = 0 \tag{1.1}$$

$$g_k^N(X;\theta) := \frac{1}{N^{\beta_2}}\sum_{i=1}^N C^i S_{k+1}^{i,N}(X;\theta), \tag{1.2}$$

where

- N is the number of hidden units in the memory states $S_k^{i,N}(X;\theta)$,
- $\beta_1 = 1$ and β_2 determine the scaling used to normalise the outputs of the network,
- $C \in \mathbb{R}^N$, with C^i representing the i-th component of C,
- $W \in \mathbb{R}^{N\times d}$, with W^i representing the i-th column of W as a column vector, and
- $B \in \mathbb{R}^{N\times N}$, with B^{ij} is the (i,j)-entries of B.

The data samples X_k, which are elements of a data sequence, are *not* i.i.d. (unlike for feedforward neural networks). In our mathematical analysis, we will make the simplifying assumption that all columns of B are equal, i.e. for all j we have $B^{ij} = B^j$ for some B^j. The memory states $S_k^{N,i}(X;\theta)$ is a nonlinear representation of the history of the data sequence $(X_j)_{j=0}^{k-1}$. Using this nonlinear representation – which is learned from the data by training the parameters θ – the RNN generates a prediction $\hat{Y}_k^N$ for the target data Y_k. Notice that if we fix $X_k = x$ and $B^{ij} = 0$, (1.2) becomes a standard feedforward network (i.e., the network does not dynamically evolve over time k and the network output is a static prediction). Limits for gradient-descent-trained feedforward networks as the number of hidden units $N \to \infty$ can been established when $\frac{1}{2} \leq \beta_2 < 1$ (the NTK limit [3]) or $\beta_2 = 1$ (the mean-field limit [4, 5, 6, 7]).

A "typical" limit ODE from mean-field analysis will not occur in for the RNNs and standard mean-field techniques (see, for example, [9]) cannot be directly applied. As an illustrative example, standard mean-field techniques would be applicable to a neural network with the following updates:

$$S_{k+1}^{i,N}(X;\theta) := S_k^{i,N}(X;\theta) + \frac{1}{N}\sigma\left((W^i)^\top X_k + \frac{1}{N^{\beta_1}}\sum_{j=1}^N B^{ij} S_k^{j,N}(X;\theta)\right), \quad S_0^{i,N}(X;\theta) = 0. \tag{1.3}$$

(1.3) is an Euler approximation of an ODE with step size $\mathcal{O}(1/N)$. (1.3) is a standard mean-field framework and it can clearly be proven as $N \to \infty$ that (1.3) converges to an appropriate infinite-dimensional ODE. However, the $S_k^{i,N}(X;\theta)$ and $1/N$ do not appear in the RNN (1.2). This changes the analysis: the RNN hidden layer (1.2) is not an Euler approximation to an ODE.

Although (1.2) is not a standard mean-field equation, we can observe mean-field behaviour for the distribution of the hidden layer. This is illustrated by Figure 1 where, for varying N, we simulated paths of the

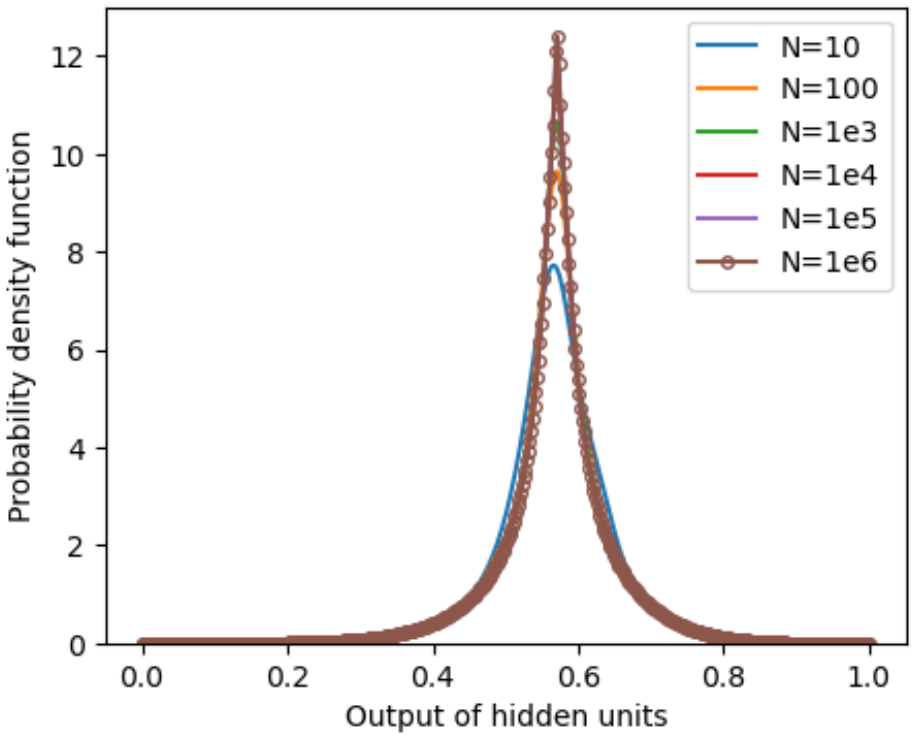

Figure 1: Each curve represents the *overall* empirical distributions of the untrained hidden units in the memory states (the *hidden memory units*) from all simulation instances $\ell = 1, \ldots, 100$ for $N = 10^2, ...10^6$ and time step $k \approx 50000$.

hidden layer $(S_k^{i,N}(X;\theta))_{i=1}^N$, based upon a common θ and independent paths of the input sequence X. The empirical distributions of the hidden units in the memory layer at a large, fixed time step k, is displayed as $N \to +\infty$. The empirical distributions converge as $N \to \infty$. Figure 1 suggests that a mean-field limit as $N \to \infty$ does exist. Further details of the simulation are provided in subsection 3.3.

Furthermore, numerical simulations suggest that the hidden layer is ergodic as the time steps $k \to \infty$. Figure 2 displays the time-averaged first and second moments of the hidden layer. The formal definition of the time averages is provided in subsection 3.3. Figures 1 and 2 together motivate an analysis of the RNN (1.2) as both $k, N \to \infty$.

Our mean-field analysis studies an appropriate fixed point for the untrained hidden units in the memory states, and use it to study the evolution of the RNN. The case for a constant input sequence $X_k = x$ (with a finite number of hidden units) has been study in [10, 11, 12] for developing a more efficient gradient-descent-type algorithm. However, the fixed point analysis is more complex when the input sequence X_k is non-constant and random. As a result, we would expect the RNN to have a *random* fixed point and that the *distribution* of the untrained hidden unit in the memory states should converge *weakly*. The random fixed point should also depend on the distribution of the data sequence X_k.

We are able to prove such random fixed point exists if the data sequence X_k is ergodic. Section 3 shows that the ergodicity of X actually leads to the ergodicity of $S_k^{i,N}(X;\theta)$.

The fixed point analysis is further complicated since the parameters θ will be simultaneously trained using the truncated backpropagation through time (tBPTT) algorithm as the hidden layer is updated at each time k; see Section 2. Both the hidden layers $S_k^N(X;\theta_k)$ and the parameters θ_k (which govern the transition function for the hidden layer S_k^N) will jointly evolve in time. Therefore, the dynamics of the RNN

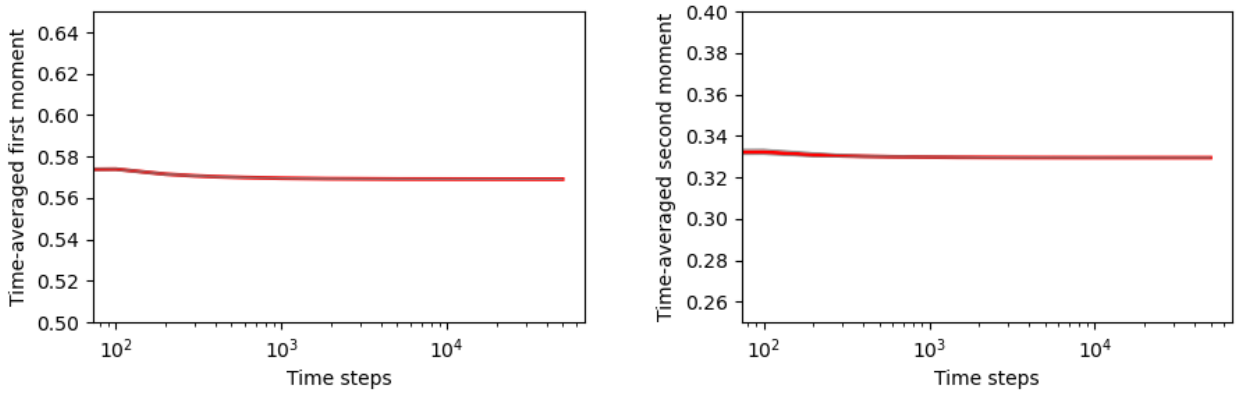

Figure 2: The plots of the time-averaged first and second moments of the hidden units for a sufficiently large N (chosen to be 10^6) and $p = 1, 2$. The x-axis represents the number of time steps. We summarise the minimum/maximum of the simulated first and second moments of the time-averages for independent input sequences X using a (seemingly invisible) grey band. The red line represents the mean of the time-averaged moments for all input sequences X, thus providing a Monte-Carlo estimate for the moments of the *random* fixed point. The fact that the realisations of the time averages all converge as $k \to \infty$ illustrates the ergodicity of the sequence $S_k^{i,N}(X;\theta)$.

will be changing over time. Fortunately, the changes due to the parameter updates will be of magnitude $\mathcal{O}(1/N)$. Therefore, the evolution of the output layer of the network ($g_k^N(X;\theta) = \frac{1}{N^{\beta_2}} \sum_{i=1}^N C^i S_{k+1}^{i,N}(X;\theta)$) can be represented as an Euler approximation of an appropriate infinite-dimensional ODE whose dynamics are a function of the RNN hidden layer's random fixed point, which it will converge to as $N \to \infty$. We emphasise that the evolution of the RNN hidden layer itself (i.e., $S_{k+1}^{i,N}(X;\theta)$) cannot be represented as a discretization of an ODE/PDE since the RNN hidden layer updates are $\mathcal{O}(1)$.

The RNN hidden layer is studied as a function in a Sobolev space, whose evolution is governed by the data sequence X_k, the parameter updates, and its dependence on the RNN hidden layer at the previous time step. Due to the correlation between updates (including both the hidden layer and data samples), a Poisson equation must be used to bound the fluctuations of the RNN around its limit equation.

The rest of the paper is organized as follows. In Section 2 we present the model that we study and our main assumptions. In Section 3 we characterise the dynamics of the RNN memory layer as a sequence of random functions of the parameters. We prove that the sequence of memory layers of the trained network is geometrically ergodic. In Section 4 we prove the convergence of the trained RNN. Our main result is presented in Theorem 4.2. Section 5 uses our limit result to prove a useful guarantee for the training of RNN models: the RNN training algorithm will asymptotically decrease the loss function (i.e., the RNN model will be updated in a descent direction). In Section 6 we present a number of technical lemmas and preliminary estimates that are used in the detailed proof of Theorem 4.2 which is presented in Section 7. In the Appendix A we present a recursive inequality that is used in various places throughout the paper as well as the construction of a clipping function.

2 Assumptions, Data, and Model Architecture

2.1 Data generation

Our paper focus on the problem of recovering the map from an input data sequence to an output data sequence. We assume the input data sequence $X = (X_k)_{k\geq 0}, X_k \in \mathbb{R}^d$ and output data sequence $Y = (Y_k)_{k\geq 0}, Y_k \in \mathbb{R}$ are jointly governed by the following update equation

$$(X_{k+1}, Z_{k+1}) = g(X_k, Z_k) + \epsilon_k, \tag{2.1}$$

$$Y_k = f(X_k, Z_k) + \eta_k, \tag{2.2}$$

where $(\epsilon_k, \eta_k)_{k\geq 0}, \epsilon_k \in \mathbb{R}^{d+1}, \eta_k \in \mathbb{R}$ are independent, identically distributed (iid) and mean-zero noises (but ϵ_k doesn't need to be independent from η_k). The map $g : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ drives the dynamics of the input sequence, and $f : \mathbb{R}^{d+1} \to \mathbb{R}$ maps the input sequence to the output sequence. The variables $Z_k \in \mathbb{R}$ are *hidden*, i.e. only the inputs $(X_k)_{k\geq 0}$ and outputs $(Y_k)_{k\geq 0}$ are observed, and it is not guaranteed whether the observation $(X_k, Y_k)_{k\geq 0}$ is actually a Markov process.

Let us make the following assumption on the functions f, g, as well as the background noise of the dynamics. As we will see in the next section, these assumptions are essential in ensuring that the dynamics of the input chain is ergodic.

Assumption 2.1 (On governing dynamics and background noises of the data sequences)**.**

1. We assume that the function (f, g) is L-globally Lipschitz (with respect to the Euclidean norm) with $L < 1$. That is for any $(x, z), (x', z') \in \mathbb{R}^{d+1}$ we have
$$\left| \begin{pmatrix} g(x,z) - g(x',z') \\ f(x,z) - f(x',z') \end{pmatrix} \right| \leq L \left| (x,z) - (x',z') \right|.$$
2. We assume that both $(\epsilon_k)_{k\geq 0}$ and $(\eta_k)_{k\geq 0}$ are *mean-zero* iid random variables that are uniformly bounded by some constant C_ϵ.
3. We finally assume that the sequence $(X_k, Z_k)_{k\geq 0}$ is bounded with $|(X_k, Z_k)| \leq 1$. This could be achieved by e.g. assuming
$$\max\left(\sup_{x,z} |g(x,z)|, \sup_k |\epsilon_k| \right) \leq \frac{1}{2}.$$
As a result, the output $(Y_k)_{k\geq 0}$ is uniformly bounded by some constant $C_y < +\infty$.

2.2 Recurrent Neural Network

A recurrent neural network is used to approximate the function g as defined above. Specifically, we will study the standard recurrent network in equation (1.2) but with the following simplifying assumption for its weight matrix:

Assumption 2.2 (Simplifying assumption of memory weights)**.** We assume all columns of B are equal, i.e. for all j we have $B^{ij} = B^j$ for some B^j.

Assumption 2.3 (Regularity of activation function)**.** We assume that $\sigma \in C_b^2(\mathbb{R})$ (being a twice-continuously differentiable), with output bounded by one and first and second derivatives both uniformly bounded by C_σ where $C_\sigma^2 < \min\{\frac{1}{2}, \frac{1-L^2}{8}\}$.

Assumption 2.3 on C_σ is instrumental in our proof of ergodicity of the memory states. One of the crucial results is Lemma 3.2 and the related results on erogodicity are in Section 3.1.

Example 2.4. An example of such activation function is the usual sigmoid function, defined as
$$z \in \mathbb{R} \mapsto \sigma(z) = \frac{1}{1+e^{-z}} \implies \sup_{z\in\mathbb{R}} |\sigma(z)| \leq 1.$$
One could show (see e.g. [13]) that $\sigma'(z) \in [0, 1/4]$ and $|\sigma''(z)| \leq 1/4$ for any $z \in \mathbb{R}$. As a result, the sigmoid function satisfies our assumption when $L < 1/2$.

We emphasise that the above assumptions are included for the simplicity of our proof, and we conjecture from our simulation that the RNN exhibits the desired mean-field behaviour under weaker assumptions.

Assumption 2.5 (Initialisation of parameters)**.** We assume that the trained parameters (C^i, W^i, B^i) are initialised independently from the data sequence, so that the sequence $(C_0^i, W_0^i, B_0^i)_{i=1}^N$ is composed of independent and identically distributed random vectors with a distribution $(C_0^i, W_0^i, B_0^i) \sim \lambda_0 := \lambda(dc, dw, db)$, such that C_0^i and W_0^i are independent from each other with C_0^i having zero mean. We further assume that the measure $\lambda(dc, dw, db)$ is absolutely continuous with respect to Lebesgue measure, i.e., it has a density. In addition, we assume that $|C_0^i| \leq 1$, $|B_0^i| \leq 1$ and $\mathbb{E}\left\|W_0^i\right\|^2 \leq 1$.

Remark 2.6. For simplicity of notation, we define λ_0^N to be the empirical measure of the trained parameters at initialisation on the space $\mathbb{R}^{1+d+1}$, i.e.

$$\lambda_0^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{C_0^i, W_0^i, B_0^i}.$$

Finally, in our notation later in the paper, the dependence of the RNN hidden layer $S_k^{i,N}(X;\theta)$ on X may be dropped in the later sections.

2.3 Training the RNN parameters

The parameters θ are in practice trained with an online SGD algorithm seeking to minimize the function

$$\mathcal{L}(\theta) = \frac{1}{T}\sum_{k=1}^{T} \mathcal{L}_k(\theta), \quad \text{where } \mathcal{L}_k(\theta) = \frac{1}{2}(g_k^N(X;\theta) - Y_k)^2.$$

The computational cost of evaluating the RNN network up to step T and performing one full evaluation of gradient by back-propagation through time (BPTT) grows in $\mathcal{O}(T)$, which become intractable if we want to perform gradient-based algorithms for large T for minimising the loss function. We instead update the parameters through the *online stochastic gradient descent with truncated back-propagation through time* (online SGD with tBPTT) [14] [15]. A detailed explanation is given in the next section, but the main idea is to truncate the computational graph up to τ time steps for $\tau \ll T$ when computing the outputs of the RNN and gradients with respect to the parameters. For simplicity and without loss of generality, we restrict our discussion to the case when $\tau = 1$.

Assume that the network (estimated) outputs at each step is $\hat{Y}_k^N$, which are used as a comparison with the actual observations Y_k for the computation of loss function and the estimated gradient. As it turns out, it is challenging to prove the boundedness of $\hat{Y}_k^N$. To resolve this issue, we will clip the gradients used in the parameter updates. Gradient clipping is a standard method in deep learning and training RNNs [13] [16] [17]. The gradient clipping will actually disappear in the final approximation argument as the number of hidden units $\to \infty$. Once the gradients in the parameter updates are clipped, the output $\hat{Y}_k^N$ can be proven to be bounded. We use the following clipping function:

Definition 2.7 (Smooth clipping function [18])**.** A family of functions $(\psi^N)_{N\in\mathbb{N}}$, $\psi^N \in C_b^2(\mathbb{R})$ is a family of *smooth clipping function* with parameter γ if the following are satisfied:

1. $\left|\psi^N(x)\right|$ is bounded by $2N^\gamma$,
2. $\psi^N(x) = x$ for $x \in [-N^\gamma, N^\gamma]$,
3. $\left|\dfrac{d}{dx}\psi^N(x)\right| \le 1$.

A construction of such clipping function is provided in appendix B following the arguments in [19]. By fundamental theorem of calculus we also know that $|\psi^N(x)| \le |x|$. The parameter γ is to be chosen in the next section, and it should be small to ensure that $|\psi^N(\hat{Y}_k^N)|$ does not blow up too quickly as $N \to \infty$.

Assumptions for the wide-network limit In our paper we set $\beta_1 = 1$. Depending on the choice of β_2, we will get different limits. Preliminary works for feedforward NN (including our prior work), [7, 20, 21, 22, 23] demonstrate that when $N \to +\infty$, the evolution of the output of the feedforward NN converges to a limit equation:

1. for $\beta_2 = 1$, the limit equation is expected to be a PDE,
2. for $\beta_2 = \beta \in (1/2, 1)$, the limit equation is expected to be an infinite-dimensional ODE, and
3. for $\beta_2 = 1/2$, the limit equation is expected to be a *random* infinite-dimensional ODE.

Algorithm 1 Online SGD with tBPTT ($\tau = 1$)

procedure SGDtBPTT(N, λ, T) ▷ network size, initial parameters distribution, running time
  Initalise: initial parameters $\theta = (C_0, W_0, B_0) \sim \lambda$, initial memories $\forall i, \hat{S}_0^{i,N} = 0$, step $k = 0$
  while $k \leq NT$ **do**
    for all $i \in \{1, 2, ..., N\}$ **do** ▷ Truncated forward propagation
      $\hat{S}_{k+1}^{i,N} \leftarrow \sigma\left((W_k^i)^\top X_k + \frac{1}{N}\sum_{j=1}^{N} B_k^j \hat{S}_k^{j,N}\right)$ ▷ Updating memory
    end for
    $\hat{Y}_k^N \leftarrow \frac{1}{N^\beta}\sum_{i=1}^{N} C_k^i \hat{S}_{k+1}^{i,N}$ ▷ Updating output
    $\hat{L}_k = \frac{1}{2}(\hat{Y}_k^N - Y_k)^2$ ▷ Computing loss
    for all $i \in \{1, 2, ..., N\}$ **do** ▷ Truncated backward propagation on $\hat{L}_k$
      $\Delta\hat{S}_{k+1}^{i,N} \leftarrow \sigma'\left((W_k^i)^\top X_k + \frac{1}{N}\sum_{j=1}^{N} B_k^j \hat{S}_k^{i,N}\right)$
      $C_{k+1}^i = C_k^i - \frac{\alpha}{N^{2-\beta}}(\psi^N(\hat{Y}_k^N) - Y_k)\hat{S}_{k+1}^{j,N}$
      $W_{k+1}^i = W_k^i - \frac{\alpha C_k^i}{N^{2-\beta}}(\psi^N(\hat{Y}_k^N) - Y_k)\Delta\hat{S}_{k+1}^{i,N}$
      $B_{k+1}^i = B_k^i - \frac{\alpha}{N^{3-\beta}}(\psi^N(\hat{Y}_k^N) - Y_k)\sum_{\ell=1}^{N} C_k^\ell \Delta\hat{S}_{k+1}^{\ell,N}$
    end for
  end while
end procedure

Our paper will focus on the case when $\beta_2 = \beta \in (1/2, 1)$. We are aware that some analyses have been done for the case when $\beta_2 = 1$ in [24], with the RNN trained *offline* by continuous gradient descent after observing a fixed number of steps of the sequence $(X_k)_{k\geq 0}$. We emphasise that in the present work the RNN is trained *online*, so that we update the parameters every time we observe a new step of our sequences (X_k, Y_k). We study the asymptotics of the training of RNN as the training time (and hence number of observations made for the input and output sequences) grow with the width of the hidden layer, i.e. $N \to +\infty$.

In order to derive a well-defined typical behaviour of the neural network when training our RNN in the limit as $N \to +\infty$, we choose the learning rate to be

$$\alpha^N := \frac{\alpha}{N^{2-2\beta}}, \tag{2.3}$$

for some constant $\alpha > 0$. The learning rate is constant in time, and with this we may explicitly state the online SGD with tBPTT, see Algorithm 1.

Lemma 2.8 yields a bound (uniform in i) for how much the parameters move.

Lemma 2.8. *Fix $T > 0$. If we choose $\gamma > 0$ such that $\beta + 2\gamma < 1$, then for all k such that $k/N \leq T$ we have the following sure bounds:*

$$|C_k^i - C_0^i| + \left\|W_k^i - W_0^i\right\| + |B_k^i - B_0^i| < \frac{C_T}{N^{1-\beta-\gamma}}, \tag{2.4}$$

where $C_T > 0$ is a constant that depends on T.

Proof. See Section 6.1. □

Remark 2.9. One can also show, with the absence of clipping function ψ^N, that

$$|C_{k+1}^i - C_k^i| + \left\|W_{k+1}^i - W_k^i\right\| + |B_{k+1}^i - B_k^i| \leq \frac{C_T}{N}, \tag{2.5}$$

and hence by a telescoping sum argument,

$$|C_k^i - C_0^i| + \left\|W_k^i - W_0^i\right\| + |B_k^i - B_0^i| \leq C_T.$$

The proof of this remark is also very similar - see section 6.1. However, this does not reflect our intuition that the trained parameters are closed to the initialisation. The clipping function here is crucial in justifying the linearisation of the sample outputs with respect to the parameters in our analysis.

Define the empirical measure of (C_k^i, W_k^i, B_k^i) as λ_k^N, i.e.

$$\lambda_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{C_k^i, W_k^i, B_k^i}.$$

Lemma 2.8 says that the parameters at step k should be close to the initial parameters on average, so λ_k^N should be well-approximated by λ^N in some sense. We therefore expect that for $k \leq \lfloor TN \rfloor$, λ_k^N converges to λ (distribution of initialisation) in a sense to be specified. However, establishing such convergence using our traditional way of studying its limiting ODE (see e.g. [20]) is difficult. To that end, the evolution of

$$\left\langle f, \lambda_k^N \right\rangle = \int_{\mathbb{R}^{1+d+1}} f(c,w,b)\, \lambda_k^N(dc, dw, db) = \frac{1}{N}\sum_{i=1}^N f(C_k^i, W_k^i, B_k^i),$$

i.e. the inner-product of the empirical distribution with a smooth test function $f \in C^\infty(\mathbb{R}^{1+d+1})$ does not look like a discretization of a differential equation. A new mathematical approach is therefore required to analyze the infinite-width RNN.

Our result is divided into three parts: a result regarding the dynamics of sampled memories through the network, the limiting ODE of the process $\hat{Y}_k^N$, and how one could gain insight on how the tBPTT decreases the loss function by studying the limiting ODE.

3 Dynamics of the memory units

We characterise the dynamics of the memories as a sequence of a random function with respect to the weights. Specifically, we have $\hat{S}_{k+1}^{i,N} = v_{k+1}^N(W_k^i)$, where $v_{k+1}^N(w)$ satisfies the following recursion:

$$v_{k+1}^N(w) = \sigma\left(w^\top X_k + \frac{1}{N}\sum_{i=1}^N B_k^i v_k^N(W_{k-1}^i)\right), \quad v_0^N(w) = 0. \tag{3.1}$$

We will show that the sequence of random functions $\left\{v_k^N(\cdot)\right\}_{k\geq 0}$ converges in distribution to some stationary distribution over a specific function space, as a result of the ergodicity of the underlying input and output dynamics, and the parameters do not move too far from their initial position. Taking these observations into account, we study the joint process $V_k^N = (X_k, Z_k, Y_k, v_k^N(\cdot)) \sim \nu_k^N$ as an element of the product space

$$\mathcal{X} = \overline{B_1(0)} \times [-C_y, C_y] \times \overline{\mathcal{H}},$$

where

$$\overline{B_1(0)} = \{(x,z)\,|\,|(x,z)| \leq 1\}, \tag{3.2}$$

$\mathcal{H}$ is the set

$$\mathcal{H} = \left\{f(w) = \sigma(w^\top a + b) \,:\, a \in \mathbb{R}^d,\, |a| \leq 1,\, b \in \mathbb{R}\right\},$$

and $\overline{\mathcal{H}}$ is the closure of $\mathcal{H}$ with the Sobolev norm (with respect to the distribution of initialisation of w (i.e. λ)

$$\|h\|^2_{H^1(\lambda)} = \int_{\mathbb{R}^d} \left(|h(w)|^2 + |\nabla h(w)|^2\right) \lambda(dw). \tag{3.3}$$

We equip the space $\mathcal{X}$ with the norm $\|\cdot\|_{\mathcal{X}}$:

$$\|(x,z,y,h)\|^2_{\mathcal{X}} = |x|^2 + |z|^2 + |y|^2 + \|h\|^2_{H^1(\lambda)}, \tag{3.4}$$

and the usual Borel σ-algebra $\mathcal{B}(\mathcal{X})$ that contains the open subsets of $\mathcal{X}$ with respect to this norm.

Lemma 3.1. *An element $v \in \mathcal{H}$ satisfies*

$$\sup_w |v(w)| \leq 1, \quad \sup_w |\nabla v(w)| \leq C_\sigma < 1, \quad \max_{i,j} \sup_w [|\partial_{ij} v(w)|] \leq C_\sigma < 1.$$

As a result, we know that $v \in H^1(\mathbb{R}^d;\lambda) =: H^1(\lambda)$, i.e. $\mathcal{H}$ is a (proper) subspace of $H^1(\lambda)$.

Proof. The first inequality is immediately true by Assumption 2.3. The second inequality is true as

$$\nabla v(w) = \sigma'(w^\top a + b)a \implies |\nabla v(w)| \leq |\sigma'(w^\top a + b)||a| \leq C_\sigma.$$

The second inequality can also be proven by noting that

$$\mathsf{Hess}\, v(w) = \sigma''(w^\top a + b)aa^\top \implies |\partial_{ij} v(w)| \leq |\sigma''(w^\top a + b)||a_i||a_j| \leq C_\sigma.$$

Notice that this implies that $v \in H^1(\lambda)$, as the $\|v\|_{H^1(\lambda)}$ norm is well-defined and bounded by the following

$$\begin{aligned}\|v\|^2_{H^1(\lambda)} \leq \int_{\mathbb{R}^d} \left(|v(w)|^2 + |\nabla v(w)|^2\right) \lambda(dw) &= \int_{\mathbb{R}^d} \left[\left|\sigma(w^\top a + b)\right|^2 + \left|\sigma'(w^\top a + b)\right|^2 |a|^2\right] \lambda(dw) \\ &\leq \int_{\mathbb{R}^d} \left(1 + C_\sigma^2\right) \lambda(dw) = 1 + C_\sigma^2 < +\infty. \end{aligned} \tag{3.5}$$

In fact this yields the control $\|v\|_{H^1(\lambda)} \leq \sqrt{1 + C_\sigma^2}$, so $\mathcal{H}$ is bounded in $H^1(\lambda)$-norm. □

As a result, the closure $\overline{\mathcal{H}}$ is a proper closed subspace of the Polish space $\mathbb{R}^{d+1+1} \times H^1(\lambda)$ in the $\mathcal{X}$-norm, hence it is a complete normed space as well.

As the parameters do not move too much from their initial position, we would expect λ_k^N to be close to λ^N in some sense when $k \leq \lfloor NT \rfloor$. We also expect λ^N to be close to λ when N gets large. This motivates us to consider the following auxiliary processes:

$$H_k = (X_k, Z_k, Y_k, h_k(\cdot)) \sim \mu_k,$$

where h_k is governed by the following random iterations:

$$h_{k+1}(w) = \sigma(w^\top X_k + \langle b' h_k(w'), \lambda\rangle) \in \mathcal{H}, \quad h_0(w) = 0. \tag{3.6}$$

We will quantify the distance between $v_k^N(w)$ and $h_k(w)$ in the later part of this section, but let us focus on this auxiliary process. The process is a Markov chain of the form $H_{k+1} = F(H_k, \epsilon_k, \eta_k)$, where

$$\begin{aligned} F : (x,z,y,h,\epsilon,\eta) &\mapsto (g(x,z) + \epsilon, f(x,z) + \eta, \varsigma(x,z,y,h)) \in \mathcal{X} \\ \varsigma : (x,z,y,h) &\mapsto \left[w \mapsto \sigma(w^\top x + \langle bh(w), \lambda\rangle)\right] \in \mathcal{H} \end{aligned} \tag{3.7}$$

As a result, the Markov chain possesses the following transition kernel $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$, where:

$$P(H, A) = \mathbb{E}_{\epsilon,\eta}(\mathbb{I}_A(F(H,\epsilon,\eta))).$$

We make note that

Lemma 3.2. *For any fixed realisation of (ϵ, η), the function F is q_0-Lipschitz continuous with $q_0 \leq \sqrt{L^2 + 8C_\sigma^2} < 1$. Due to Assumptions 2.1 and 2.3 on L and C_σ respectively , we have $q_0 < 1$.*

Proof. See section 6.2. □

This enables us to show that the Markov chain $(H_k)_{k\geq 1}$ is ergodic, which is the objective of subsection 3.1.

3.1 Geometric Ergodicity of the auxiliary process

Given a norm $\|\cdot\|$ on a vector space $\mathcal{X}$ such that $(\mathcal{X}, \|\cdot\|)$ is Banach (i.e. complete) and separable, we can define the associated p-Wasserstein distance for $p \geq 1$ [25]:

Definition 3.3 (Wasserstein Metric)**.** Let ρ_1, ρ_2 are measures in the p-Wasserstein space of $(\mathcal{X}, \|\cdot\|)$:

$$\rho_1, \rho_2 \in \mathcal{P}_p^{\|\cdot\|}(\mathcal{X}) = \left\{ \rho \,:\, \int_{\mathcal{X}} \|u\|_{\mathcal{X}}^p \; \rho(dx) < +\infty \right\}$$

A measure ρ on $\mathcal{X} \times \mathcal{X}$ is a *coupling* between ρ_1 and ρ_2 if

$$\rho_1(\cdot) = \rho(\cdot \times \mathcal{X}) \quad \rho_2(\cdot) = \rho(\mathcal{X} \times \cdot).$$

The p-Wasserstein distance between ρ_1 and ρ_2 with respect to norm d is defined as $\mathsf{Wass}_p^{\|\cdot\|}(\rho_1, \rho_2)$ that satisfies

$$\left[\mathsf{Wass}_p^{\|\cdot\|}(\rho_1, \rho_2)\right]^p = \inf_{\gamma \in \Gamma(\rho_1, \rho_2)} \int_{\mathcal{X}\times\mathcal{X}} \|u - v\|_{\mathcal{X}}^p \; \gamma(du\, dv), \tag{3.8}$$

where $\Gamma(\rho_1, \rho_2)$ is the set of all couplings between ρ_1 and ρ_2.

The norm $\|\cdot\|$ might be omitted in writing $\mathsf{Wass}_p^{\|\cdot\|}(\rho_1, \rho_2)$ and $\mathcal{P}_p^{\|\cdot\|}(\mathcal{X})$ if no confusion is created as a result.

Remark 3.4. Note that the infimum in (3.8) is a minimum [25], i.e. there is an optimal coupling $\gamma^* \in \Gamma(\rho_1, \rho_2)$ such that

$$\left[\mathsf{Wass}_p^{\|\cdot\|}(\rho_1, \rho_2)\right]^p = \int_{\mathcal{X}\times\mathcal{X}} \|u - v\|_{\mathcal{X}}^p \; \gamma^*(du\, dv). \tag{3.9}$$

This fact will be used to prove bounds for the Wasserstein metric between successive terms of sequence of distribution μ_k.

Recall that the transition kernel P induces an operator on the space of all probability measure on $\mathcal{X}$:

$$\mu \overset{P^\vee}{\mapsto} [P^\vee(\mu)](\cdot) = \int_{\mathcal{X}} P(\mathsf{H}, \cdot)\, \mu(d\mathsf{H}).$$

In particular, the sequence of measures $(\mu_k)_{k\geq 0}$ could now be written as $\mu_k = (P^\vee)^k \mu_0$. Thanks to the q_0-Lipschitzness of the function F, we could show the following control:

Lemma 3.5.

$$\mathsf{Wass}_2(P(H, \cdot), P(H', \cdot)) \leq q_0 \left\| H - H' \right\|_{\mathcal{X}}.$$

Proof. We consider the joint random variable $\tilde{F}_{H,H'} = (F(H, \epsilon, \eta), F(H', \epsilon, \eta))$. We note that the first component has the same distribution as $P(H, \cdot)$, and the second component has the same distribution as $P(H', \cdot)$. Therefore the distribution of $\tilde{F}_{H,H'}$, denoted as $\tilde{\gamma}_{H,H'}$, is both a coupling between the measures $P(H, \cdot)$ and $P(H', \cdot)$ and a valid ((H, H')-measurable) transition kernel. Therefore, by Lemma 3.2 we have

$$\mathsf{Wass}_2(P(H, \cdot), P(H', \cdot)) \leq \left[\mathbb{E}_{F_{H,H'}} |F(H, \epsilon, \eta) - F(H', \epsilon, \eta)|^2\right]^{1/2} \leq q_0 \|H - H'\|_{\mathcal{X}}. \tag{3.10}$$

□

This leads to the following formal result on the convergence of μ_k.

Theorem 3.6. *For all $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$,*

$$\mathsf{Wass}_2(P^\vee(\mu), P^\vee(\nu)) \le q_0 \mathsf{Wass}_2(\mu, \nu),$$

where $q_0 < 1$, and the 2-Wasserstein distance is taken with respect to the metric induced by the $\|\cdot\|_{\mathcal{X}}$ norm defined in (3.4)*. As a result the operator $P^\vee$ has a unique contractive fixed point μ, which is the stationary measure of the Markov chain $(H_k)_{k\ge 0}$.*

Proof. Let γ^* be the optimal coupling between μ and ν, such that

$$(\mathsf{Wass}_2(\mu, \nu))^2 = \int_{\mathcal{X}} \|\mathsf{H} - \tilde{\mathsf{H}}|^2_{\mathcal{X}} \, \gamma^*(d\mathsf{H}, d\tilde{\mathsf{H}}). \tag{3.11}$$

We look at the coupling

$$\gamma(\cdot) = \int_{\mathcal{X}\times\mathcal{X}} \tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(\cdot) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}).$$

The integral is well defined as $\tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(\cdot)(A)$ is a $(\mathsf{H}, \bar{\mathsf{H}})$-measurable function for any sets $A \in \mathcal{B}(\mathcal{X}\times\mathcal{X})$. One can directly check that γ is a measure. In particular, γ is a coupling of $P^\vee(\mu)$ and $P^\vee(\nu)$, as for all $A, B \in \mathcal{B}(\mathcal{X})$

$$\begin{aligned}
\gamma(A\times\mathcal{X}) &= \int_{\mathcal{X}\times\mathcal{X}} \tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(A\times\mathcal{X}) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}) = \int_{\mathcal{X}\times\mathcal{X}} P(\mathsf{H}, A) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}) = [P^\vee(\mu)](A), \\
\gamma(\mathcal{X}\times B) &= \int_{\mathcal{X}\times\mathcal{X}} \tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(\mathcal{X}\times B) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}) = \int_{\mathcal{X}\times\mathcal{X}} P(\bar{\mathsf{H}}, B) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}) = [P^\vee(\nu)](B).
\end{aligned} \tag{3.12}$$

We therefore have $\gamma \in \Gamma(P^\vee(\mu), P^\vee(\nu))$. Finally, since

$$\gamma(A\times B) = \int_{\mathcal{X}\times\mathcal{X}} \tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(A\times B) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}) = \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}\times\mathcal{X}} \mathbb{I}_{A\times B}(H, \bar{H}) \, \tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(dH, d\bar{H}) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}})$$

one could extend the above formula to define an integral of $(H, \bar{H})$-measurable functions $c(h, \bar{h})$ with respect to the coupling γ:

$$\int_{\mathcal{X}\times\mathcal{X}} c(h, \bar{h}) \, \gamma(dh, d\bar{h}) = \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}\times\mathcal{X}} c(H, \bar{H}) \, \tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(dH, d\bar{H}) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}),$$

so we could use this formula with $c(h, \bar{h}) = \|h - \bar{h}\|^2_{\mathcal{X}}$ to conclude that

$$\begin{aligned}
[\mathsf{Wass}_2(P^\vee(\mu), P^\vee(\nu))]^2 &\le \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}\times\mathcal{X}} \|H - \bar{H}\|^2_{\mathcal{X}} \, \tilde{\gamma}_{\mathsf{H},\bar{\mathsf{H}}}(dH, d\bar{H}) \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}), \\
&\overset{(3.10)}{\le} \int_{\mathcal{X}\times\mathcal{X}} q_0^2 \|\mathsf{H} - \bar{\mathsf{H}}\|^2_{\mathcal{X}} \, \gamma^*(d\mathsf{H}, d\bar{\mathsf{H}}) \\
&\overset{(3.11)}{=} q_0^2 [\mathsf{Wass}_2(\mu, \nu)]^2,
\end{aligned} \tag{3.13}$$

Therefore $P^\vee$ is indeed a contraction in the space $\mathcal{P}_2(\mathcal{X})$. Since $\mathcal{X}$ is a complete space with respect to the $\|\cdot\|_{\mathcal{X}}$ norm, so is $\mathcal{P}_2(\mathcal{X})$ with respect to the Wass_2 metric [25], so by Banach fixed point theorem $P^\vee$ admits a unique fixed point μ, such that

$$P^\vee(\mu) = \mu.$$

By definition of P, we see that μ is the unique stationary measure of the Markov chain $(H_k)_{k\ge 0}$, so that for any initial distribution μ_0 and induced sequence $(\mu_k)_{k\ge 0}$ we have

$$\mathsf{Wass}_2(\mu_k, \mu) = q_0^k \mathsf{Wass}_2(\mu_0, \mu).$$

This completes the proof of Theorem 3.6. □

3.2 Bounds on auxiliary processes

We will introduce a further auxiliary sequence $H_k^N = (X_k, Z_k, Y_k, h_k^N(\cdot))$ for our proof, where the random sequence $\left(h_k^N(\cdot)\right)_{k\geq 0}$ evolves according to the equation

$$h_{k+1}^N(w) = \sigma(w^\top X_k + \left\langle b' h_k^N(w'), \lambda^N \right\rangle), \quad h_0^N(w) \equiv 0. \tag{3.14}$$

Remark 3.7. We could now express the hidden memory units in terms of the auxiliary processes. In fact,

- The trained hidden memory units $\hat{S}_k^{i,N} = v_k^N(W_{k-1}^i)$ for $k \geq 1$. The base case is justified as $\hat{S}_1^{i,N} = v_1^N(W_0^i) = \sigma((W_0^i)^\top X_1)$, and we could inductively prove that

$$\hat{S}_{k+1}^{i,N} = v_{k+1}^N(W_k^i) = \sigma\left((W_k^i)^\top X_k + \frac{1}{N}\sum_{i=1}^N B_k^i \hat{S}_k^{i,N}\right) = \sigma\left((W_k^i)^\top X_k + \frac{1}{N}\sum_{i=1}^N B_k^i v_k^N(W_{k-1}^i)\right),$$

- We also show that the untrained hidden memory units $S_k^{i,N}(X;\theta_0) = h_k^N(W^i)$ for $k \geq 1$ and $\theta_0 = (C^i, W^i, B^i)$. This is shown by noting that for $k = 1$, we have $S_1^{i,N}(X;\theta) = h_1^N(W^i) = \sigma((W^i)^\top X_0)$, and inductively,

$$\begin{aligned} S_{k+1}^{i,N}(X;\theta) &= \sigma\left((W^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B^j S_k^{j,N}(X;\theta)\right) \\ &= \sigma\left((W^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B^j h_k^N(W^j)\right) \\ &= \sigma\left((W^i)^\top X_k + \left\langle b' h_k^N(w'),\, \lambda^N \right\rangle\right) = h_{k+1}^N(W^i). \end{aligned}$$

Since we already know that the individual parameters C_k^i, W_k^i, B_k^i are close to their initial values C^i, W^i, B^i by Lemma 2.8, we should expect that the random functions $v_k^N(\cdot)$ to get close to the functions $h_k^N(\cdot)$ as $N \to +\infty$. In addition, we should also expect that $h_k^N(\cdot)$ to be close to the random functions $h_k(\cdot)$ as $N \to +\infty$ as well. This is useful in proving the weak convergence results in the next section.

Let us formalise the above heuristics by studying the difference

$$\Gamma_k^N(w) = v_k^N(w) - h_k(w) = \Gamma_k^{N,1}(w) + \Gamma_k^{N,2}(w),$$

where

$$\begin{aligned} \Gamma_k^{N,1}(w) &= h_k^N(w) - h_k(w), \\ \Gamma_k^{N,2}(w) &= v_k^N(w) - h_k^N(w). \end{aligned} \tag{3.15}$$

We will show the following

Proposition 3.8. *If $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$, we have for $k \leq \lfloor NT \rfloor$,*

$$\mathbb{E}\left\|\Gamma_k^{N,1}(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^{N,1}(W^i)\right]^2 \leq \frac{C}{N}, \tag{3.16}$$

for some constant $C > 0$, and

$$\mathbb{E}\left\|\Gamma_k^{N,2}(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^{N,2}(W^i)\right]^2 \leq \frac{C_T}{N^{2-2\beta-2\gamma}}, \tag{3.17}$$

where $C_T > 0$ is another constant depending on T. Therefore for $k \leq \lfloor NT \rfloor$,

$$\mathbb{E}\left\|\Gamma_k^N(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^N(W^i)\right]^2 \leq 4C_\sigma^2\left(\frac{C}{N} + \frac{C_T}{N^{2-2\beta-2\gamma}}\right) \leq \frac{C_T}{N^{2-2\beta-2\gamma}}. \tag{3.18}$$

Proof. See section 6.3. □

As an immediate result of the above proposition, we see that the empirical distribution of the untrained memory units converges. This gives a formal statement of the mean-field behaviour of the distribution of the untrained memory units.

3.3 Numerical simulations

To illustrate the above results, we simulated 100 paths of the untrained hidden memory units, denoted as $h_k^{N,\mathsf{path}}(W^i) = S_k^{i,N}(X^{\mathsf{path}};\theta)$, each based on the common parameters $\theta = (C^i, W^i, B^i)$ and an independent instance of the input sequence X^{path} for $\mathsf{path} = 1, 2, ..., 100$. The actual input sequence is simulated from the following 2D random dynamical systems:

$$\begin{bmatrix} X_{k+1} \\ Z_{k+1} \end{bmatrix} = \frac{1}{2}\tanh\left(P\begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}P^{-1}\begin{bmatrix} X_k \\ Z_k \end{bmatrix}\right), \quad P = \begin{bmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{bmatrix}, \; X_0, Z_0 \overset{\text{iid}}{\sim} \mathsf{Uniform}[0,1]. \tag{3.19}$$

We choose the standard sigmoid function as our activation function, and (B^i, W^i) are simulated so that $B^i \overset{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$ and $W^i \overset{\text{iid}}{\sim} \mathsf{Normal}[0,1]$.[1]

The mean-field behaviour could be studied by looking at the empirical distributions of the untrained hidden units for each set (represented by the grey lines in figure 3):

$$\nu_k^{N,\mathsf{path}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{h_k^{N,\mathsf{path}}(W^i)},$$

as well as the overall distributions of the hidden units from all sets as a Monte-carlo estimate of the expectation of the distribution of $h_k^N(W^i)$ (represented by the red lines in figure 3):

$$\nu_k^{N,\mathsf{overall}} = \frac{1}{100}\sum_{\mathsf{path}=1}^{100}\delta_{h_k^{N,\mathsf{path}}(W^i)}.$$

The convergence of distribution of $h_k^N(W^i)$ as $N \to \infty$ is illustrated by noting that the above plots are similar to each other for large N. Figure 1 in the introduction is plotted by stacking all overall empirical distributions $\nu_k^{N,\mathsf{overall}}$ on the same plot for easier comparison.

The intermediate proposition provides a solid foundation for our numerical experiment. Showing the ergodicity of $h_k^N(W^i) = S^{i,N}(X;\theta)$ is easier than showing the ergodicity of the function $h_k(\cdot)$ as the former does not involve the exact computation of $\langle b' h_k(w'), \lambda\rangle$. The above proposition implies that for sufficiently large N, the untrained memory units $h_k^N(W^i) = S^{i,N}(X;\theta)$ exhibit similar ergodic behaviour due to the ergodicity of h_k. To illustrate this, we compute the time average of the empirical first and second moments of the sample hidden memory units $(S_k^{i,N}(X;\theta))_{k\geq 0}$ for each sets, defined as

$$\mathsf{timeAvg}_T^{N,p,\mathsf{path}} = \frac{1}{NT}\sum_{k=0}^{T}\sum_{i=1}^{N}\left[h_k^{N,\mathsf{path}}(W^i)\right]^p, \quad p = 1, 2, \quad \mathsf{path} = 1, ..., 100,$$

as well as the overall time average of the empirical first and second moments:

$$\mathsf{timeAvg}_T^{N,p,\mathsf{overall}} = \frac{1}{100}\sum_{\mathsf{path}=1}^{100}\mathsf{timeAvg}_T^{N,p,\mathsf{path}}.$$

They are both plotted in Figure 4, with the redline being $\mathsf{timeAvg}_T^{N,p,\mathsf{overall}}$, and that all $\mathsf{timeAvg}_T^{N,p,\mathsf{path}}$ lies within the greyband. The shrinking of the grayband towards the red line is another evidence of mean-field

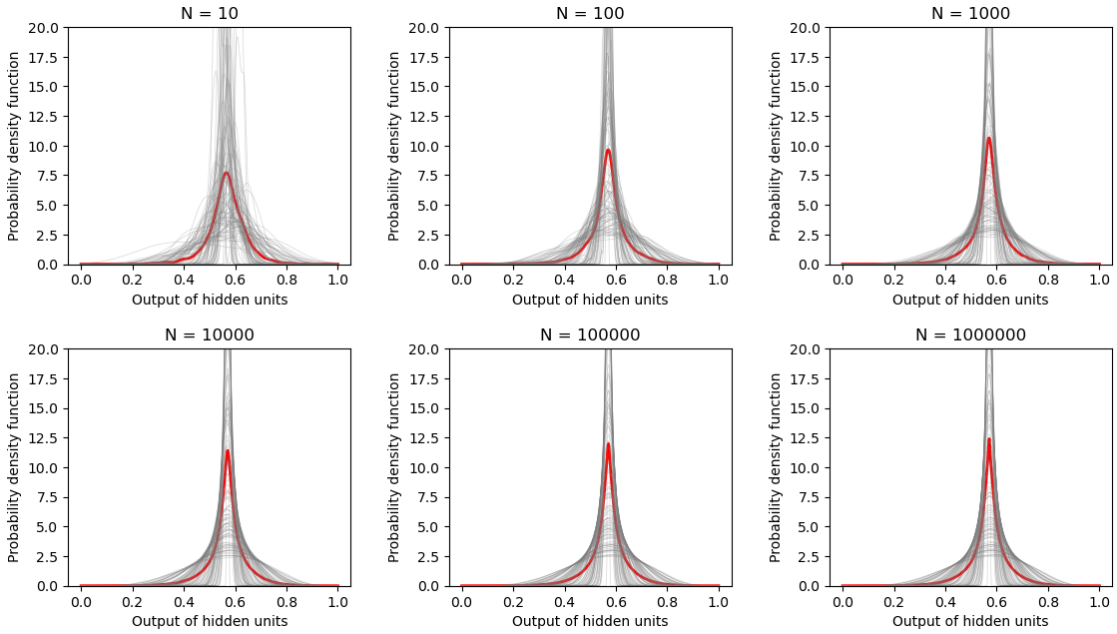

Figure 3: Empirical distributions of the untrained hidden memory units $h_k^N(W^i)$ for varying N and large time step $k \approx 50000$. The grey lines represents the empirical distribution for a *single* set of the untrained hidden memory units $\nu_k^{N,\mathsf{path}}$, and the red line represents the empirical distribution of all untrained hidden memory units from *all* sets $\nu_k^{N,\mathsf{overall}}$.

behaviour, and the convergence of $\mathsf{timeAvg}_T^{N,p,\mathsf{overall}}$ demonstrates the ergodicity of the untrained hidden memory units. Figure 2 in the introduction is the enlarged version of 4 for $N = 10^6$.

Please refer to `https://github.com/Samuel-CHLam/rnn_ergodicity` for the code of the simulation.

4 Dynamics of the RNN output

4.1 Pre-limit evolution

To study the dynamics of the sequence $\hat{Y}_k^N$, we first study the auxiliary sequence of functions

$$g_k^N : h \in \mathcal{H} \mapsto g^N(h; \theta_k) := \frac{1}{N^\beta} \sum_{i=1}^N C_k^i h(W_k^i) = N^{1-\beta} \left\langle h(w)c,\, \lambda_k^N \right\rangle,$$

Notice that $g_k^N(\cdot)$ is a sequence of linear functionals on $\mathcal{H}$ that evolves with the time step k. We further note that $\hat{S}_{k+1}^{i,N} = v_{k+1}^N(W_k^i)$, so

$$\hat{Y}_k^N = g_k^N(v_{k+1}^N(w)).$$

Using Taylor expansion, we derive the following evolution equation for $g_k^N(h)$ for a fixed test function h:

[1]The setting does not fully satisfy our strong assumptions, but it gives a good illustration of the general mean-field behaviour.

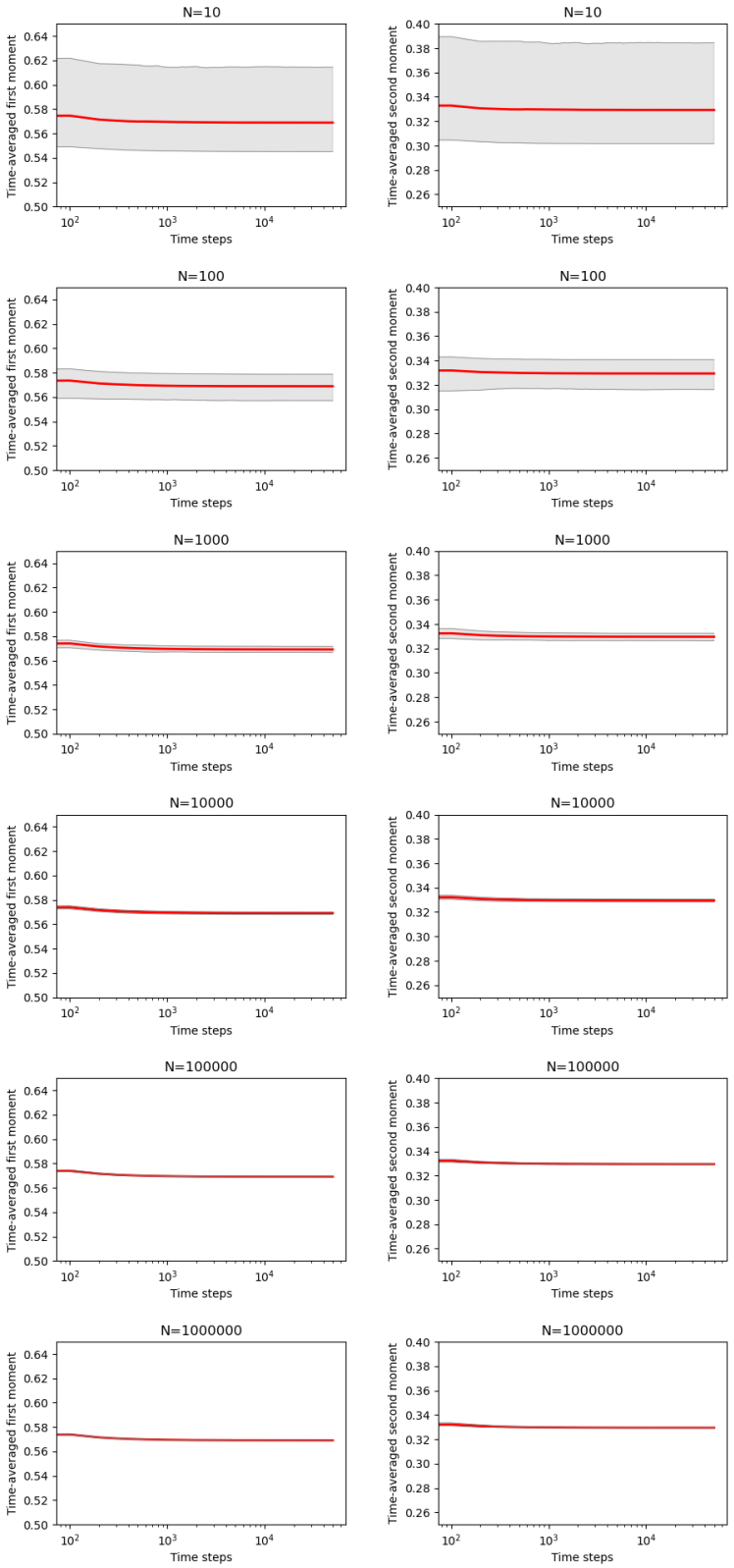

Figure 4: The plot of time averages for $N = 10^k, k = 2, 3, 4, 5, 6$ and $p = 1, 2$. The actual realisations of $\mathsf{timeAvg}_T^{N,p,\mathsf{path}}$ lie in the grey band, and the red line is the overall time average $\mathsf{timeAvg}_T^{N,p,\mathsf{overall}}$ as the Monte-carlo estimate of $\mathbb{E}[\mathsf{timeAvg}_T^{N,p}]$. The desired converging behaviour is only exhibited when N is sufficiently large.

Lemma 4.1. *For all $h \in \mathcal{H}$, let*

$$\begin{aligned}
\triangle g_k^N(h) &= g_{k+1}^N(h) - g_k^N(h), \\
\delta^{(1)} g_k^N(h) &= -\frac{\alpha}{N^2}(\psi^N(\hat{Y}_k^N) - Y_k) \sum_{i=1}^N \left[\hat{S}_{k+1}^{i,N} h(W_k^i) + (C_k^i)^2 \Delta \hat{S}_{k+1}^i \nabla h(W_k^i)^\top X_k \right], \\
&= -\frac{\alpha}{N}(\psi^N(\hat{Y}_k^N) - Y_k)\big(\langle v_{k+1}^N(w) h(w) + c^2 \sigma' \left(w \cdot X_k + \langle b v_k^N(w),\, \lambda_k^N \rangle \right) \nabla h(w)^\top X_k,\, \lambda_k^N \rangle \big). \quad (4.1)
\end{aligned}$$

Then for all $k \le \lfloor NT \rfloor$,

$$\left| \triangle g_k^N(h) - \delta^{(1)} g_k^N(h) \right| \le \frac{C_T}{N^{3-\beta-2\gamma}}.$$

Proof. See section 6.4. □

We will prove that the RNN converges to a continuous-time limit equation on $[0, T]$ as the number of hidden units $N \to \infty$ and where the number of time steps is $k = 0, 1, \ldots, \lfloor TN \rfloor$. The following time-rescaling is defined to embed (4.1) in continuous time: $g_t^N(h) = g_{\lfloor Nt \rfloor}^N(h; \theta_{\lfloor Nt \rfloor})$. Using a telescoping series applied to (4.1) and re-writing in terms of the empirical measure λ_k^N, one can conclude that

$$\begin{aligned}
\left| g_t^N(h) - g_0^N(h) - \sum_{k=0}^{\lfloor Nt \rfloor - 1} \delta^{(1)} g_k^N(h) \right| &= \left| \sum_{k=0}^{\lfloor Nt \rfloor - 1} (\triangle g_k^N(h) - \delta^{(1)} g_k^N(h)) \right| \\
&\le \sum_{k=0}^{\lfloor Nt \rfloor - 1} \left| \triangle g_k^N(h) - \delta^{(1)} g_k^N(h)) \right| \\
&\le Nt \times \frac{C_t}{N^{3-\beta-2\gamma}} = \frac{C_t}{N^{2-\beta-2\gamma}}. \quad (4.2)
\end{aligned}$$

The sum in (4.2) is an approximation to a Riemann time integral with error $\mathcal{O}(N^{-1})$, so one would expect that for $t \in [0, T]$, the evolution of $g_t^N(h)$ for fixed h would converge in distribution to a solution of the linear infinite-dimensional ODE equation. Formally, we have the following:

Theorem 4.2. *Let $T < \infty$ be given and $t \le T$. Define the infinite dimensional equation for $g_t \in (H^1(\lambda))^*$, the dual space of $H^1(\lambda)$, such that for the test function $h \in H^1(\lambda)$:*

$$\begin{aligned}
g_t(h) &= -\alpha \int_0^t \left[\int_{\mathcal{X}} (g_s(\varsigma(\mathsf{H})) - \mathsf{y})) \mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h})\, \mu(d\mathsf{H}) \right] ds, \quad g_0(h) = 0, \\
\mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h}) &= \left\langle \sigma(w^\top \mathsf{x} + \langle b' \mathsf{h}(w'), \lambda \rangle) h(w) + c^2 \sigma'(w^\top \mathsf{x} + \langle b' \mathsf{h}(w'), \lambda \rangle) \nabla h(w)^\top x,\, \lambda \right\rangle, \\
[\varsigma(\mathsf{H})](w) &= \sigma(w^\top \mathsf{x} + \langle b' \mathsf{h}(w'),\, \lambda \rangle) \quad (4.3)
\end{aligned}$$

where $\mathsf{H} = (\mathsf{x}, \mathsf{z}, \mathsf{y}, \mathsf{h})$, and μ is the stationary distribution of Markov chain $(H_k)_{k \ge 0}$ obtained in Theorem (3.6). Then, for $\gamma > 0$ sufficiently small, as $N \to \infty$ and $T < +\infty$, we have

$$\sup_{t \in [0,T]} \sup_{h \in \mathcal{H}} \mathbb{E}\left[\left| g_t^N(h) - g_t(h) \right| \right] \le \frac{C_T}{N^\epsilon},$$

where $\epsilon = (1 - \beta - 2\gamma) \wedge \gamma \wedge (\beta - 1/2) \wedge 1/2 > 0$.

Proof. The well-posedness of the ODE is discussed in section 6.5, and the actual proof of weak convergence is deferred to section 7. □

5 Limit RNN minimises the average loss

Let us analyse the deterministic limiting ODE. Again we let $\mathsf{H} = (\mathsf{x}, \mathsf{z}, \mathsf{y}, \mathsf{h})$ and $\tilde{\mathsf{H}} = (\tilde{\mathsf{x}}, \tilde{\mathsf{z}}, \tilde{\mathsf{y}}, \tilde{\mathsf{h}})$. If we integrate the limit equation (4.3) with respect to $\mathcal{X}$ again with the limiting measure μ, we will obtain that

$$\begin{aligned}\int_{\mathcal{X}} g_t(\varsigma(\tilde{\mathsf{H}}))\,\mu(d\tilde{\mathsf{H}}) &= \int_{\mathcal{X}} \left[-\alpha \int_0^t \int_{\mathcal{X}} (g_s(\varsigma(\mathsf{H})) - \mathsf{y}) \mathcal{K}_{\mathsf{x},\lambda}(\varsigma(\tilde{\mathsf{H}}), \mathsf{h})\,\mu(d\mathsf{H})\,ds\right] \mu(d\tilde{\mathsf{H}}) \\ &\overset{\text{(Fubini)}}{=} -\alpha \int_0^t \int_{\mathcal{X}} \int_{\mathcal{X}} (g_s(\varsigma(\mathsf{H})) - \mathsf{y}) \mathcal{K}_{\mathsf{x},\lambda}(\varsigma(\tilde{\mathsf{H}}), \mathsf{h})\,\mu(d\mathsf{H})\,\mu(d\tilde{\mathsf{H}})\,ds, \end{aligned} \tag{5.1}$$

which yields

$$\frac{d}{dt}\int_{\mathcal{X}} g_t(\varsigma(\tilde{\mathsf{H}}))\,\mu(d\tilde{\mathsf{H}}) = -\alpha \int_{\mathcal{X}} \int_{\mathcal{X}} (g_t(\varsigma(\mathsf{H})) - \mathsf{y}) \mathcal{K}_{\mathsf{x},\lambda}(\varsigma(\tilde{\mathsf{H}}), \mathsf{h})\,\mu(d\mathsf{H})\,\mu(d\tilde{\mathsf{H}}).$$

Notice that the integrand on the left hand side of (5.1) is $g_t(\varsigma(\tilde{\mathsf{H}}))$ and not $g_t(\varsigma(\mathsf{h}))$. This is because $\varsigma(\tilde{\mathsf{H}})](w) = \sigma(w^\top \tilde{\mathsf{x}} + \left\langle b' \tilde{\mathsf{h}}(w'),\, \lambda \right\rangle)$ mimics the limiting behavior of the pre-limit memory state (1.1).

Now, we can further compute that

$$\begin{aligned}\frac{d}{dt}\left[\frac{1}{2}\int_{\mathcal{X}} (g_t(\varsigma(\tilde{\mathsf{H}})) - \tilde{\mathsf{y}})^2\,\mu(d\tilde{\mathsf{H}})\right] &= \int_{\mathcal{X}} \left[(g_t(\varsigma(\tilde{\mathsf{H}})) - \tilde{\mathsf{y}}) \frac{dg_t(\varsigma(\tilde{\mathsf{H}}))}{dt}\right] \mu(d\tilde{\mathsf{H}}) \\ &= -\alpha \int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{K}_{\mathsf{x},\lambda}(\varsigma(\tilde{\mathsf{H}}), \mathsf{h})(g_t(\varsigma(\tilde{\mathsf{H}})) - \tilde{\mathsf{y}})(g_t(\varsigma(\mathsf{H})) - \mathsf{y})\,\mu(d\mathsf{H})\,\mu(d\tilde{\mathsf{H}}) \end{aligned} \tag{5.2}$$

For convenience we define the kernel on $\mathcal{X}$:

$$\begin{aligned}\tilde{\mathcal{K}}(\tilde{\mathsf{H}}, \mathsf{H}) &= \mathcal{K}_{\mathsf{x},\lambda}(\varsigma(\tilde{\mathsf{H}}), \mathsf{h}) \\ &= \left\langle \sigma(w^\top \mathsf{x} + \langle b' \mathsf{h}(w'),\, \lambda\rangle)[\varsigma(\tilde{\mathsf{H}})](w) + c^2 \sigma'(w^\top \mathsf{x} + \langle b' \mathsf{h}(w'),\, \lambda\rangle) \nabla[\varsigma(\tilde{\mathsf{H}})](w),\, \lambda \right\rangle \\ &= \langle \sigma(w^\top \mathsf{x} + \langle b' \mathsf{h}(w'),\, \lambda\rangle) \sigma(w^\top \tilde{\mathsf{x}} + \langle b' \tilde{\mathsf{h}}(w'),\, \lambda\rangle) \\ &\quad + c^2 \sigma'(w^\top \mathsf{x} + \langle b' \mathsf{h}(w'),\, \lambda\rangle) \sigma'(w^\top \tilde{\mathsf{x}} + \langle b' \tilde{\mathsf{h}}(w'),\, \lambda\rangle) \mathsf{x}^\top \tilde{\mathsf{x}},\, \lambda \rangle \end{aligned}$$

We can now prove that the loss is monotonically decreasing during training.

Proposition 5.1 (Minimisation of averaged square loss)**.**

$$\frac{d}{dt}\left[\frac{1}{2}\int_{\mathcal{X}} (g_t(\varsigma(\tilde{\mathsf{H}})) - \tilde{\mathsf{y}})^2\,\mu(d\tilde{\mathsf{H}})\right] \leq 0. \tag{5.3}$$

Proof. By straight forward computation, we have

$$
\begin{aligned}
&\frac{d}{dt}\left[\frac{1}{2}\int_{\mathcal{X}}(g_t(\varsigma(\tilde{\mathsf{H}}))-\tilde{\mathsf{y}})^2\,\mu(d\tilde{\mathsf{H}})\right]\\
&=-\alpha\int_{\mathcal{X}}\int_{\mathcal{X}}\tilde{\mathcal{K}}(\tilde{\mathsf{H}},\mathsf{H})(g_t(\varsigma(\tilde{\mathsf{H}}))-\tilde{\mathsf{y}})(g_t(\varsigma(\mathsf{H}))-\mathsf{y})\,\mu(d\mathsf{H})\,\mu(d\tilde{\mathsf{H}})\\
&=-\alpha\int_{\mathcal{X}}\int_{\mathcal{X}}\Bigg[\langle\sigma(w^\top\mathsf{x}+\langle b'\mathsf{h}(w'),\,\lambda\rangle)\sigma(w^\top\tilde{\mathsf{x}}+\langle b'\tilde{\mathsf{h}}(w'),\,\lambda\rangle),\,\lambda\rangle(g_t(\varsigma(\tilde{\mathsf{H}}))-\tilde{\mathsf{y}})(g_t(\varsigma(\mathsf{H}))-\mathsf{y})\\
&\quad+\langle c^2\sigma'(w^\top\mathsf{x}+\langle b'\mathsf{h}(w'),\,\lambda\rangle)\sigma'(w^\top\tilde{\mathsf{x}}+\langle b'\tilde{\mathsf{h}}(w'),\,\lambda\rangle)\mathsf{x}^\top\tilde{\mathsf{x}},\,\lambda\rangle(g_t(\varsigma(\tilde{\mathsf{H}}))-\tilde{\mathsf{y}})(g_t(\varsigma(\mathsf{H}))-\mathsf{y})\,\mu(d\mathsf{H})\,\mu(d\tilde{\mathsf{H}})\Bigg]\\
&=-\alpha\Bigg\langle\left(\int_{\mathcal{X}}\sigma(w^\top\mathsf{x}+\langle b'\mathsf{h}(w'),\,\lambda\rangle)(g_t(\varsigma(\mathsf{H}))-\mathsf{y})\,\mu(d\mathsf{H})\right)^2+\\
&\qquad+c^2\left|\int_{\mathcal{X}}\sigma'(w^\top\mathsf{x}+\langle b'\mathsf{h}(w'),\,\lambda\rangle)(g_t(\varsigma(\mathsf{H}))-\mathsf{y})\mathsf{x}\,\mu(d\mathsf{H})\right|^2,\,\lambda\Bigg\rangle\\
&\leq 0,
\end{aligned}
$$

proving the claim. □

Therefore, the function $t\mapsto\frac{1}{2}\int_{\mathcal{X}}(g_t(\varsigma(\tilde{\mathsf{H}}))-\tilde{\mathsf{y}})^2\,\mu(d\tilde{\mathsf{H}})$ is decreasing. This is a useful theoretical guarantee that emerges from the limit analysis. The pre-limit training algorithm (tBPPT) truncates the chain rule – see the algorithm (1) – and therefore it is not guaranteed that parameter updates are in a descent direction for the loss function. That is, in principle, the loss (for the long-run distribution of the data sequence) may actually increase when the parameters are updated. Proposition 5.1 proves that, as $N,k\to\infty$, the RNN model will be updated in a descent direction for the loss function.

6 Technical Lemmas

The unimportant constants $C, C_T > 0$ may change from line to line, and we allow the constants C_T to depend on T.

6.1 A-priori control of the parameter increments

Proof of Lemma 2.8 Assumption 2.3 on the activation functions σ implies for all $k\in\mathbb{N}$

$$\hat{S}^{i,N}_{k+1}=\sigma\left((W^i_k)^\top X_k+\frac{1}{N}\sum_{j=1}^N B^j_k\hat{S}^{j,N}_k\right)\leq 1 \tag{6.1}$$

$$\Delta\hat{S}^{i,N}_{k+1}=\sigma'\left((W^i_k)^\top X_k+\frac{1}{N}\sum_{j=1}^N B^j_k\hat{S}^{i,N}_k\right)\leq C_\sigma, \tag{6.2}$$

where $(\hat{S}^{i,N}_k)_{k\geq 0}$ and $(\Delta\hat{S}^{i,N}_k)_{k\geq 1}$ are as specified in algorithm 1. Therefore, we have

$$\left|C^i_{k+1}-C^i_k\right|\leq\frac{\alpha}{N^{2-\beta}}\left|\hat{S}^{i,N}_{k+1}\right|\left(|\psi^N(\hat{Y}^N_k)|+|Y_k|\right)\leq\frac{\alpha}{N^{2-\beta}}(2N^\gamma+C_y)\leq\frac{C}{N^{2-\beta-\gamma}}, \tag{6.3}$$

so by a telescopic sum argument we have for all $k\leq\lfloor NT\rfloor$

$$\left|C^i_k-C^i_0\right|\leq\sum_{j=0}^{k-1}\left|C^i_{j+1}-C^i_j\right|\leq\frac{CNT}{N^{2-\beta-\gamma}}=\frac{CT}{N^{1-\beta-\gamma}},$$

and

$$|C_k^i| \le |C_0^i| + |C_k^i - C_0^i| \le 1 + \frac{CT}{N^{1-\beta-\gamma}} \le C_T.$$

We may further conclude that

$$\begin{aligned}
\left\|W_{k+1}^i - W_k^i\right\| &\le \frac{\alpha}{N^{2-\beta}} \left|C_k^i\right| \left[|\psi^N(\hat{Y}_k^N)| + |Y_k|\right] \left|\Delta \hat{S}_{k+1}^{i,N}\right| \le \frac{\alpha C_\sigma C_T}{N^{2-\beta}} \left[2N^\gamma + C_y\right] \le \frac{C_T}{N^{2-\beta-\gamma}}, \\
\left|B_{k+1}^i - B_k^i\right| &\le \frac{\alpha}{N^{3-\beta}} \left(|\psi^N(\hat{Y}_k^N)| + |Y_k|\right) \sum_{\ell=1}^N |C_k^\ell| |\Delta \hat{S}_{k+1}^{\ell,N}| \le \frac{\alpha C_\sigma C_T}{N^{2-\beta}} \left[2N^\gamma + C_y\right] \le \frac{C_T}{N^{2-\beta-\gamma}},
\end{aligned}$$

so by following the telescoping sum argument as above we conclude, as desired,

$$|C_k^i - C_0^i| + \left\|W_k^i - W_0^i\right\| + |B_k^i - B_0^i| \le \frac{C_T}{N^{1-\beta-\gamma}}.$$

Proof of Remark 2.9 The proof of remark relies on the observation that

$$\left|C_{k+1}^i\right| - \left|C_k^i\right| \le \left|C_{k+1}^i - C_k^i\right| \le \frac{\alpha}{N^{2-\beta}} \left|S_{k+1}^i\right| \left(|\hat{Y}_k^N| + |Y_k|\right) \le \frac{\alpha C_\sigma C_y}{N^{2-\beta}} + \frac{\alpha C_\sigma^2}{N^2} \sum_{i=1}^N |C_k^i|, \tag{6.4}$$

so by letting $\bar{C}_k := \frac{\sum_{i=1}^N |C_k^i|}{N}$, one has

$$\bar{C}_{k+1} - \bar{C}_k \le \frac{\alpha C_\sigma C_y}{N^{2-\beta}} + \frac{\alpha C_\sigma^2}{N} \bar{C}_k \implies \bar{C}_{k+1} \le \left(1 + \frac{\alpha C_\sigma^2}{N}\right) \bar{C}_k + \frac{\alpha C_\sigma C_y}{N^{2-\beta}},$$

and by Lemma A.1, for all $k \le \lfloor NT \rfloor$, one has

$$\bar{C}_k \le \left(1 + \frac{\alpha C_\sigma^2}{N}\right)^{\lfloor TN \rfloor} \left(1 + \frac{\alpha C_\sigma C_y / N^{2-\beta}}{\alpha C_\sigma^2 / N}\right) \le C \exp(\alpha C_\sigma^2 T) =: C_T.$$

One can that make use of equation (6.4) to conclude that

$$\left|C_{k+1}^i - C_k^i\right| \le \frac{\alpha C_\sigma C_y}{N^{2-\beta}} + \frac{\alpha C_\sigma \bar{C}_k}{N} \le \frac{C_T}{N},$$

and that $\left|C_k^i\right| \le NTC_T/N = C_T$ is bounded whenever $k \le \lfloor NT \rfloor$ by a telescoping sum arguments. This enables us to show further that

$$\begin{aligned}
\left\|W_{k+1}^i - W_k^i\right\| &\le \frac{\alpha}{N^{2-\beta}} \left|C_k^i\right| \left[|\hat{Y}_k^N| + |Y_k|\right] \left|\Delta \hat{S}_{k+1}^{i,N}\right| \\
&\le \frac{\alpha C_\sigma \left|C_k^i\right|}{N^{2-\beta}} \left[|Y_k| + \frac{1}{N^\beta} \sum_{\bullet=1}^N |C_k^\bullet|\right] \\
&\le \frac{C_T}{N^{2-\beta}} \left[|C_y| + C_T N^{1-\beta}\right] \le \frac{C_T}{N}, \\
\left|B_{k+1}^i - B_k^i\right| &\le \frac{\alpha}{N^{3-\beta}} \left(|\hat{Y}_k^N| + |Y_k|\right) \sum_{\ell=1}^N |C_k^\ell| |\Delta \hat{S}_{k+1}^{\ell,N}| \\
&\le \frac{\alpha C_\sigma}{N^{3-\beta}} \left(C_y + \frac{1}{N^\beta} \sum_{\bullet=1}^N |C_k^\bullet| |\hat{S}_{k+1}^{\bullet,N}|\right) \left(\sum_{\ell=1}^N |C_k^\ell|\right) \\
&= \frac{\alpha C_\sigma}{N^{3-\beta}} \left(C_y + N^{1-\beta} C_\sigma C_T\right) N C_T \le \frac{C_T}{N}.
\end{aligned}$$

6.2 Proof of Lemma 3.2

$$
\begin{aligned}
&\left\|F(x,z,y,h,\epsilon,\eta) - F(\tilde{x},\tilde{z},\tilde{y},\tilde{h},\epsilon,\eta)\right\|_{\mathcal{X}}^2 \\
&= |g(x,z) - g(\tilde{x},\tilde{z})|^2 + |f(x,z) - f(\tilde{x},\tilde{z})|^2 + \left\|[\varsigma(x,z,y,h)](w) - [\varsigma(\tilde{x},\tilde{z},\tilde{y},\tilde{h})](w)\right\|_{H^1(\lambda)}^2 \\
&\leq L^2\left(|x-\tilde{x}|^2 + |z-\tilde{z}|^2\right) + \left\|[\varsigma(x,z,y,h)](w) - [\varsigma(\tilde{x},\tilde{z},\tilde{y},\tilde{h})](w)\right\|_{H^1(\lambda)}^2
\end{aligned}
$$

Note

$$
\begin{aligned}
&\left|[\varsigma(x,z,y,h)](w) - [\varsigma(\tilde{x},\tilde{z},\tilde{y},\tilde{h})](w)\right| \\
&= \left|\sigma(w^\top x + \langle b' h(w'), \lambda\rangle) - \sigma(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\right| \\
&\leq C_\sigma \left|w^\top(x-\tilde{x}) + \langle b'(h(w') - \tilde{h}(w'), \lambda\rangle\right| \\
&\leq C_\sigma \left(|w|\,|x-\tilde{x}| + \int_{\mathbb{R}^d} \left|h(w') - \tilde{h}(w')\right| \lambda(dw')\right) \\
&\leq C_\sigma \left(|w|\,|x-\tilde{x}| + \left(\int_{\mathbb{R}^d} \left|h(w') - \tilde{h}(w')\right|^2 \lambda(dw')\right)^{1/2}\right) \\
&\leq C_\sigma \left(|w|\,|x-\tilde{x}| + \|h-\tilde{h}\|_{H^1(\lambda)}\right). \\
&\quad \implies \left|\sigma(w^\top x + \langle b' h(w'), \lambda\rangle) - \sigma(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\right|^2 \\
&\leq 2C_\sigma^2 \left(|w|^2\,|x-\tilde{x}|^2 + \|h-\tilde{h}\|_{H^1(\lambda)}^2\right).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\left|\nabla[\varsigma(x,z,y,h)](w) - [\nabla\varsigma(\tilde{x},\tilde{z},\tilde{y},\tilde{h})](w)\right| \\
&= \left|\nabla\sigma(w^\top x + \langle b' h(w'), \lambda\rangle) - \nabla\sigma(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\right| \\
&= \left|\sigma'(w^\top x + \langle b' h(w'), \lambda\rangle)x - \sigma'(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\tilde{x}\right| \\
&\leq \left|\sigma'(w^\top x + \langle b' h(w'), \lambda\rangle) - \sigma'(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\right| |x| + \left|\sigma'(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\tilde{x}\right| |x-\tilde{x}| \\
&\leq C_\sigma \left|w^\top(x-\tilde{x}) + \langle b'(h(w') - \tilde{h}(w'), \lambda\rangle\right| + C_\sigma\,|x-\tilde{x}| \\
&\leq C_\sigma \left(|w|\,|x-\tilde{x}| + \|h-\tilde{h}\|_{H^1(\lambda)}\right) + C_\sigma\,|x-\tilde{x}|. \\
\implies &\left|\nabla\sigma(w^\top x + \langle b' h(w'), \lambda\rangle) - \nabla\sigma(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\right|^2 \\
&\leq 3C_\sigma^2 \left((1+|w|^2)|x-\tilde{x}|^2 + \|h-\tilde{h}\|^2\right),
\end{aligned}
$$

so, using Assumptions 2.1 and 2.5 we have the bound

$$
\begin{aligned}
\|[\varsigma(x,z,y,h)] - [\varsigma(\tilde{x},\tilde{z},\tilde{y},\tilde{h})]\|_{H^1(\lambda)}^2 &= \left\|\sigma(w^\top x + \langle b' h(w'), \lambda\rangle) - \sigma(w^\top \tilde{x} + \langle b'\tilde{h}(w'), \lambda\rangle)\right\|_{H^1(\lambda)}^2 \\
&\leq C_\sigma^2 |x-\tilde{x}|^2 \int_{\mathbb{R}^d} (5|w|^2 + 3)\,\lambda(dw) + 5C_\sigma^2 \left\|h-\tilde{h}\right\|_{H^1(\lambda)}^2 \\
&\leq 8C_\sigma^2\,|x-\tilde{x}|^2 + 5C_\sigma^2 \left\|h-\tilde{h}\right\|_{H^1(\lambda)}^2, \qquad (6.5)
\end{aligned}
$$

and

$$\left\|F(x,z,y,h,\epsilon,\eta)-F(\tilde{x},\tilde{z},\tilde{y},\tilde{h},\epsilon,\eta)\right\|_{\mathcal{X}}^2 \leq (L^2+8C_\sigma^2)\left(|x-\tilde{x}|^2+|z-\tilde{z}|^2+|y-\tilde{y}|^2+\left\|h-\tilde{h}\right\|^2\right)$$

By Assumption 2.3 we have that the Lipschitz constant is bounded by

$$q_0^2 := L^2+8C_\sigma^2 < 1,$$

and that $F(\cdot,\epsilon,\eta)$ is a continuous contraction for any realisations of ϵ and η.

6.3 Analysis of the auxiliary processes (proof of proposition 3.8)

Preliminary bounds for the auxiliary processes Let us begin by establishing some preliminary bounds for $h_k(w)$, $h_k^N(w)$ and $v_k^N(w)$.

Lemma 6.1. *All of the $h_k(w)$, $h_k^N(w)$ and $v_k^N(w)$ are elements in $\mathcal{H}$. As a result, the following holds:*

$$\sup_w |h_k(w)| \vee \sup_w |h_k^N(w)| \vee \sup_w |v_k^N(w)| \leq 1, \tag{6.6}$$

$$\sup_w |\nabla h(w)| \vee \sup_w |\nabla h_k^N(w)| \vee \sup_w |\nabla v_k^N(w)| \leq C_\sigma < 1, \tag{6.7}$$

$$\max_{i,j}\sup_w \left[|\partial_{ij}h_k(w)| \vee |\partial_{ij}h_k^N(w)| \vee |\partial_{ij}v_k^N(w)|\right] \leq C_\sigma < 1.$$

Proof. To see this, we record the following recursive formula for the gradients of the above random sequences of functions for our convenience:

$$\begin{aligned}
\nabla v_{k+1}^N(w) &= \sigma'\left(w^\top X_k + \frac{1}{N}\sum_{i=1}^N B_k^i v_k^N(W_{k-1}^i)\right)X_k, \quad \nabla v_0^N(w)\equiv 0,\\
\nabla h_{k+1}^N(w) &= \sigma'(w^\top X_k + \langle b' h_k^N(w'), \lambda^N\rangle)X_k, \quad \nabla h_0^N(w)\equiv 0,\\
\nabla h_{k+1}(w) &= \sigma'(w^\top X_k + \langle b' h_k(w'), \lambda\rangle)X_k, \quad \nabla h_0(w)\equiv 0.
\end{aligned}$$

We further note that

$$\left|\frac{1}{N}\sum_{i=1}^N B_k^i v_k^N(W_{k-1}^i)\right| \leq \frac{1}{N}\sum_{i=1}^N \left|B_k^i\right|\left|v_k^N(W_k^i)\right| \leq C_T,$$

where C_T depends on the size of the increments as derived in Lemma 2.8. Similarly, we have

$$\begin{aligned}
\left|\langle b' h_k^N(w'), \lambda^N\rangle\right| &= \left|\frac{1}{N}\sum_{i=1}^N B^i h_k^N(W^i)\right| \leq \frac{1}{N}\sum_{i=1}^N \left|B^i\right|\left|h_k^N(W^i)\right| \leq 1, \qquad (6.8)\\
|\langle b' h_k(w'), \lambda\rangle| &= \left|\int_{\mathbb{R}\times\mathbb{R}^d} b' h_k(w')\,\lambda(db', dw')\right| \leq \int_{\mathbb{R}\times\mathbb{R}^d} |b'|\,|h_k(w')|\;\lambda(db', dw') \leq 1.
\end{aligned}$$

So we know that for finite k, N, all of the $h_k(w)$, $h_k^N(w)$ and $v_k^N(w)$ are elements of $H^1(\lambda)$. □

Bounds for the differences between the auxiliary processes Recall that

$$\Gamma_k^N(w) = v_k^N(w) - h_k(w) = \Gamma_k^{N,1}(w) + \Gamma_k^{N,2}(w),$$

where

$$\begin{aligned}
\Gamma_k^{N,1}(w) &= h_k^N(w) - h_k(w), \qquad (6.9)\\
\Gamma_k^{N,2}(w) &= v_k^N(w) - h_k^N(w).
\end{aligned}$$

Since the activation function is regular, we know that

$$\begin{aligned}\left|\Gamma_k^{N,1}(w)\right|^2 + \left|\nabla\Gamma_k^{N,1}(w)\right|^2 &\leq \left|\sigma(w^\top X_k + \left\langle b' h_k^N, \lambda^N\right\rangle) - \sigma(w^\top X_k + \langle b' h_k, \lambda\rangle)\right|^2 \\ &\quad + \left|\sigma'(w^\top X_k + \left\langle b' h_k^N, \lambda^N\right\rangle) - \sigma'(w^\top X_k + \langle b' h_k, \lambda\rangle)\right|^2 |X_k|^2 \\ &\leq (1+|X_k|^2) C_\sigma^2 \left(\left\langle b' h_k^N(w'), \lambda^N\right\rangle - \langle b' h_k(w'), \lambda\rangle\right)^2 \\ &\leq 2C_\sigma^2 \left(E_k^{N,1}\right)^2,\end{aligned}$$

where

$$E_k^{N,1} := \left\langle b' h_k^N(w'), \lambda^N\right\rangle - \langle b' h_k(w'), \lambda\rangle .$$

Similarly,

$$\left|\Gamma_k^{N,2}(w)\right|^2 + \left|\nabla\Gamma_k^{N,2}(w)\right|^2 \leq 2C_\sigma^2 \left(E_k^{N,2}\right)^2,$$

where for $k \geq 1$ we define

$$E_k^{N,2} = \frac{1}{N}\sum_{i=1}^N B_k^i v_k^N(W_{k-1}^i) - \left\langle b' h_k^N(w'), \lambda^N\right\rangle,$$

and $E_0^{N,2} = 0$[2]. Finally, we have that

$$\left|\Gamma_k^N(w)\right|^2 + \left|\nabla\Gamma_k^N(w)\right|^2 \leq 2C_\sigma^2 \left(E_k^{N,1} + E_k^{N,2}\right)^2 \leq 4C_\sigma^2 \left(\left(E_k^{N,1}\right)^2 + \left(E_k^{N,2}\right)^2\right).$$

This yields the following lemma

Lemma 6.2. *For all $k \geq 0$,*

$$\mathbb{E}\left\|\Gamma_k^{N,1}(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^{N,1}(W^i)\right]^2 \leq 2C_\sigma^2 \mathbb{E}\left[E_k^{N,1}\right]^2 \tag{6.10}$$

$$\mathbb{E}\left\|\Gamma_k^{N,2}(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^{N,2}(W^i)\right]^2 \leq 2C_\sigma^2 \mathbb{E}\left[E_k^{N,2}\right]^2 \tag{6.11}$$

$$\mathbb{E}\left\|\Gamma_k^N(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^N(W^i)\right]^2 \leq 4C_\sigma^2 \left(\mathbb{E}\left[E_k^{N,1}\right]^2 + \mathbb{E}\left[E_k^{N,2}\right]^2\right) \tag{6.12}$$

where $a \vee b = \max(a,b)$.

Proof. The proof for (6.10)-(6.12) are similar, so we will only prove (6.10) here. To that end, we note that the bound $E_k^{N,1}$ is uniform in w, so for all i,

$$\mathbb{E}\left[\Gamma_k^{1,N}(W_0^i)\right]^2 \leq 2C_\sigma^2 \mathbb{E}\left[E_k^{N,1}\right]^2.$$

Moreover,

$$\begin{aligned}\mathbb{E}\left\|\Gamma_k^{N,1}(w)\right\|_{H^1(\lambda)}^2 &= \mathbb{E}\left[\int_{\mathbb{R}^d}\left[\left(\Gamma_k^{N,1}(w)\right)^2 + \left(\nabla\Gamma_k^{N,1}(w)\right)^2\right]\lambda(dw)\right] \leq 2C_\sigma^2 \mathbb{E}\left[\int_{\mathbb{R}^d}\left(E_k^{N,1}\right)^2 \lambda(dw)\right] \\ &= 2C_\sigma^2 \mathbb{E}\left[E_k^{N,1}\right]^2.\end{aligned}$$

This completes the proof. □

It remains for us to control $\mathbb{E}\left[E_k^{N,1}\right]^2$ and $\mathbb{E}\left[E_k^{N,2}\right]^2$.

[2]We note that when $k = 0$ then $v_k^N(w) = h_k^N(w) \equiv 0$, so the value of $E_0^{N,2} = 0$ regardless of what value we set for W_{-1}^i.

Lemma 6.3. *For all $k \geq 0$,*

$$\mathbb{E}\left[E_k^{N,1}\right]^2 \leq \frac{C}{N},$$

for an unimportant constant $C < \infty$.

Proof. This is trivial for $k = 0$ since $h_0^N(w) = h_0(w) \equiv 0$, so $E_0^{N,1} = 0$. For general k, we note by Young's inequality that

$$\begin{aligned}\left(E_k^{N,1}\right)^2 &= \left(\left\langle b'(h_k^N(w') - h_k(w')), \lambda\right\rangle + \left\langle b' h_k^N(w'), \lambda^N - \lambda\right\rangle\right)^2 \\ &\leq 2\left(\left\langle b'(h_k^N(w') - h_k(w')), \lambda\right\rangle\right)^2 + 2\left(\left\langle b' h_k^N(w'), \lambda^N - \lambda\right\rangle\right)^2.\end{aligned}$$

Now note that

$$\begin{aligned}\mathbb{E}\left(\left\langle b' h_k^N(w'), \lambda^N - \lambda\right\rangle^2\right) &= \mathbb{E}\left[\mathbb{E}\left[\left\langle b' h_k^N(w'), \lambda^N - \lambda\right\rangle^2 \mid h_k^N\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \left(B^i h_k^N(W^i) - \int_{\mathbb{R}^{1+d}} b' h_k^N(w')\,\lambda(db'\,dw')\right)\right)^2 \mid h_k^N\right]\right].\end{aligned}$$

Since $B^i h_k^N(W^i) \mid h_k^N$ is conditionally mutually iid for $i = 1, \cdots, N$,

$$\begin{aligned}\mathbb{E}\left(\left\langle b' h_k^N(w'), \lambda^N - \lambda\right\rangle\right)^2 &= \frac{1}{N}\mathbb{E}\left[\mathbb{E}\left[\left(B^1 h_k^N(W^1) - \int_{\mathbb{R}^{1+d}} b' h_k^N(w')\,\lambda(db'\,dw')\right)^2 \mid h_k^N\right]\right] \\ &= \frac{1}{N}\mathbb{E}\left[\mathbb{E}\left[\left(B^1 h_k^N(W^1)\right)^2 \mid h_k^N\right] - \left(\int_{\mathbb{R}^{1+d}} b' h_k^N(w')\,\lambda(db'\,dw')\right)^2\right] \\ &\leq \frac{1}{N}\mathbb{E}\left[\mathbb{E}\left[\left(B^1 h_k^N(W^1)\right)^2 \mid h_k^N\right]\right] + \frac{1}{N}\mathbb{E}\left(\int_{\mathbb{R}^{1+d}} b' h_k^N(w')\,\lambda(db'\,dw')\right)^2 \\ &\leq \frac{1}{N}\mathbb{E}\left[\left(B^1 h_k^N(W^1)\right)^2\right] + \frac{1}{N}\mathbb{E}\left(\int_{\mathbb{R}^{1+d}} b' h_k^N(w')\,\lambda(db'\,dw')\right)^2 \\ &\leq \frac{C}{N},\end{aligned}$$

for an unimportant constant $C < \infty$. In addition, we have

$$\begin{aligned}\mathbb{E}\left(\left\langle b'(h_k^N(w') - h_k(w')), \lambda\right\rangle\right)^2 &\leq \int_{\mathbb{R}^{1+d}} (b')^2\,\lambda(db'\,dw')\mathbb{E}\int_{\mathbb{R}^{1+d}} |h_k^N(w') - h_k(w')|^2\,\lambda(db'\,dw') \\ &\leq \mathbb{E}\int_{\mathbb{R}^{1+d}} \Big(\sigma(w^\top X_{k-1} + \left\langle b'' h_{k-1}^N(w''), \lambda^N\right\rangle) \\ &\quad - \sigma(w^\top X_{k-1} + \langle b'' h_{k-1}(w''), \lambda\rangle)\Big)^2 \lambda(db'\,dw') \\ &\leq \mathbb{E}\int_{\mathbb{R}^{1+d}} C_\sigma^2 \left(E_{k-1}^{N,1}\right)^2 \lambda(db'\,dw') = C_\sigma^2 \mathbb{E}\left(E_{k-1}^{N,1}\right)^2.\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left(E_k^{N,1}\right)^2 \leq \frac{2C}{N} + 2C_\sigma^2 \mathbb{E}\left(E_{k-1}^{N,1}\right)^2. \tag{6.13}$$

By lemma A.1, for all $k \geq 0$ we have

$$\mathbb{E}\left(E_k^{N,1}\right)^2 \leq \frac{2C}{N(1 - 2C_\sigma^2)} = \frac{C}{N},$$

noting that $2C_\sigma^2 < 1$. The unimportant constant C changes from line to line. This completes the proof. □

Similarly,

Lemma 6.4. *For all $k \le \lfloor NT \rfloor$,*

$$\mathbb{E}\left(E_k^{N,2}\right)^2 \le \frac{C_T}{N^{2-2\beta-2\gamma}}.$$

Proof. Again, we have $v_0^N(w) = h_k^N(w) \equiv 0$ so $E_0^{N,2} = 0$. For general k, we conclude by Young's inequality:

$$\begin{aligned}
\left(E_k^{N,2}\right)^2 &= \left(\left\langle b'(v_k^N(w') - h_k^N(w')), \lambda^N \right\rangle + \frac{1}{N}\sum_{i=1}^N \left(B_k^i v_k^N(W_{k-1}^i) - B_0^i v_k^N(W_0^i)\right)\right)^2 \\
&\le 2\left(\left\langle b'(v_k^N(w') - h_k^N(w')), \lambda^N \right\rangle\right)^2 + 2\left(\frac{1}{N}\sum_{i=1}^N \left(B_k^i v_k^N(W_{k-1}^i) - B_0^i v_k^N(W_0^i)\right)\right)^2 \\
&\le 2\left(\left\langle b'(v_k^N(w') - h_k^N(w')), \lambda^N \right\rangle\right)^2 + \frac{2}{N}\sum_{i=1}^N \left(B_k^i v_k^N(W_{k-1}^i) - B_0^i v_k^N(W_0^i)\right)^2
\end{aligned}$$

We first analyse the second term, and we emphasise that this term vanishes at $k = 0$ as v_0^N is defined to be zero. For $k \ge 1$, we have

$$\begin{aligned}
\left|B_0^i v_k^N(W_0^i) - B_k^i v_k^N(W_{k-1}^i)\right| &= \left|B_0^i(v_k^N(W_0^i) - v_k^N(W_{k-1}^i)) - v_k^N(W_{k-1}^i)(B_0^i - B_k^i)\right| \\
&\le \left(\left|B_0^i\right|\left|v_k^N(W_0^i) - v_k^N(W_{k-1}^i)\right| + \left|v_k^N(W_{k-1}^i)\right|\left|(B_0^i - B_k^i)\right|\right) \\
&\le \left(C_\sigma\left|B_0^i\right|\left|X_{k-1}^\top(W_0^i - W_{k-1}^i)\right| + \left|(B_0^i - B_k^i)\right|\right) \\
&\le \left(C_\sigma\left|W_0^i - W_{k-1}^i\right| + \left|B_0^i - B_k^i\right|\right), \\
\implies \mathbb{E}\left|B_0^i v_k^N(W^i) - B_k^i v_k^N(W_{k-1}^i)\right|^2 &\le 2\mathbb{E}\left(C_\sigma^2\left|W_0^i - W_{k-1}^i\right|^2 + \left|B_0^i - B_k^i\right|^2\right) \\
&\le \frac{C_T}{N^{2-2\beta-2\gamma}},
\end{aligned}$$

where in the last inequality, we used Lemma 2.8. In addition, for $k \ge 1$ we have

$$\begin{aligned}
\left\langle b'(v_k^N(w') - h_k^N(w')), \lambda^N \right\rangle &\le \frac{1}{N}\sum_{i=1}^N \left|B^i\right|\left|v_k^N(W^i) - h_k^N(W^i)\right| \\
&\le \frac{C_\sigma}{N}\sum_{i=1}^N \left|\frac{1}{N}\sum_{j=1}^N B_{k-1}^j v_{k-1}^N(W_{k-2}^j) - \left\langle b' h_{k-1}^N(w'), \lambda^N \right\rangle\right| \\
&= C_\sigma\left|E_{k-1}^{N,2}\right|
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left(E_k^{N,2}\right)^2 \le \frac{2C_T}{N^{2-2\beta-2\gamma}} + 2C_\sigma^2\mathbb{E}\left(E_{k-1}^{N,2}\right)^2. \tag{6.14}$$

By Lemma A.1, for all $k \ge 0$ we have

$$\mathbb{E}\left(E_k^{N,2}\right)^2 \le \frac{2C_T}{N^{2-2\beta-2\gamma}(1-2C_\sigma^2)} = \frac{C_T}{N^{2-2\beta-2\gamma}},$$

noting that $2C_\sigma^2 < 1$. □

Proof. (Proof of proposition 3.8) Lemma 6.2 and 6.3 imply that for any $k \ge 0$, there is a constant $C > 0$ such that

$$\mathbb{E}\left\|\Gamma_k^{N,1}(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^{N,1}(W^i)\right]^2 \le \frac{C}{N},$$

Similarly, lemma 6.2 and 6.4 imply that for all $k \leq \lfloor NT \rfloor$, there is another constant $C_T > 0$, depending on T, such that

$$\mathbb{E}\left\|\Gamma_k^{N,2}(w)\right\|_{H^1(\lambda)}^2 \vee \max_i \mathbb{E}\left[\Gamma_k^{N,2}(W^i)\right]^2 \leq \frac{C_T}{N^{2-2\beta-2\gamma}}.$$

This completes the proof. □

6.4 Proof of Pre-Limit Evolution (Lemma 4.1)

We expand $h(W_{k+1}^i) - h(W_k^i)$ using Taylor series as followed:

$$\begin{aligned} h(W_{k+1}^i) - h(W_k^i) &= \nabla h(W_k^{i,*})^\top (W_{k+1}^i - W_k^i) \\ &= \nabla h(W_k^i)^\top (W_{k+1}^i - W_k^i) + (W_{k+1}^i - W_k^i)^\top \mathsf{Hess}\, h(W_k^{i,**})(W_{k+1}^i - W_k^i), \end{aligned}$$

where $W_k^{i,*}, W_k^{i,**}$ are points in the line segment connecting the points W_k^i and W_{k+1}^i. The following remainder terms in the Taylor's expansion is small by Lemma 2.8 (specifically by (6.3)):

$$\begin{aligned} \left|(C_{k+1}^i - C_k^i)\nabla h(W_k^{i,*})^\top (W_{k+1}^i - W_k^i)\right| &\leq \frac{C_T}{N^{4-2\beta-2\gamma}} \\ \left|C_k^i (W_{k+1}^i - W_k^i)\mathsf{Hess}\, h(W_k^{i,*})(W_{k+1}^i - W_k^i)\right| &\leq C_T(Cd^2)\left|W_{k+1}^i - W_k^i\right|^2 \leq \frac{C_T}{N^{4-2\beta-2\gamma}}, \end{aligned} \tag{6.15}$$

noting that the derivatives of h are bounded. From this, we obtain the following evolution equation

$$\begin{aligned} \triangle g_k^N(h) &= g_{k+1}^N(h) - g_k^N(h) \\ &= \frac{1}{N^\beta}\sum_{i=1}^N (C_{k+1}^i h(W_{k+1}^i) - C_k^i h(W_k^i)) \\ &= \frac{1}{N^\beta}\sum_{i=1}^N [(C_{k+1}^i - C_k^i)h(W_k^i) + C_k^i(h(W_{k+1}^i) - h(W_k^i)) + (C_{k+1}^i - C_k^i)(h(W_{k+1}^i) - h(W_k^i))] \\ &= \frac{1}{N^\beta}\sum_{i=1}^N [(C_{k+1}^i - C_k^i)h(W_k^i) + C_k^i \nabla h(W_k^i)(W_{k+1}^i - W_k^i) + C_k^i (W_{k+1}^i - W_k^i)\mathsf{Hess}\, h(W_k^{i,**})(W_{k+1}^i - W_k^i) \\ &\quad + (C_{k+1}^i - C_k^i)\nabla h(W_k^{i,*})^\top (W_{k+1}^i - W_k^i)]. \end{aligned}$$

Notice that with the specific choice of the learning rate $\alpha^N = \frac{\alpha}{N^{2-2\beta}}$ by (2.3) we get after training is taken into account

$$\begin{aligned} &\frac{1}{N^\beta}\sum_{i=1}^N [(C_{k+1}^i - C_k^i)h(W_k^i) + C_k^i \nabla h(W_k^i)^\top (W_{k+1}^i - W_k^i)] \\ &= -\frac{\alpha}{N^2}(\psi^N(\hat{Y}_k^N) - Y_k)\sum_{i=1}^N \left[\hat{S}_{k+1}^i h(W_k^i) + C_k^i \Delta\hat{S}_{k+1}^{i,N}\nabla h(W_k^i)^\top X_k\right] = \delta^{(1)} g_k^N(h). \end{aligned}$$

Therefore, for all $k \leq \lfloor NT \rfloor$,

$$\begin{aligned} &\left|\triangle g_k^N(h) - \delta^{(1)} g_k^N(h)\right| \\ &= \left|\frac{1}{N^\beta}\sum_{i=1}^N \left[C_k^i (W_{k+1}^i - W_k^i)\mathsf{Hess}\, h(W_k^{i,**})(W_{k+1}^i - W_k^i) + (C_{k+1}^i - C_k^i)\nabla h(W_k^{i,*})^\top (W_{k+1}^i - W_k^i)\right]\right| \\ &\leq \frac{1}{N^\beta}\sum_{i=1}^N \left[\left|C_k^i (W_{k+1}^i - W_k^i)\mathsf{Hess}\, h(W_k^{i,**})(W_{k+1}^i - W_k^i)\right| + \left|(C_{k+1}^i - C_k^i)\nabla h(W_k^{i,*})^\top (W_{k+1}^i - W_k^i)\right|\right] \\ &\leq \frac{N}{N^\beta} \times \frac{C_T}{N^{4-2\beta-2\gamma}} = \frac{C_T}{N^{3-\beta-2\gamma}}. \end{aligned}$$

It is important to note that $3-\beta-2\gamma > 2+1-\beta-2\gamma > 2$, so $\left|\triangle g_k^N(h) - \delta^{(1)} g_k^N(h)\right| = o(N^{-2})$ uniformly for $k \le \lfloor NT \rfloor$ and $h \in \mathcal{H}$.

6.5 Well-posedness of the limit ODE

The affine ODE on the dual space $(H^1(\lambda))^*$ (which is a Banach space) could be written as

$$\frac{d}{dt} g_t = \mathcal{A}(g_t) + b, \quad g_0 = 0,$$

where $\mathcal{A}$ is a linear operator from $(H^1(\lambda))^*$ to $(H^1(\lambda))^*$ as defined below,

$$\mathcal{A} : g \in (H^1(\lambda))^* \mapsto \left[\mathcal{A}(g) : h \mapsto -\int_{\mathcal{X}} g(\varsigma(\mathsf{H})) \mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h})\, \mu(d\mathsf{H}) \right];$$

and b is the following linear functional:

$$b : h \mapsto \int \mathsf{y} \mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h})\, \mu(d\mathsf{H})$$

Therefore the ODE admits a unique solution if $\mathcal{A}$ is bounded. In fact, for $\mathsf{H} \in \mathcal{X}$, we have

$$|\mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h})| \le \int_{\mathbb{R}^d} \left[|h(w)| + |\nabla h(w)| \right] \lambda(dw) \le \|h\|_{H^1(\lambda)},$$

so

$$\begin{aligned} |[\mathcal{A}(g)](h)| &\le \int_{\mathcal{X}} |g(\mathsf{h})| |\mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h})|\, \mu(d\mathsf{H}) \\ &\le \int_{\mathcal{X}} \|g\|_{(H^1(\lambda))^*} \|\varsigma(\mathsf{H})\|_{H^1(\lambda)} \|h\|_{H^1(\lambda)}\, \mu(d\mathsf{H}) \\ &\le \sqrt{1 + C_\sigma^2}\, \|g\|_{(H^1(\lambda))^*} \|h\|_{H^1(\lambda)}, \end{aligned} \tag{6.16}$$

thanks to the fact that any elements of $\varsigma(\mathsf{H})$ is in $\mathcal{H}$ so the control (3.5) holds. Therefore, we have

$$\|\mathcal{A}(g)\|_{(H^1(\lambda))^*} \le \sqrt{1 + C_\sigma^2}\, \|g\|_{(H^1(\lambda))^*} \implies \|\mathcal{A}\| \le \sqrt{1 + C_\sigma^2}.$$

Moreover, b is bounded the following sense:

$$|b(h)| \le \int |\mathsf{y}| |\mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h})|\, \mu(d\mathsf{x}, d\mathsf{z}, d\mathsf{y}, d\mathsf{h}) \le C_y \|h\|_{H^1(\lambda)} \implies \|b\|_{(H^1(\lambda))^*} \le C_y.$$

Therefore, the following exponential operator is well defined for any $T > 0$:

$$\exp(t\mathcal{A}) = \sum_{i=0}^{\infty} \frac{t^i \mathcal{A}^{\circ i}}{i!}, \quad \mathcal{A}^{\circ i} = \underbrace{\mathcal{A} \circ ... \circ \mathcal{A}}_{i \text{ times}},$$

and therefore

Proposition 6.5. *The ODE* (4.3) *admits the following classical solution:*

$$g_t = \int_0^t \exp((t-s)\mathcal{A}) b\, ds, \tag{6.17}$$

and when acting on $h \in H^1(\lambda)$ we have

$$g_t(h) = \int_0^t \exp((t-s)\mathcal{A}) b(h)\, ds.$$

In particular, we have the following control over the operator norm

$$\|g_t\|_{(H^1(\lambda))^*} \leq T\exp(T\|\mathcal{A}\|)\|b\|_{(H^1(\lambda))^*} \leq C_y T e^{\sqrt{1+C_\sigma^2}T} := C_T \tag{6.18}$$

where $\|\mathcal{A}\|$ is the operator norm of the operator $\mathcal{A}$, which is proven to be bounded by (6.16). So for all h we have $|g_t(h)| \leq C_T \|h\|_{H^1(\lambda)}$ for any finite $T > 0$.

Proof. We follow [26] for our discussion. Firstly, Theorem 1.2-3 of [26] states that $\mathcal{A}$ induces a unique uniformly continuous semigroup $\exp(t\mathcal{A})$. Indeed, consider the sequence of operators from $(H^1(\lambda))^*$ to $(H^1(\lambda))^*$:

$$S^N(t;\mathcal{A}) = \sum_{i=0}^{N} \frac{t^i \mathcal{A}^{\circ i}}{i!}$$

By triangle inequality we have the operator norm control $\|S^N(t;\mathcal{A})\| \leq \exp(t\|\mathcal{A}\|) < +\infty$, uniform in N. Therefore the partial sums $(S^N(t,\mathcal{A}))_{N\geq 0}$ must converge absolutely. Since the space of operators from $(H^1(\lambda))^*$ to $(H^1(\lambda))^*$ is Banach, the partial sum $(S^N(t,\mathcal{A}))_{N\geq 0}$ must also converge to an operator in operator norm, for which we define it to be $\exp(t\mathcal{A})$.

Since $b \in (H^1(\lambda))^*$ is constant in t, then by Corollary 2.5 of [26] the ODE 4.3 admits a classical solution, given by the formula (6.17) according to Corollary 2.2 of [26]. □

7 Proof of weak convergence

The actual convergence of the ODE is broken mainly into three steps. Firstly, we show in subsection 7.1 that one could modify the increments $\delta^{(1)}g_k^N(h)$ by replacing the current parameters (C_k^i, W_k^i, B_k^i) to the initial parameters (C^i, W^i, B^i) without introducing too much error, thanks to the heuristic provided in Lemma 2.8 and Proposition 3.8. Then we remove the clipping in the evolution in subsection 7.2 by noticing the initial value $g_0^N(h)$ is L^2 integrable for all $h \in \mathcal{H}$. Finally, we establish in subsection 7.3 the weak convergence by using analysing the corresponding Poisson equation [27] of the Markov chain $(H_k)_{k\geq 0}$. The proof of the main result of this paper, Theorem 4.2, then follows in subsection 7.4.

7.1 Replacing the trained memory units

Recall that $g_t^N(h) = g_{\lfloor Nt\rfloor}^N(h;\theta_{\lfloor Nt\rfloor})$, and that we have now established

$$\left| g_t^N(h) - g_0^N(h) - \sum_{k=0}^{\lfloor Nt\rfloor - 1} \delta^{(1)} g_k^N(h) \right| \leq \frac{C_t}{N^{2-\beta-2\gamma}},$$

where

$$\delta^{(1)}g_k^N(h) = -\frac{\alpha}{N^2}(\psi^N(\hat{Y}_k^N) - Y_k)\sum_{i=1}^{N}\left[\hat{S}_{k+1}^{i,N}h(W_k^i) + (C_k^i)^2\Delta\hat{S}_{k+1}^i \nabla h(W_k^i)^\top X_k\right]. \tag{7.1}$$

In this section, we would like to show that one could study a simpler increment for the evolution of $g_t^N(h)$, in light of the a-priori bounds for the increments of parameters (Lemma 2.8) and our current analysis of the trained hidden memory units in appendix 3. We begin by showing that $\Psi^N(Y_k)$ could be replaced by $\Psi^N(g_k^N(h_k))$. This relies on the following lemma:

Lemma 7.1. *If $u, v \in \mathcal{H}$, then for all $k \leq \lfloor TN\rfloor$, there are constants C_T depending on T such that*

$$\mathbb{E}|g_k^N(u) - g_k^N(v)|^2 \leq C_T N^{2\gamma}\|u-v\|_{H^1(\lambda)}^2 + \frac{C_T}{N^{2-2\beta-4\gamma}}.$$

Proof. We start with

$$\begin{aligned}
&\left|(g^N_{k+1}(u) - g^N_{k+1}(v)) - (g^N_k(u) - g^N_k(v))\right| \\
&= \left|(g^N_{k+1}(u) - g^N_k(u)) - (g^N_{k+1}(v) - g^N_k(v))\right| \\
&= \left|\triangle g^N_k(u) - \triangle g^N_k(v)\right| \\
&\leq \left|\delta^{(1)} g^N_k(u) - \delta^{(1)} g^N_k(v)\right| + \left|\triangle g^N_k(u) - \delta^{(1)} g^N_k(u)\right| + \left|\delta^{(1)} g^N_k(v) - \triangle g^N_k(v)\right|,
\end{aligned}$$

so by Young's inequality,

$$\begin{aligned}
&\mathbb{E}\left|(g^N_{k+1}(u) - g^N_{k+1}(v)) - (g^N_k(u) - g^N_k(v))\right|^2 \\
&\leq 3\left[\mathbb{E}\left|\delta^{(1)} g^N_k(u) - \delta^{(1)} g^N_k(v)\right|^2 + \mathbb{E}\left|\triangle g^N_k(u) - \delta^{(1)} g^N_k(u)\right|^2 + \mathbb{E}\left|\delta^{(1)} g^N_k(v) - \triangle g^N_k(v)\right|^2\right], \\
&\leq 3\mathbb{E}\left|\delta^{(1)} g^N_k(u) - \delta^{(1)} g^N_k(v)\right|^2 + \frac{C_T}{N^{6-2\beta-4\gamma}}, \qquad (7.2)
\end{aligned}$$

with the constant C_T depends on the constant in the pre-limit equation in Lemma 4.1. We further note from the definition of $\delta^{(1)} g^N_k(\cdot)$ in equation (7.1) that

$$\begin{aligned}
&\left|\delta^{(1)} g^N_k(u) - \delta^{(1)} g^N_k(v)\right| \\
&\leq \frac{\alpha}{N^2}\left|\psi^N(\hat{Y}^N_k) - Y_k\right|\left|\sum_{i=1}^N \left[\hat{S}^{i,N}_{k+1}(u(W^i_k) - v(W^i_k)) + C^i_k \Delta\hat{S}^{i,N}_{k+1}(\nabla u(W^i_k) - \nabla v(W^i_k))^\top X_k\right]\right|,
\end{aligned}$$

so

$$\begin{aligned}
&\mathbb{E}\left|\delta^{(1)} g^N_k(u) - \delta^{(1)} g^N_k(v)\right|^2 \\
&\leq \frac{\alpha^2}{N^4}\mathbb{E}\left[\left|\psi^N(\hat{Y}^N_k) - Y_k\right|\left|\sum_{i=1}^N \left[\hat{S}^{i,N}_{k+1}(u(W^i_k) - v(W^i_k)) + C^i_k \Delta\hat{S}^{i,N}_{k+1}(\nabla u(W^i_k) - \nabla v(W^i_k))^\top X_k\right]\right|\right]^2 \\
&\leq \frac{C}{N^{4-2\gamma}}\mathbb{E}\left[\sum_{i=1}^N \left[\hat{S}^{i,N}_{k+1}(u(W^i_k) - v(W^i_k)) + C^i_k \Delta\hat{S}^{i,N}_{k+1}(\nabla u(W^i_k) - \nabla v(W^i_k))^\top X_k\right]\right]^2 \\
&\leq \frac{C}{N^{3-2\gamma}}\mathbb{E}\left[\sum_{i=1}^N \left[\left(\hat{S}^{i,N}_{k+1}(u(W^i_k) - v(W^i_k))\right)^2 + \left(C^i_k \Delta\hat{S}^{i,N}_{k+1}(\nabla u(W^i_k) - \nabla v(W^i_k))^\top X_k\right)^2\right]\right] \\
&\leq \frac{C_T}{N^{3-2\gamma}}\mathbb{E}\left[\sum_{i=1}^N \left[\left(u(W^i_k) - v(W^i_k)\right)^2 + \left|\nabla u(W^i_k) - \nabla v(W^i_k)\right|^2\right]\right] \\
&\leq \frac{C_T}{N^{3-2\gamma}}\mathbb{E}\Bigg[\sum_{i=1}^N \Big[\left(u(W^i_k) - u(W^i_0)\right)^2 + \left(v(W^i_k) - v(W^i_0)\right)^2 + \left(u(W^i_0) - v(W^i_0)\right)^2 \\
&\quad + \left|\nabla u(W^i_k) - \nabla u(W^i_0)\right|^2 + \left|\nabla v(W^i_k) - \nabla v(W^i_0)\right|^2 + \left|\nabla u(W^i_0) - \nabla v(W^i_0)\right|^2\Big]\Bigg] \\
&\leq \frac{C_T}{N^{3-2\gamma}}\mathbb{E}\Bigg[\sum_{i=1}^N \Big[\left|W^i_k - W^i_0\right|^2 + \left(u(W^i_0) - v(W^i_0)\right)^2 + \left|\nabla u(W^i_0) - \nabla v(W^i_0)\right|^2\Big]\Bigg] \\
&\leq \frac{C_T}{N^{2-2\gamma}}\left[\frac{C_T}{N^{2-2\beta-2\gamma}} + \|u - v\|^2_{H^1(\lambda)}\right] = \frac{C_T}{N^{2-2\gamma}}\|u - v\|^2_{H^1(\lambda)} + \frac{C_T}{N^{4-2\beta-4\gamma}}
\end{aligned}$$

Combining the above with equation (7.2) yields

$$\mathbb{E}\left|(g^N_{k+1}(u) - g^N_k(u)) - (g^N_{k+1}(v) - g^N_k(v))\right| \leq \frac{C_T}{N^{2-2\gamma}}\|u - v\|^2_{H^1(\lambda)} + \frac{C_T}{N^{4-2\beta-4\gamma}},$$

so by summing the telescoping sums:

$$\mathbb{E}\left|g_k^N(u)-g_k^N(v)\right|^2=\mathbb{E}\left|\sum_{j=0}^{k-1}\left(\triangle g_j^N(u)-\triangle g_j^N(v)\right)\right|^2\leq N\mathbb{E}\left[\sum_{j=0}^{k-1}\left(\triangle g_j^N(u)-\triangle g_j^N(v)\right)^2\right]$$
$$\leq C_T N^{2\gamma}\left\|u-v\right\|_{H^1(\lambda)}^2+\frac{C_T}{N^{2-2\beta-4\gamma}}.$$

□

This leads us to the following lemma.

Lemma 7.2. *Define*

$$\delta^{(2)}g_k^N(h)=-\frac{\alpha}{N^2}(\psi^N(g_k^N(h_{k+1}))-Y_k)\sum_{i=1}^N\left[\hat{S}_{k+1}^{i,N}h(W_k^i)+(C_k^i)^2\Delta\hat{S}_{k+1}^i\nabla h(W_k^i)^\top X_k\right]$$
$$=-\frac{\alpha}{N^2}(\psi^N(g_k^N(\varsigma(X_k,Z_k,Y_k,h_k))-Y_k)\sum_{i=1}^N\left[\hat{S}_{k+1}^{i,N}h(W_k^i)+(C_k^i)^2\Delta\hat{S}_{k+1}^i\nabla h(W_k^i)^\top X_k\right],$$

then for all $h\in\mathcal{H}$, the following holds for all $k\leq\lfloor NT\rfloor-1$

$$\mathbb{E}\left|\delta^{(1)}g_k^N(h)-\delta^{(2)}g_k^N(h)\right|^2\leq\frac{C_T}{N^{4-2\beta-4\gamma}}.$$

Proof. We first note the uniform bound in i for $k\leq\lfloor NT\rfloor$: thanks to the boundedness of the test function h, boundedness of activation function (see Assumption 2.3 and equations (6.1)-(6.2)) and the boundedness of C_k^i (see Assumption 2.5 and Lemma 2.8), one has

$$\hat{S}_{k+1}^{i,N}h(W_k^i)+(C_k^i)^2\Delta\hat{S}_{k+1}^{i,N}\nabla h(W_k^i)^\top X_k\leq C+C_TC_\sigma C=:C_T,$$

where the constant C_T may change from line to line. By the boundedness of derivative of ψ^N,

$$\left|\delta^{(1)}g_k^N(h)-\delta^{(2)}g_k^N(h)\right|\leq\frac{C_T}{N}\left|\psi^N(\hat{Y}_k^N)-\psi^N(g_k^N(h_{k+1}))\right|\leq\frac{C_T}{N}\left|\hat{Y}_k^N-g_k^N(h_{k+1})\right|.\tag{7.3}$$

Observing that $\hat{Y}_k^N=g_k^N(v_{k+1}^N)$, and that we know a-priori that the h_k and v_k^N has output bounded by 1 and partial derivatives bounded by C_σ, we may invoke Lemma 7.1 with our a-priori bound 7.3 and Proposition 3.8 (see also equation (3.18)) to bound our required expectation:

$$\begin{aligned}\mathbb{E}\left|\delta^{(1)}g_k^N(h)-\delta^{(2)}g_k^N(h)\right|^2&\overset{(7.3)}{\leq}\frac{C_T}{N^2}\mathbb{E}\left[g_k^N(v_{k+1}^N)-g_k^N(h_{k+1})\right]^2\\&=\frac{C_T}{N^2}\mathbb{E}\left[\mathbb{E}\left[\left(g_k^N(v_{k+1}^N)-g_k^N(h_{k+1})\right)^2\,\middle|\,v_{k+1}^N,h_{k+1}\right]\right]\\&\overset{(7.1)}{\leq}\frac{C_T}{N^2}\mathbb{E}\left[C_TN^{2\gamma}\left\|v_{k+1}^N-h_{k+1}\right\|_{H^1(\lambda)}^2+\frac{C_T}{N^{2-2\beta-4\gamma}}\right]\\&\overset{(3.18)}{\leq}\frac{C_T}{N^{4-2\beta-4\gamma}}\end{aligned}\tag{7.4}$$

The result holds by recalling that $h_{k+1}=\varsigma(X_k,Z_k,Y_k,h_k)$. □

We could then sum up all the differences between the increments to conclude that

$$\begin{aligned}
&\mathbb{E}\left|g_t^N(h) - g_0^N(h) - \sum_{k=0}^{\lfloor Nt\rfloor-1} \delta^{(2)} g_k^N(h)\right|^2 \\
&\leq 2\mathbb{E}\left[\left(g_t^N(h) - g_0^N(h) - \sum_{k=0}^{\lfloor Nt\rfloor-1} \delta^{(1)} g_k^N(h)\right)^2 + \left(\sum_{k=0}^{\lfloor Nt\rfloor-1}\left(\delta^{(1)} g_k^N(h) - \delta^{(2)} g_k^N(h)\right)\right)^2\right] \\
&\leq \frac{C_T}{N^{4-2\beta-4\gamma}} + 2N \sum_{k=0}^{\lfloor Nt\rfloor-1} \mathbb{E}\left[\delta^{(1)} g_k^N(h) - \delta^{(2)} g_k^N(h)\right]^2 \leq \frac{C_T}{N^{2-2\beta-4\gamma}}.
\end{aligned}$$

To understand the next step, we revisit the formula for $\delta^{(2)} g_k^N(h)$:

$$\begin{aligned}
\delta^{(2)} g_k^N(h) &= -\frac{\alpha}{N^2}(\psi^N(g_k^N(h_{k+1})) - Y_k) \sum_{i=1}^N \left[\hat{S}_{k+1}^{i,N} h(W_k^i) + (C_k^i)^2 \Delta \hat{S}_{k+1}^i \nabla h(W_k^i)^\top X_k\right] \\
&= -\frac{\alpha}{N^2}(\psi^N(g_k^N(h_{k+1})) - Y_k) \sum_{i=1}^N \left[v_{k+1}^N(W_k^i) h(W_k^i) + (C_k^i)^2 \sigma'\left((W_k^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\right) \nabla h(W_k^i)^\top X_k\right] \\
&= -\frac{\alpha}{N}(\psi^N(g_k^N(h_{k+1})) - Y_k) \left\langle v_{k+1}^N(w) h(w) + c^2 \sigma'\left(w^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\right) \nabla h(w)^\top X_k,\, \lambda_k^N\right\rangle
\end{aligned}$$

In light of the results proven regarding the increments of parameters (Lemma 2.8) and the trained memory units (Section 3), we would expect that one could study the new increments with the average of empirical distribution of updated parameters λ_k^N replaced by an average with respect to the initial parameter distribution λ. This is formalised by the following technical lemma:

Lemma 7.3. *Define*

$$\begin{aligned}
\delta^{(3)} g_k^N(h) &= -\frac{\alpha}{N}(\psi^N(g_k^N(h_{k+1})) - Y_k) \left\langle h_{k+1}(w) h(w) + c^2 \sigma'\left(w^\top X_k + \langle b' h_k(w'), \lambda\rangle\right) \nabla h(w)^\top X_k,\, \lambda\right\rangle, \\
&= -\frac{\alpha}{N}(\psi^N(g_k^N(\varsigma(X_k, Z_k, Y_k, h_k))) - Y_k) \left\langle h_{k+1}(w) h(w) + c^2 \sigma'\left(w^\top X_k + \langle b' h_k(w'), \lambda\rangle\right) \nabla h(w)^\top X_k,\, \lambda\right\rangle
\end{aligned}$$

then for all $h \in \mathcal{H}$, the following holds for all $k \leq \lfloor NT\rfloor - 1$

$$\mathbb{E}\left|\delta^{(2)} g_k^N(h) - \delta^{(3)} g_k^N(h)\right|^2 \leq \frac{C_T}{N^{4-2\beta-4\gamma}}.$$

Proof. We break down the fluctuation term into different components

$$\delta^{(2)} g_k^N(h) - \delta^{(3)} g_k^N(h) = -\frac{\alpha}{N}(\psi^N(g_k^N(h_{k+1})) - Y_k) \sum_{\bullet=1}^6 M_k^{\bullet,N},$$

where

$$
\begin{aligned}
M_k^{1,N} &= \left\langle v_{k+1}^N(w)h(w),\, \lambda_k^N - \lambda^N \right\rangle \\
M_k^{2,N} &= \left\langle c^2\sigma'\left(w^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\right)\nabla h(w)^\top X_k,\, \lambda_k^N - \lambda^N \right\rangle \\
M_k^{3,N} &= \left\langle (v_{k+1}^N(w) - h_{k+1}(w))h(w),\, \lambda^N \right\rangle \\
M_k^{4,N} &= \left\langle c^2\left(\sigma'\left(w^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\right) - \sigma'\left(w^\top X_k + \langle b' h_k(w'),\, \lambda\rangle\right)\right)\nabla h(w)^\top X_k,\, \lambda^N \right\rangle \\
M_k^{5,N} &= \left\langle h_{k+1}(w)h(w), \lambda^N - \lambda \right\rangle \\
M_k^{6,N} &= \left\langle c^2\sigma'\left(w^\top X_k + \langle b' h_k(w'),\, \lambda\rangle\right)\nabla h(w)^\top X_k, \lambda^N - \lambda \right\rangle
\end{aligned}
$$

This implies

$$
\mathbb{E}\left|\delta^{(2)} g_k^N(h) - \delta^{(3)} g_k^N(h)\right|^2 \le \frac{\alpha^2}{N^2}\mathbb{E}\left[\left|\psi^N(g_k^N(h_{k+1})) - Y_k\right|^2\left(\sum_{\bullet=1}^6 M_k^{\bullet,N}\right)^2\right] \le \frac{6C}{N^{2-2\gamma}}\left(\sum_{\bullet=1}^6 \mathbb{E}\left[M_k^{\bullet,N}\right]^2\right)
$$

Now note that:

$$
\begin{aligned}
\left|M_k^{1,N}\right| &= \left|\frac{1}{N}\sum_{i=1}^N \left(v_{k+1}^N(W_k^i)h(W_k^i) - v_{k+1}^N(W_0^i)h(W_0^i)\right)\right| \\
&\le \frac{1}{N}\sum_{i=1}^N \left|v_{k+1}^N(W_k^i)h(W_k^i) - v_{k+1}^N(W_0^i)h(W_0^i)\right| \\
&\le \frac{1}{N}\sum_{i=1}^N \left[\left|v_{k+1}^N(W_k^i)(h(W_k^i) - h(W_0^i))\right| + \left|(v_{k+1}^N(W_k^i) - v_{k+1}^N(W_0^i))h(W_0^i)\right|\right] \\
&\le \frac{C}{N}\sum_{i=1}^N \left|W_k^i - W_0^i\right| \overset{(2.4)}{\le} \frac{C_T}{N^{1-\beta-\gamma}}
\end{aligned} \tag{7.5}
$$

$$
\begin{aligned}
\left|M_k^{2,N}\right| &= \left|\frac{1}{N}\sum_{i=1}^N \left((C_k^i)^2\sigma'\left((W_k^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\right)\nabla h(W_k^i)^\top X_k\right.\right.\\
&\quad \left.\left. - (C_0^i)^2\sigma'\left((W_0^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\right)\nabla h(W_0^i)^\top X_k\right)\right|\\
&\le \frac{1}{N}\sum_{i=1}^N\left[\left|(C_k^i)^2-(C_0^i)^2\right|\underbrace{\left|\sigma'((W_k^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\nabla h(W_k^i)^\top X_k)\right|}_{\le C}\right.\\
&\quad + \underbrace{(C_0^i)^2}_{\le 1}\left|\sigma'\Big((W_k^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\Big)\nabla h(W_k^i)^\top X_k\right.\\
&\quad \left.\left. - \sigma'\Big((W_0^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\Big)\nabla h(W_0^i)^\top X_k\right|\right]\\
&\le \frac{1}{N}\sum_{i=1}^N\left[\left|C_k^i - C_0^i\right|\underbrace{\left|C_k^i + C_0^i\right|}_{\le C_T}\right.\\
&\quad + \underbrace{\left|\sigma'\Big((W_k^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\Big) - \sigma'\Big((W_0^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\Big)\right|}_{\le C_\sigma|W_k^i - W_0^i|}\underbrace{\left|\nabla h(W_k^i)^\top X_k\right|}_{\le C}\\
&\quad \left. + \underbrace{\left|\sigma'\Big((W_0^i)^\top X_k + \frac{1}{N}\sum_{j=1}^N B_k^j v_k^N(W_{k-1}^i)\Big)\right||X_k|}_{\le dC_\sigma}\underbrace{\left|\nabla h(W_k^i) - \nabla h(W_0^i)\right|}_{\le Cd^2|W_k^i - W_0^i|}\right]\\
&\le \frac{C_T}{N}\sum_{i=1}^N\left[|C_k^i - C_0^i| + |W_k^i - W_0^i|\right] \overset{(2.4)}{\le} \frac{C_T}{N^{1-\beta-\gamma}}.
\end{aligned}
\tag{7.6}
$$

Furthermore,

$$
\left|M_k^{3,N}\right| \le \frac{1}{N}\sum_{i=1}^N \left|v_{k+1}^N(W_0^i) - h_{k+1}(W_0^i)\right|\left|h(W_0^i)\right| \le \frac{C}{N}\sum_{i=1}^N\left|\Gamma_{k+1}^N(W_0^i)\right|,
$$

where $\Gamma_k^N(w) = v_k^N(w) - h_k(w)$. By Proposition 3.8, and recalling the definitions of $E^{N,1}, E^{N,2}$, we have

$$
\begin{aligned}
\mathbb{E}\left|M_k^{3,N}\right|^2 &\le \frac{C}{N}\sum_{i=1}^N \mathbb{E}\left|\Gamma_{k+1}^N(W_0^i)\right|^2\\
&\overset{(6.12)}{\le} C\left[\mathbb{E}\left[E_{k+1}^{N,1}\right]^2 + \mathbb{E}\left[E_{k+1}^{N,2}\right]^2\right]\\
&\overset{(6.13),(6.14)}{\le} C\left[\frac{2}{N} + 2C_\sigma^2\mathbb{E}\left(E_{k-1}^{N,1}\right)^2 + \frac{2C_T}{N^{2-2\beta-2\gamma}} + 2C_\sigma^2\mathbb{E}\left(E_{k-1}^{N,2}\right)^2\right]\\
&\overset{(3.18)}{\le} \frac{C_T}{N^{2-2\beta-2\gamma}}.
\end{aligned}
$$

Next, we have

$$
\left|M_k^{4,N}\right| \le C\left[\left|E_k^{N,1}\right| + \left|E_k^{N,2}\right|\right] \implies \mathbb{E}\left[M_k^{4,N}\right]^2 \overset{(3.18)}{\le} \frac{C_T}{N^{2-2\beta-2\gamma}}. \tag{7.7}
$$

The remaining two terms could be easily bounded using similar arguments in establishing equation (6.13). Formally, we note that $h_{k+1}(W_0^i)h(W_0^i) \mid h_{k+1}$ is conditionally mutually iid for $i = 1, \cdots, N$, so

$$\begin{aligned}
\mathbb{E}\left|M_k^{5,N}\right|^2 &= \mathbb{E}\left(\left\langle h_{k+1}(w)h(w), \lambda^N - \lambda\right\rangle^2\right) \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle h_{k+1}(w)h(w), \lambda^N - \lambda\right\rangle^2 \,\middle|\, h_{k+1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \left(h_{k+1}(W_0^i)h(W_0^i) - \int h_{k+1}(w')h(w')\,\lambda(db'\,dw')\right)\right)^2 \,\middle|\, h_{k+1}\right]\right] \\
&= \frac{1}{N}\mathbb{E}\left[\mathbb{E}\left[\left(h_{k+1}(W^1)h(W^1)\right)^2 \mid h_{k+1}\right] - \left(\int h_{k+1}(w')h(w')\,\lambda(db'\,dw')\right)^2\right] \\
&\leq \frac{C}{N}.
\end{aligned}$$

Noting also that $(C_0^i)^2\sigma'\left((W_0^i)^\top X_k + \langle b' h_k(w'),\, \lambda\rangle\right)\nabla h(W_0^i)^\top X_k \,|\, h_k, X_k$ is conditionally mutually iid, we can apply the exact same argument to show that

$$\mathbb{E}\left|M_k^{6,N}\right|^2 = \frac{\mathbb{E}\left[\left((C_0^1)^2\sigma'\left((W_0^1)^\top X_k + \langle b' h_k(w'),\, \lambda\rangle\right)\nabla h(W_0^1)^\top X_k\right)\right]^2}{N} \leq \frac{C}{N}.$$

Finally, since $\beta > 1/2$ and $\gamma > 0$, we have $2-2\beta-2\gamma < 1-2\gamma < 1$, so $C_T N^{-1} = o(N^{-(2-2\beta-2\gamma)})$. Therefore, one could sum up all fluctuation term to have

$$\mathbb{E}\left|\delta^{(2)}g_k^N(h) - \delta^{(3)}g_k^N(h)\right|^2 \leq \frac{C}{N^{2-2\gamma}}\frac{C_T}{N^{2-2\beta-2\gamma}} = \frac{C_T}{N^{4-2\beta-4\gamma}}.$$

Arguing as before, one could now conclude that

$$\mathbb{E}\left|g_t^N(h) - g_0^N(h) - \sum_{k=0}^{\lfloor Nt\rfloor-1}\delta^{(3)}g_k^N(h)\right|^2 \leq \frac{C_T}{N^{2-2\beta-4\gamma}}. \tag{7.8}$$

□

7.2 Removal of the clipping function

So far, we have shown that one could study the process $g_m^N(h)$ by looking at the increment $\delta^{(3)}g_k^N$. To simplify our computation, let us recall the fact that

$$h_{k+1}(w) = [\varsigma(X_k, Z_k, Y_k, h_k)](w) = \sigma(w^\top X_k + \langle b' h_k(w'),\, \lambda\rangle),$$

and define the kernel $\mathcal{K}_{x,\lambda}(\cdot,\cdot) : \mathcal{H}\times\mathcal{H} \to \mathbb{R}$ for all x:

$$\mathcal{K}_{x,\lambda}(h, \mathsf{h}) = \left\langle \sigma(w^\top x + \langle b'\mathsf{h}(w')\rangle)h(w) + c^2\sigma'(w^\top x + \langle b'\mathsf{h}(w'), \lambda\rangle)\nabla h(w)^\top x,\, \lambda\right\rangle,$$

then

$$\begin{aligned}
\delta^{(3)}g_k^N(h) &= -\frac{\alpha}{N}(\psi^N(g_k^N(h_{k+1})) - Y_k)\left\langle h_{k+1}(w)h(w) + c^2\sigma'\left(w^\top X_k + \langle b' h_k(w'), \lambda\rangle\right)\nabla h(w)^\top X_k,\, \lambda\right\rangle, \\
&\overset{(3.6)}{=} -\frac{\alpha}{N}(\psi^N(g_k^N(h_{k+1})) - Y_k)\left\langle \sigma(w^\top X_k + \langle b' h_k(w'),\, \lambda\rangle)h(w) + c^2\sigma'\left(w^\top X_k + \langle b' h_k(w'), \lambda\rangle\right)\nabla h(w)^\top X_k,\, \lambda\right\rangle \\
&= -\frac{\alpha}{N}(\psi^N(g_k^N(\varsigma(X_k, Z_k, Y_k, h_k))) - Y_k)\,\mathcal{K}_{X_k,\lambda}(h, h_k).
\end{aligned} \tag{7.9}$$

We have assumed that $|(X_k, Z_k)| \leq 1$ in Assumption 2.1, which implies $|X_k| \leq 1$, so we could limit our discussion of the kernel $\mathcal{K}_{x,\lambda}$ for $|x| \leq 1$. By Assumption 2.3 on activation function $\sigma(\cdot)$ and the space $\mathcal{H}$, we have

$$|\mathcal{K}_{x,\lambda}(h, \mathsf{h})| \leq C.$$

The main objective of this section is to remove the clipping function in $\delta^{(3)} g_k^N(h)$, and set the initial value of the output evolution to zero, without introducing too much error.

Lemma 7.4 (Uniform integrability of $g_m^N(h_{m+1})$)**.** *For $m \leq \lfloor NT \rfloor - 1$, we have $\mathbb{E}\left[g_m^N(h_{m+1})\right]^2 \leq C_T$.*

Proof. Using the independence and identical distribution of (C_0^i, W_0^i), for $i = 1, \cdots, N$ and the fact that C_0^i and W_0^i are independent from each other with C_0^i having zero mean, we have

$$\begin{aligned}
\mathbb{E}|g_0^N(h_1)|^2 &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N^\beta}\sum_{i=1}^N C_0^i \sigma((W_0^i)^\top X_0)\right)^2 \,\middle|\, X_0\right]\right] \\
&= \mathbb{E}\left[\frac{1}{N^{2\beta}}\sum_{i=1}^N \mathbb{E}\left[\left(C_0^i \sigma((W_0^i)^\top X_0)\right)^2 \,\middle|\, X_0\right]\right] \\
&\leq \mathbb{E}\left[\frac{1}{N^{2\beta}}\sum_{i=1}^N 1\right] = \frac{1}{N^{2\beta-1}}
\end{aligned}$$

Therefore by (7.8) we have

$$\begin{aligned}
\mathbb{E}|g_m^N(h_{m+1})|^2 &\leq 2\mathbb{E}\left|\sum_{k=0}^{m-1} \delta^{(3)} g_k^N(h_{m+1})\right|^2 + \frac{C_T}{N^{2-2\beta-4\gamma}} \\
&\leq 2m \sum_{k=0}^{m-1} \mathbb{E}\left[\delta^{(3)} g_k^N(h_{m+1})\right]^2 + \frac{C_T}{N^{2-2\beta-4\gamma}} \\
&\leq \frac{2CT}{N}\sum_{k=0}^{m-1} \mathbb{E}\left|\psi^N(g_k^N(h_{k+1})) - Y_k\right|^2 + \frac{C_T}{N^{2-2\beta-4\gamma}} \\
&\leq \frac{2CT}{N}\sum_{k=0}^{m-1} 2\left[\mathbb{E}\left|\psi^N(g_k^N(h_{k+1}))\right|^2 + \mathbb{E}\left|Y_k\right|^2\right] + \frac{C_T}{N^{2-2\beta-4\gamma}} \\
&\leq \frac{4CT}{N}\sum_{k=0}^{m-1} \left[\mathbb{E}\left|g_k^N(h_{k+1})\right|^2 + C_y^2\right] + \frac{C_T}{N^{2-2\beta-4\gamma}} \\
&\leq \frac{4CT}{N}\sum_{k=0}^{m-1} \left[\mathbb{E}\left|g_k^N(h_{k+1})\right|^2\right] + 4CC_y^2 T + \frac{C_T}{N^{2-2\beta-4\gamma}} \\
&\leq \frac{4CT}{N}\sum_{k=0}^{m-1} \mathbb{E}\left|g_k^N(h_{k+1})\right|^2 + C_T
\end{aligned}$$

We may then use discrete Gronwall's inequality to conclude that

$$\mathbb{E}|g_m^N(h_{m+1})|^2 \leq C_T \exp\left(\frac{4CmT}{N}\right) \leq C_T.$$

□

Because the expectation $\mathbb{E}|g_m^N(h_{m+1})|^2$ is uniformly bounded for $m \leq \lfloor NT \rfloor$, we may invoke Markov's inequality to prove the following:

Lemma 7.5 (Removal of clipping function in $\delta^{(3)}$)**.** *Define*

$$\delta^{(4)} g_k^N(h) = \frac{\alpha}{N}(g_k^N(h_{k+1}) - Y_k)\mathcal{K}_{X_k,\lambda}(h, \mathsf{h}),$$

then for all $h \in \mathcal{H}$, the following holds for all $k \leq \lfloor NT \rfloor$

$$\mathbb{E}\left|\delta^{(3)} g_k^N(h) - \delta^{(4)} g_k^N(h)\right| \leq \frac{C_T}{N^{1+\gamma}}.$$

Proof. Note that

$$\left|\delta^{(3)}g_k^N(h)-\delta^{(4)}g_k^N(h)\right|=\frac{\alpha}{N}\left|\psi^N(g_k^N(h_{k+1}))-g_k^N(h_{k+1})\right|\left|\mathcal{K}_{X_k,\lambda}(h,h_k)\right|\leq\frac{C}{N}\left|\psi^N(g_k^N(h_{k+1}))-g_k^N(h_{k+1})\right|$$

Recall that when $|g_k^N(h_{k+1})|\leq N^\gamma$ then $\psi^N(g_k^N(h_{k+1}))=g_k^N(h_{k+1})$, so

$$\begin{aligned}\psi^N(g_k^N(h_{k+1}))-g_k^N(h_{k+1})&=(\psi^N(g_k^N(h_{k+1}))-g_k^N(h_{k+1}))\left(\mathbf{1}_{\left\{|g_k^N(h_{k+1})|>N^\gamma\right\}}+\mathbf{1}_{\left\{|g_k^N(h_{k+1})|\leq N^\gamma\right\}}\right)\\&=(\psi^N(g_k^N(h_{k+1}))-g_k^N(h_{k+1}))\mathbf{1}_{\left\{|g_k^N(h_{k+1})|>N^\gamma\right\}}.\end{aligned}$$

Combining with $|\psi^N(x)|\leq|x|$, we have

$$\begin{aligned}\mathbb{E}\left|\delta^{(3)}g_k^N(h)-\delta^{(4)}g_k^N(h)\right|&\leq\frac{C}{N}\mathbb{E}\left|\psi^N(g_k^N(h_{k+1}))-g_k^N(h_{k+1})\right|\mathbf{1}_{\left\{|g_k^N(h_{k+1})|>N^\gamma\right\}}\\&\leq\frac{C}{N}\mathbb{E}\left[\left|\psi^N(g_k^N(h_{k+1}))\right|+\left|g_k^N(h_{k+1})\right|\right]\mathbf{1}_{\left\{|g_k^N(h_{k+1})|>N^\gamma\right\}}\\&\leq\frac{2C}{N}\mathbb{E}\left|g_k^N(h_{k+1})\right|\mathbf{1}_{\left\{|g_k^N(h_{k+1})|>N^\gamma\right\}}\\&\leq\frac{2C}{N}\sqrt{\mathbb{E}\left|g_k^N(h_{k+1})\right|^2}\sqrt{\mathbb{E}\left[\mathbf{1}_{\left\{|g_k^N(h_{k+1})|>N^\gamma\right\}}\right]}\\&\leq\frac{C_T}{N}\sqrt{\frac{\mathbb{E}\left|g_k^N(h_{k+1})\right|^2}{N^{2\gamma}}}=\frac{C_T}{N^{1+\gamma}}.\end{aligned}$$

□

Arguing as before, we have

$$\mathbb{E}\left|g_t^N(h)-g_0^N(h)-\sum_{k=0}^{\lfloor Nt\rfloor-1}\delta^{(4)}g_k^N(h)\right|\leq\frac{C_T}{N^{(1-\beta-2\gamma)\wedge\gamma}},\tag{7.10}$$

We now record a simple observation regarding the boundeness of the second moment of $g_0^N(h)$.

Lemma 7.6. *For all $h\in\mathcal{H}$,*

$$\mathbb{E}\left[g_0^N(h)\right]^2\leq\frac{1}{N^{2\beta-1}}$$

Proof. Using the independence and identical distribution of (C_0^i,W_0^i), for $i=1,\cdots,N$ and the fact that C_0^i and W_0^i are independent from each other with C_0^i having zero mean, we have

$$\mathbb{E}\left[g_0^N(h)\right]^2=\frac{1}{N^{2\beta}}\mathbb{E}\left[\sum_{i=1}^N C_0^ih(W_0^i)\right]^2=\frac{\mathbb{E}\left[C_0^1h(W_0^1)\right]^2}{N^{2\beta-1}}\leq\frac{1}{N^{2\beta-1}}$$

□

Remark 7.7. The above is also true if we replace the fixed h with the elements in random sequence $(h_k)_{k\geq0}$, as the sequence h_k is independent of the initialisation $(C_0^i,W_0^i)_{i=1}^N$.

This allows us to consider the new evolution that $\varphi_k^N(h)$ that satisfies the evolution equation

$$\begin{aligned}\varphi_m^N(h)&=\sum_{k=0}^{m-1}\triangle\varphi_k^N(h),\quad\varphi_0^N(h)=0\\\triangle\varphi_k^N(h)&=-\frac{\alpha}{N}(\varphi_k^N(h_{k+1})-Y_k)\mathcal{K}_{X_k,\lambda}(h,h_k).\\&=-\frac{\alpha}{N}(\varphi_k^N(\varsigma(X_k,Z_k,Y_k,h_k))-Y_k)\mathcal{K}_{X_k,\lambda}(h,h_k).\end{aligned}\tag{7.11}$$

Lemma 7.8 ($\varphi_m^N(h)$ approximates $g_m^N(h)$)**.** *For all $m \le \lfloor NT \rfloor$ and $h \in \mathcal{H}$,*

$$\sup_{h\in\mathcal{H}} \mathbb{E}\left|g_m^N(h) - \varphi_m^N(h)\right| \le \frac{C_T}{N^{(1-\beta-2\gamma)\wedge\gamma\wedge(\beta-1/2)}}. \tag{7.12}$$

Proof. In light of Lemma 7.6, we know that

$$\begin{aligned}
\mathbb{E}\left|g_m^N(h) - \varphi_m^N(h)\right| &\le \mathbb{E}\left[g_m^N(h) - g_0^N(h) - \varphi_m^N(h)\right] + \mathbb{E}\left|g_0^N(h)\right| \\
&\le \mathbb{E}\left|\sum_{k=0}^{m-1}(\delta^{(4)}g_k^N(h) - \triangle\varphi_k^N(h))\right| + \frac{1}{N^{\beta-1/2}} + \frac{C_T}{N^{1-\beta-2\gamma}} \\
&\le \frac{C}{N}\sum_{k=0}^{m-1}\mathbb{E}\left|g_k^N(h_{k+1}) - \varphi_k^N(h_{k+1})\right| + \frac{1}{N^{\beta-1/2}} + \frac{C_T}{N^{(1-\beta-2\gamma)\wedge\gamma}}.
\end{aligned} \tag{7.13}$$

The above inequality is also true when replacing $h = h_{m+1}$, in particular

$$\mathbb{E}\left|g_m^N(h_{m+1}) - \varphi_m^N(h_{m+1})\right| \le \frac{C}{N}\sum_{k=0}^{m-1}\mathbb{E}\left|g_k^N(h_{k+1}) - \varphi_k^N(h_{k+1})\right| + \frac{1}{N^{\beta-1/2}} + \frac{C_T}{N^{(1-\beta-2\gamma)\wedge\gamma}}; \tag{7.14}$$

so by discrete Gronwall's inequality,

$$\mathbb{E}\left|g_m^N(h_{m+1}) - \varphi_m^N(h_{m+1})\right| \le \frac{C_T\exp(Cm/N)}{N^{(1-\beta-2\gamma)\wedge\gamma\wedge(\beta-1/2)}} \le \frac{C_T}{N^{(1-\beta-2\gamma)\wedge\gamma\wedge(\beta-1/2)}}. \tag{7.15}$$

We can then plug in (7.15) into (7.13) to obtain our desired result. □

With the absence of initialisation in the sequence $\varphi_m^N(h)$, one can prove that $\varphi_m^N(h)$ is surely bounded for any $h \in \mathcal{H}$.

Lemma 7.9. *For $m \le \lfloor NT \rfloor$ and $h \in \mathcal{H}$, $|\varphi_m^N(h)| \le C_T$ surely. Moreover, $\varphi_m^N(h)$ is C_T globally Lipschitz with respect to the $H^1(\lambda)$ norm whenever $m \le \lfloor NT \rfloor$. Finally, consider the map for $\mathsf{H} = (\mathsf{x}, \mathsf{z}, \mathsf{y}, \mathsf{h})$*

$$G^{m,h} : \mathsf{H} \in \mathcal{X} \mapsto -\alpha(\varphi_m^N(\varsigma(\mathsf{H})) - \mathsf{y})\mathcal{K}_{\mathsf{x},\lambda}(h, \mathsf{h}),$$

so that

$$N \times \triangle\varphi_m^N(h) = G^{m,h}(X_m, Z_m, Y_m, h_m). \tag{7.16}$$

Then for fixed $h \in \mathcal{H}$ and $m \le \lfloor NT \rfloor$,

- *$\int |G^{m,h}|\, d\mu < +\infty$, where μ is the invariant measure of chain (H_k) as induced in Theorem 3.6, and*
- *$G^{m,h}$ is C_T-Lipschitz in $\mathcal{X}$.*

Proof. From the evolution equation, one sees that

$$\left|\varphi_m^N(h)\right| \le \frac{C}{N}\sum_{k=0}^{m-1}(\left|\varphi_k^N(h_{k+1})\right| + C), \tag{7.17}$$

The above equation is true for $h = h_{k+1}$, so by discrete Gronwall's inequality one has

$$\left|\varphi_m^N(h_{m+1})\right| \le C\exp(Cm/N) \le Ce^{CT} =: C_T.$$

Substituting this into (7.17) yields our estimate $|\varphi_m^N(h)| \le C$. To show the global Lipschitzness of $\varphi_m^N(\cdot)$, we note that for any $h, \tilde{h} \in H^1(\lambda)$

$$\begin{aligned}
|\varphi_m^N(h) - \varphi_m^N(\tilde{h})| &\le \sum_{k=0}^{m}\frac{\alpha}{N}|\varphi_k^N(h_{k+1}) - Y_k|\int\left[|h(w) - \tilde{h}(w)| + |\nabla h(w) - \nabla\tilde{h}(w)|\right]\lambda(dw) \\
&\le C_T\left\|h - \tilde{h}\right\|_{H^1(\lambda)}.
\end{aligned} \tag{7.18}$$

It is then immediately true that $\int |G^{m,h}(\mathsf{x},\mathsf{z},\mathsf{y},\mathsf{h})|\,\mu(d\mathsf{x},d\mathsf{z},d\mathsf{y},d\mathsf{h}) < +\infty$. Finally we recall that if $(x,z,y,h)\in\mathcal{X}$ then $|x|\leq 1$ and $|y|\leq C_y$, and that $|\mathcal{K}_{x,\lambda}(h,\mathsf{h})|\leq C$. Therefore, for any $\mathsf{H}=(\mathsf{x},\mathsf{z},\mathsf{y},\mathsf{h}),\tilde{\mathsf{H}}=(\tilde{\mathsf{x}},\tilde{\mathsf{z}},\tilde{\mathsf{y}},\tilde{\mathsf{h}})$ with $|\mathsf{x}|\vee|\tilde{\mathsf{x}}|\leq 1$ and $|\mathsf{y}|\vee|\tilde{\mathsf{y}}|\leq C_y$,

$$\begin{aligned}
|\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})-\mathcal{K}_{\tilde{\mathsf{x}},\lambda}(h,\tilde{\mathsf{h}})| &\leq \alpha\Big[\Big\langle |h(w)||\sigma(w^\top\mathsf{x}+\langle b'\mathsf{h}(w'),\lambda\rangle)-\sigma(w^\top\tilde{\mathsf{x}}+\langle b'\tilde{\mathsf{h}}(w'),\,\lambda\rangle)|,\,\lambda\Big\rangle \\
&\quad+\Big\langle c^2|\nabla h(w)||\sigma'(w^\top\mathsf{x}+\langle b'\mathsf{h}(w'),\lambda\rangle)-\sigma'(w^\top\tilde{\mathsf{x}}+\langle b'\tilde{\mathsf{h}}(w'),\,\lambda\rangle)|,\,\lambda\Big\rangle\Big] \\
&\leq 4\alpha C_\sigma\left[\int w^\top(\mathsf{x}-\tilde{\mathsf{x}})\,\lambda(dw)+\int|\mathsf{h}(w')-\tilde{\mathsf{h}}(w')|\,\lambda(dw')\right] \\
&\leq 4\alpha C_\sigma\left[|\mathsf{x}-\tilde{\mathsf{x}}|+\|\mathsf{h}-\tilde{\mathsf{h}}\|_{H^1(\lambda)}\right].
\end{aligned}$$

We also recall the controls from inequality (6.5)

$$\begin{aligned}
\|[\varsigma(\mathsf{H})](w)-\varsigma(\tilde{\mathsf{H}})](w)\|^2_{H^1(\lambda)} &= \Big\|\sigma(w^\top\mathsf{x}+\langle b'\mathsf{h}(w'),\lambda\rangle)-\sigma(w^\top\tilde{\mathsf{x}}+\langle b'\tilde{\mathsf{h}}(w'),\lambda\rangle)\Big\|^2_{H^1(\lambda)} \\
&\leq 8C_\sigma^2\left[|\mathsf{x}-\tilde{\mathsf{x}}|^2+\|\mathsf{h}-\tilde{\mathsf{h}}\|^2_{H^1(\lambda)}\right], \\
&\leq 8C_\sigma^2\|\mathsf{H}-\tilde{\mathsf{H}}\|^2 \qquad (7.19)
\end{aligned}$$

so by (7.18) we have

$$\left|\varphi_m^N(\varsigma(\mathsf{H}))-\varphi_m^N(\varsigma(\tilde{\mathsf{H}}))\right|^2\leq C_T\|\mathsf{H}-\tilde{\mathsf{H}}\|^2_{\mathcal{X}},$$

This yields the following control

$$\begin{aligned}
\implies |G^{m,h}(\mathsf{H})-G^{m,h}(\tilde{\mathsf{H}})| &\leq \alpha|\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})|\left[\left|\varphi_m^N(\varsigma(\mathsf{H}))-\varphi_m^N(\varsigma(\tilde{\mathsf{H}}))\right|+|\mathsf{y}-\tilde{\mathsf{y}}|\right] \\
&\quad+\alpha|\varphi_m^N(\varsigma(\tilde{\mathsf{H}}))-\tilde{\mathsf{y}}||\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})-\mathcal{K}_{\tilde{\mathsf{x}},\lambda}(h,\tilde{\mathsf{h}})| \\
&\leq C_T\|\mathsf{H}-\tilde{\mathsf{H}}\|_{\mathcal{X}},
\end{aligned}$$

which completes our proof. In particular, the Lipschitz constant for $G^{0,h}$ is independent from T as $\varphi_0^N(\cdot)$ is set to zero. □

7.3 Weak convergence analysis

We study the difference between the random evolution

$$\varphi_m^N(h)=\sum_{k=0}^{m-1}\triangle\varphi_k^N(h)$$

and the evolution

$$\bar{\varphi}_m^N(h)=\sum_{k=0}^{m-1}\delta\varphi_k^N(h),$$

where

$$\delta\varphi_k^N(h)=-\frac{\alpha}{N}\int_{\mathcal{X}}(\varphi_k^N(\varsigma(\mathsf{H}))-\mathsf{y})\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{H}).$$

We study this difference by first constructing the associated Poisson equation of the chain of memories [27]. We write the transition kernel of the Markov chain $(H_k)_{k\geq 1}$ as P, and recall that the Markov chain admits a limiting invariant measure μ.

Definition 7.10. Given $G : \mathcal{X} \to \mathbb{R}$ a measurable function with $\int |G| d\mu < +\infty$, the Poisson equation is a functional equation on $\hat{G}$

$$\hat{G}(H) - \int_{\mathcal{X}} \hat{G}(\bar{H})\, P(H, d\bar{H}) = G(H) - \int_{\mathcal{X}} G(\bar{H})\, \mu(d\bar{H}), \tag{7.20}$$

where $H = (x, z, y, h)$, $\bar{H} = (\bar{x}, \bar{z}, \bar{y}, \bar{h})$, and P is the transitional kernel of the Markov Chain (H_k).

In our case, the Poisson equation is used to replace the integral with respect to the invariant measure μ with an integral with respect to the transition kernel P. We note that the following expansion converges for any $H = (x, z, y, h) \in \mathcal{X}$ if G is globally C-Lipschitz:

$$\hat{G}(H) = \sum_{k=0}^{\infty} \left(\int_{\mathcal{X}} G(\bar{H})\, P^k(H, d\bar{H}) - \int_{\mathcal{X}} G(\bar{H})\, \mu(d\bar{H}) \right), \tag{7.21}$$

where $P^k = \underbrace{P \circ ... \circ P}_{k \text{ times}}$. This is because

$$\left| \int_{\mathcal{X}} G(\bar{H}) P^k(H, d\bar{H}) - \int_{\mathcal{X}} G(\bar{H}) \mu(d\bar{H}) \right| \leq C \mathsf{Wass}_2(P^k(H, \cdot), \mu) \leq C q_0^k. \tag{7.22}$$

As a result, for all H

$$|\hat{G}(H)| = \left| \sum_{k=0}^{\infty} \left(\int_{\mathcal{X}} G(\bar{H})\, P^k(H, d\bar{H}) - \int_{\mathcal{X}} G(\bar{H})\, \mu(d\bar{H}) \right) \right| \leq \sum_{k=0}^{\infty} C q_0^k = \frac{C}{1 - q_0} < +\infty. \tag{7.23}$$

Lemma 7.11. *Let G be a globally C-Lipschitz function on $\mathcal{X}$, and define $\hat{G}$ as in* (7.21)*. Then $\hat{G}$ is the solution to the Poisson equation* (7.20)*. Moreover $\hat{G}$ is $C/(1 - q_0)$-Lipschitz, where q_0 is as defined in Theorem 3.6*

Proof. We observe that

$$\int_{\mathcal{X}} \hat{G}(\bar{H})\, P(H, d\bar{H}) = \int_{\mathcal{X}} \left[\sum_{k=0}^{\infty} \left(\int_{\mathcal{X}} G(\mathsf{H})\, P^k(\bar{H}, d\mathsf{H}) - \int_{\mathcal{X}} G(\mathsf{H})\, \mu(d\mathsf{H}) \right) \right] P(H, d\bar{H}).$$

We can exchange the sum and the integral as the integrand is summable (hence integrable). Noting that μ is invariant, we have

$$\int_{\mathcal{X}} \hat{G}(\bar{H})\, P(H, d\bar{H}) = \sum_{k=0}^{\infty} \left(\int_{\mathcal{X}} G(\bar{H}) P^{k+1}(H, d\bar{H}) - \int_{\mathcal{X}} G(\bar{H}) \mu(d\bar{H}) \right).$$

Subtracting this from $\hat{G}(H)$ yields

$$\hat{G}(H) - \int_{\mathcal{X}} \hat{G}(\bar{H})\, P(H, d\bar{H}) = \int_{\mathcal{X}} \hat{G}(\bar{H})\, P^0(H, d\bar{H}) - \int_{\mathcal{X}} G(\bar{H})\, \mu(d\bar{H}) = G(H) - \int_{\mathcal{X}} G(\bar{H})\, \mu(d\bar{H})$$

as desired. We further note that if $H' = (x', z', y', h')$, then

$$\begin{aligned} |\hat{G}(H) - \hat{G}(H')| &\leq \sum_{k=0}^{\infty} \left| \int_{\mathcal{X}} G(\bar{H})\, P^k(H, d\bar{H}) - \int_{\mathcal{X}} G(\bar{H})\, P^k(H', d\bar{H}) \right| \\ &\leq C \sum_{k=0}^{\infty} \mathsf{Wass}_2(P^k(H, \cdot), P^k(H', \cdot)) \\ &\leq C \|H - H'\|_{\mathcal{X}} \sum_{k=0}^{\infty} q_0^k \leq \frac{C}{1 - q_0} \|H - H'\|_{\mathcal{X}}, \end{aligned}$$

completing the proof. □

We now consider the Poisson equations with $G(\mathsf{H}) = G^{m,h}(\mathsf{H})$ as defined in lemma 7.9. These Poisson equations admit a solution for each $m \leq \lfloor NT \rfloor$ using the expansion in (7.21), for which we will call them $\hat{G}^{m,h}(\mathsf{H})$. In summary we have

$$\hat{G}^{m,h}(\mathsf{H}) - \int_{\mathcal{X}} \hat{G}^{m,h}(\bar{\mathsf{H}})\, P(\mathsf{H}, d\bar{\mathsf{H}}) = G^{m,h}(\mathsf{H}) - \int_{\mathcal{X}} G^{m,h}(\bar{\mathsf{H}})\, \mu(d\bar{\mathsf{H}}). \tag{7.24}$$

It is important to note from the above analysis that, since there is a constant $C_T > 0$ such that $G^{m,h}(\mathsf{H})$ is C_T-Lipschitz (Lemma 7.9) for $m \leq \lfloor NT \rfloor$ and $h \in \mathcal{H}$, the above analysis shows that Corollary 7.12 below holds.

Corollary 7.12.

$$\sup_{h,\mathsf{H}} |\hat{G}^{m,h}(\mathsf{H})| \leq C_T,$$

and that $\hat{G}^{m,h}(\cdot)$ *is* $C_T/(1-q_0)$ *Lipschitz.*

With this, we could study the difference between $\varphi_m^N(h)$ and $\tilde{\varphi}_m^N(h)$ for any $h \in \mathcal{H}$ and $m \leq \lfloor NT \rfloor$. Recalling that $H_k = (X_k, Z_k, Y_k, h_k)$ and letting $\bar{\mathsf{H}} = (\bar{\mathsf{x}}, \bar{\mathsf{z}}, \bar{\mathsf{y}}, \bar{\mathsf{h}})$, we have

$$\begin{aligned}
\varphi_m^N(h) - \tilde{\varphi}_m^N(h) &= \sum_{k=0}^{m-1} (\triangle\varphi_k^N(h) - \delta\varphi_k^N(h)) \\
&= \frac{1}{N} \sum_{k=0}^{m-1} \left(G^{k,h}(H_k) - \int_{\mathcal{X}} G^{k,h}(\bar{\mathsf{H}})\, \mu(d\bar{\mathsf{H}}) \right) \\
&= \frac{1}{N} \sum_{k=0}^{m-1} \left(\hat{G}^{k,h}(H_k) - \int_{\mathcal{X}} \hat{G}^{k,h}(\bar{\mathsf{H}})\, P(H_k, d\bar{\mathsf{H}}) \right) \\
&= Q^{N,1}(h) + Q^{N,2}(h) + R^{N,1}(h),
\end{aligned} \tag{7.25}$$

where

$$\begin{aligned}
Q_m^{N,1}(h) &= \frac{1}{N} \sum_{k=1}^{m-1} \left(\hat{G}^{k,h}(H_k) - \int_{\mathcal{X}} \hat{G}^{k,h}(\bar{\mathsf{H}})\, P(H_{k-1}, d\bar{\mathsf{H}}) \right) \\
Q_m^{N,2}(h) &= \frac{1}{N} \sum_{k=1}^{m-1} \left(\int_{\mathcal{X}} \hat{G}^{k,h}(\bar{\mathsf{H}})\, P(H_{k-1}, d\bar{\mathsf{H}}) - \int_{\mathcal{X}} \hat{G}^{k,h}(\bar{\mathsf{H}})\, P(H_k, d\bar{\mathsf{H}}) \right) \\
R^{N,1}(h) &= \frac{1}{N} \left(\hat{G}^{0,h}(H_0) - \int_{\mathcal{X}} \hat{G}^{0,h}(\bar{\mathsf{H}})\, P(H_0, d\bar{\mathsf{H}}) \right)
\end{aligned}$$

Corollary 7.12 immediately leads to

$$\sup_{h \in \mathcal{H}} |R^{N,1}(h)| \leq \frac{C_T}{N} \overset{N \to \infty}{\to} 0.$$

The second term on the RHS of (7.25), i.e. $Q_m^{N,2}$, could be analysed by observing that

$$\begin{aligned}
Q_m^{N,2}(h) &= \frac{\alpha}{N} \sum_{k=1}^{m-1} \left(\int_{\mathcal{X}} \hat{G}^{k,h}(\bar{H})\, P(H_{k-1}, d\bar{H}) - \int_{\mathcal{X}} \hat{G}^{k,h}(\bar{H})\, P(H_k, d\bar{H}) \right) \\
&= \frac{\alpha}{N} \left(\int_{\mathcal{X}} \hat{G}^{1,h}(\bar{H}) P(H_0, d\bar{H}) + \sum_{k=2}^{m-1} \int_{\mathcal{X}} \hat{G}^{k,h}(\bar{H})\, P(H_{k-1}, d\bar{H}) - \sum_{k=1}^{m-1} \int_{\mathcal{X}} \hat{G}^{k,h}(\bar{H})\, P(H_k, d\bar{H})) \right) \\
&= \frac{\alpha}{N} \left(\int_{\mathcal{X}} \hat{G}^{1,h}(\bar{H}) P(H_0, d\bar{H}) + \sum_{k=1}^{m-2} \int_{\mathcal{X}} \hat{G}^{k+1,h}(\bar{H})\, P(H_k, d\bar{H}) - \sum_{k=1}^{m-1} \int_{\mathcal{X}} \hat{G}^{k,h}(\bar{H})\, P(H_k, d\bar{H})) \right) \\
&= \frac{\alpha}{N} \sum_{k=1}^{m-2} \int_{\mathcal{X}} (\hat{G}^{k+1,h}(\bar{H}) - \hat{G}^{k,h}(\bar{H}))\, P(H_k, d\bar{H}) + R_m^{N,2}(h),
\end{aligned}$$

where

$$R_m^{N,2}(h) = \frac{\alpha}{N}\bigg(\int_{\mathcal{X}} \hat{G}^{1,h}(\bar{H})P(H_0, d\bar{H}) - \int_{\mathcal{X}} \hat{G}^{m-1,h}(\bar{H})\, P(H_{m-1}, d\bar{H}))\bigg)$$

We therefore naturally break our analysis of $Q_m^{N,2}$ into two parts:

Lemma 7.13. *If $m \le \lfloor NT \rfloor$, then $\sup_{h\in\mathcal{H}} |R_m^{N,2}(h)| \le C_T/N$.*

Proof. We note that $\hat{G}^{1,h}$ is C-Lipschitz by Lemma 7.11. Moreover, since $G^{m,h}$ is C_T-Lipschitz whenever $m \le \lfloor NT \rfloor$, we know from Corollary 7.12 that $\hat{G}^{m,h}(\cdot)$ is bounded by some constant $C_T > 0$ as well.

$$\begin{aligned}
|R_m^{N,2}(h)| &\le \frac{\alpha}{N}\bigg[\bigg|\int_{\mathcal{X}} \hat{G}^{1,h}(\bar{H})\, P(H_0, d\bar{H}) - \int_{\mathcal{X}} \hat{G}^{1,h}(\bar{H})\, P(H_{m-1}, d\bar{H})\bigg| \\
&\quad + \bigg|\int_{\mathcal{X}} \hat{G}^{1,h}(\bar{H})\, P(H_{m-1}, d\bar{H}) - \int_{\mathcal{X}} \hat{G}^{m-1,h}(\bar{H})\, P(H_{m-1}, d\bar{H})\bigg|\bigg] \\
&\le \frac{\alpha}{N}\left(C\mathsf{Wass}_2(P(H_0,\cdot), P(H_{m-1},\cdot)) + C_T + C_T\right) \le \frac{C_T}{N},
\end{aligned}$$

for a constant C_T that depends on T but changes from line to line. □

Lemma 7.14. *For $t \le T$, there are $C_T > 0$ such that for all $m \le \lfloor NT \rfloor$,*

$$\sup_{\mathsf{H}\in\mathcal{X}} |\Delta\hat{G}^{m,h}(\mathsf{H})| \le \frac{C_T}{N},$$

where $\Delta\hat{G}^{m,h}(\mathsf{H}) = \hat{G}^{m+1,h}(\mathsf{H}) - \hat{G}^{m,h}(\mathsf{H})$.

Proof. Define $\Delta G^{m,h}(\mathsf{H}) = G^{m+1,h}(\mathsf{H}) - G^{m,h}(\mathsf{H})$, then we see that $\Delta\hat{G}^{m,h}(\mathsf{H})$ is a solution to the Poisson equation for $\Delta G^{m,h}(\mathsf{H})$, i.e.

$$\Delta\hat{G}^{m,h}(\mathsf{H}) - \int_{\mathcal{X}} \Delta\hat{G}^{m,h}(\bar{H})\, P(\mathsf{H}, d\bar{H}) = G^{m,h}(\mathsf{H}) - \int_{\mathcal{X}} \Delta G^{m,h}(\bar{\mathsf{H}})\, \mu(d\bar{\mathsf{H}}),$$

As seen in (7.22)-(7.23), the boundedness of $\Delta\hat{G}^{m,h}(\mathsf{H})$ depends on the Lipschitz constants of $\Delta G^{m,h}(\mathsf{H})$. The proof is completed if we prove that $\Delta G^{m,h}(\mathsf{H})$ is C_T/N-Lipschitz. To begin, we note that

$$\begin{aligned}
\Delta G^{m,h}(\mathsf{H}) &= G^{m+1,h}(\mathsf{H}) - G^{m,h}(\mathsf{H}) \\
&= \alpha(\varphi_{m+1}^N(\varsigma(\mathsf{H})) - \varphi_m^N(\varsigma(\mathsf{H})))\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h}) \\
&\overset{(7.16)}{=} \frac{\alpha}{N} G^{m,\varsigma(\mathsf{H})}(H_m)\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h}),
\end{aligned}$$

where $H_m = (X_m, Z_m, Y_m, h_m)$. Since we know a-priori that $|(X_m, Z_m)| \le 1, |Y_m| \le C_y$ and $\|h_m\|_{C_b^2} \le C$,

we have the following control

$$
\begin{aligned}
&|\mathcal{K}_{X_m,\lambda}(\varsigma(\mathsf{H}),h_m)-\mathcal{K}_{X_m,\lambda}(\varsigma(\tilde{\mathsf{H}}),h_m)|\\
&\le \langle |\sigma(w^\top X_m+\langle b' h_m(w'),\lambda\rangle)([\varsigma(\mathsf{H})](w)-[\varsigma(\tilde{\mathsf{H}})](w))|\\
&\quad + c^2|\sigma'(w^\top X_m+\langle b' h_m(w'),\lambda\rangle)(\nabla[\varsigma(\mathsf{H})](w)-\nabla[\varsigma(\tilde{\mathsf{H}})](w))^\top X_m|,\lambda\rangle\\
&\le \int_{\mathbb{R}^d}|\varsigma(\tilde{\mathsf{H}})(w)-\varsigma(\tilde{\mathsf{H}})(w)|\,\lambda(dw)+C_\sigma\int_{\mathbb{R}^d}|\nabla\varsigma(\tilde{\mathsf{H}})(w)-\nabla\varsigma(\tilde{\mathsf{H}})(w)||X_m|\,\lambda(dw)\\
&\le (1+C_\sigma)\|\varsigma(\mathsf{H})-\varsigma(\tilde{\mathsf{H}})\|_{H^1(\lambda)} \overset{(7.19)}{\le} CC_\sigma(1+C_\sigma)\|\mathsf{H}-\tilde{\mathsf{H}}\|_{\mathcal{X}}\\
\Longrightarrow\quad &|G^{m,\varsigma(\mathsf{H})}(H_m)-G^{m,\varsigma(\tilde{\mathsf{H}})}(H_m)|\\
&\le \alpha|\varphi_m^N(\sigma(w^\top X_m+\langle b' h_m(w'),\,\lambda\rangle))-Y_m||\mathcal{K}_{X_m,\lambda}(\varsigma(\mathsf{H}),h_m)-\mathcal{K}_{X_m,\lambda}(\varsigma(\tilde{\mathsf{H}}),h_m)|\\
&\le C_T\|\mathsf{H}-\tilde{\mathsf{H}}\|_{\mathcal{X}}.\\
\Longrightarrow\quad &|\Delta G^{m,h}(\mathsf{H})-\Delta G^{m,h}(\tilde{\mathsf{H}})|\\
&=\frac{\alpha}{N}|K_{\mathsf{x},\lambda}(h,\tilde{\mathsf{h}})(G^{m,\varsigma(\mathsf{H})}(H_m)-G^{m,\varsigma(\tilde{\mathsf{H}})}(H_m))+G^{m,\varsigma(\tilde{\mathsf{H}})}(H_m)(\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})-\mathcal{K}_{\tilde{\mathsf{x}},\lambda}(h,\tilde{\mathsf{h}}))|\\
&\le \frac{C_T}{N}\|\mathsf{H}-\tilde{\mathsf{H}}\|_{\mathcal{X}}.
\end{aligned}
$$

So indeed $\Delta G^{m,h}(\mathsf{H})$ is C_T/N-Lipschitz whenever $m\le\lfloor NT\rfloor$, and therefore there is another $C_T>0$ such that $\sup_{\mathsf{H}\in\mathcal{X}}|\Delta\hat{G}^{m,h}|\le C_T$. □

The above lemmas lead to the control

$$\sup_{h\in\mathcal{H}}|Q_m^{N,2}(h)|\le\frac{\alpha}{N}\sum_{k=1}^{m-2}\frac{C_T}{N}+\sup_{h\in\mathcal{H}}|R_m^{N,2}(h)|\le\frac{C_T}{N}.$$

Finally, let's analyze the term $Q_m^{N,1}(h)$. Define the term

$$\tilde{g}_k^N(h)=\hat{G}^{k,h}(H_k)-\int_{\mathcal{X}}\hat{G}^{k,h}(\bar{\mathsf{H}})\,P(H_{k-1},d\bar{\mathsf{H}}),\quad k\ge 1,$$

so that $Q_m^{N,1}(h)=(\alpha\sum_{k=1}^{m-1}\tilde{g}_k^N(h))/N$. We define the increasing filtration $\{\mathcal{F}_m\}$, with $\mathcal{F}_m$ the smallest σ-algebra induced by the random variables $(H_k)_{k\le m}$. Then $\{\varphi_k^N(h)\}_{k\le m}$ are $\mathcal{F}_{m-1}$ measurable for any $h\in\mathcal{H}$, as they are continuous functions in terms of $(H_k)_{k\le m-1}$. Therefore, we have

$$\mathbb{E}[\tilde{g}_k^N(h)\,|\,\mathcal{F}_{k-1}]=\mathbb{E}[\mathbb{E}[\tilde{g}_k^N(h)\,|\,\varphi_k^N]\,|\,\mathcal{F}_{k-1}]=\mathbb{E}[\hat{G}^{k,h}(H_k)\,|\,H_{k-1}]-\int_{\mathcal{X}}\hat{G}^{k,h}(\bar{\mathsf{H}})\,P(H_{k-1},d\bar{\mathsf{H}})=0.$$

If we define $H_{k:m}$ being the sequence $(H_k,...,H_m)$, and $\sigma(\varphi_{k:m}^N(h))$ being the smallest σ-algebra so that each of the random variables $\varphi_{k:m}^N(h)$ is measurable, then for $m>k$ we can prove recursively

$$
\begin{aligned}
\mathbb{E}[\tilde{g}_k^N(h)\tilde{g}_m^N(h)\,|\,\mathcal{F}_{k-1}]&=\mathbb{E}[\mathbb{E}[g_k^N(h)g_m^N(h)\,|\,H_{m-1},\sigma(\varphi_{m-1}^N(h))]\,|\,\mathcal{F}_{k-1}]\\
&=\mathbb{E}[\mathbb{E}[g_k^N(h)g_m^N(h)\,|\,H_{k-1:m-1},\sigma(\varphi_{k-1:m-1}^N(h))]\,|\,\mathcal{F}_{k-1}]\\
&=\mathbb{E}[g_k^N(h)\mathbb{E}[g_m^N(h)\,|\,H_{k-1:m-1},\sigma(\varphi_{k-1:m-1}^N(h))]\,|\,\mathcal{F}_{k-1}]=0.
\end{aligned}
$$

Finally, notice that $\hat{G}^{k,h}(\cdot)$ is uniformly bounded (see Corollary 7.12), we have

$$
\begin{aligned}
\mathbb{E}\left[(\tilde{g}_k^N(h))^2\right]&\le\mathbb{E}\left[\hat{G}^{k,h}(H_k)-\int_{\mathcal{X}}\hat{G}^{k,h}(\bar{\mathsf{H}})\,P(H_{k-1},d\bar{\mathsf{H}})\right]^2\\
&\le 2\mathbb{E}\left[\hat{G}^{k,h}(H_k)\right]^2+2\mathbb{E}\left[\int_{\mathcal{X}}\hat{G}^{k,h}(\bar{\mathsf{H}})\,P(H_{k-1},d\bar{\mathsf{H}})\right]^2\le C_T.
\end{aligned}
$$

Therefore

$$\mathbb{E}[Q^{N,1}(h)]^2 = \frac{\alpha^2}{N^2}\sum_{j,k=1}^{\lfloor Nt\rfloor-1}\mathbb{E}[\tilde{g}_j^N(h)\tilde{g}_k^N(h)] = \frac{\alpha^2}{N^2}\sum_{k=1}^{\lfloor Nt\rfloor-1}\mathbb{E}[(\tilde{g}_k^N(h))^2] \le \frac{C_T}{N}.$$

The above estimates are uniform for all $h \in \mathcal{H}$, so by combining all of them we have

Lemma 7.15. *For all $m \le \lfloor NT\rfloor$,*

$$\sup_h \mathbb{E}\left[\varphi_m^N(h) - \tilde{\varphi}_m^N(h)\right]^2 \le \frac{C_T}{N}.$$

Proof. We collect the above estimates to show that

$$\begin{aligned}\sup_h \mathbb{E}\left[\varphi_m^N(h) - \tilde{\varphi}_m^N(h)\right]^2 &\le \sup_h \mathbb{E}\left[Q^{N,1}(h)\right]^2 + \sup_h \mathbb{E}\left[Q^{N,2}(h)\right]^2 + \sup_h \mathbb{E}\left[R^{N,1}(h)\right]^2\\ &\le \frac{C_T}{N} + \frac{C_T}{N^2} + \frac{C_T}{N^2} \le \frac{C_T}{N}.\end{aligned}$$

□

Let us then define a new process $\phi_t^N(h)$ that satisfies the recursion

$$\begin{aligned}\phi_m^N(h) &= \sum_{k=0}^{m-1}\delta\phi_m^N(h), \quad \phi_0^N(h) = 0,\\ \delta\phi_k^N(h) &= -\frac{\alpha}{N}\int_{\mathcal{X}}(\phi_k^N(\varsigma(\mathsf{H})) - \mathsf{y})\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{H}),\end{aligned}$$

and study the difference $\Gamma_m^N(h) = \varphi_m^N(h) - \phi_m^N(h)$. Then we have

$$\begin{aligned}\Gamma_m^N(h) &= \varphi_m^N(h) - \tilde{\varphi}_m^N(h) + \tilde{\varphi}_m^N(h) - \phi_m^N(h)\\ &= \varphi_m^N(h) - \tilde{\varphi}_m^N(h) - \frac{\alpha}{N}\sum_{k=0}^{m-1}\int_{\mathcal{X}}\Gamma_k^N(\varsigma(\mathsf{H}))\,\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{H}).\end{aligned}$$

So

$$\begin{aligned}\mathbb{E}\left[\Gamma_m^N(h)\right]^2 &\le 2\mathbb{E}[\varphi_m^N(h) - \tilde{\varphi}_m^N(h)]^2 + \frac{2\alpha}{N^2}\mathbb{E}\left[\sum_{k=0}^{m-1}\int_{\mathcal{X}}\Gamma_k^N(\varsigma(\mathsf{H}))\,\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{H})\right]^2\\ &\le \frac{C_T}{N} + \frac{2\alpha}{N}\sum_{k=0}^{m-1}\mathbb{E}\left[\int_{\mathcal{X}}\left(\Gamma_k^N(\varsigma(\mathsf{H}))\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\right)^2\mu(d\mathsf{H})\right]\\ &\overset{\text{(Tonelli)}}{=} \frac{C_T}{N} + \frac{2\alpha}{N}\sum_{k=0}^{m-1}\int_{\mathcal{X}}\mathbb{E}\left[\Gamma_k^N(\varsigma(\mathsf{H}))\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\right]^2\mu(d\mathsf{x},d\mathsf{z},d\mathsf{y},d\mathsf{h})\\ &\le \frac{C_T}{N} + \frac{C}{N}\sum_{k=0}^{m-1}\int_{\mathcal{X}}\sup_{\mathfrak{h}\in\mathcal{H}}\mathbb{E}\left[\Gamma_k^N(\mathfrak{h})\right]^2\mu(d\mathsf{x},d\mathsf{z},d\mathsf{y},d\mathsf{h}).\end{aligned}$$

Defining further $\tilde{\Gamma}_m^N = \sup_{\mathfrak{h}\in\mathcal{H}}\mathbb{E}[\Gamma_m^N(\mathfrak{h})]^2$, we have

$$\tilde{\Gamma}_m^N \le \frac{C_T}{N} + \frac{C_T}{N}\sum_{k=0}^{m-1}\tilde{\Gamma}_k^N.$$

So by discrete Gronwall inequality we have

Lemma 7.16. *For all $m \le \lfloor TN \rfloor$ we have*

$$\tilde{\Gamma}_m^N = \sup_{\mathfrak{h}\in\mathcal{H}} \mathbb{E}[\Gamma_m^N(\mathfrak{h})]^2 = \sup_{\mathfrak{h}\in\mathcal{H}} \mathbb{E}[\varphi_m^N(\mathfrak{h}) - \phi_m^N(\mathfrak{h})]^2 \le \frac{C_T}{N}\exp\left(\frac{mC_T}{N}\right) \le \frac{C_T\exp(TC_T)}{N} = \frac{C_T}{N}.$$

Finally, define $\tilde{\phi}_t^N(h) = \phi_{\lfloor Nt\rfloor}^N(h)$ be a time-rescaled evolution of $\phi_m^N(h)$, and recall our desired limiting equation is

$$g_t(h) = -\alpha\int_0^t \left[\int_{\mathcal{X}} (g_s(\varsigma(\mathsf{H})) - \mathsf{y}))\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{H})\right] ds, \quad g_0(h) = 0.$$

Lemma 7.17. *As $N \to +\infty$, we have*

$$\sup_{t\in[0,T]}\sup_{h\in\mathcal{H}} |\tilde{\phi}_t^N(h) - g_t(h)| \le \frac{C_T}{N}.$$

Proof. Let

$$\Upsilon_t^N(h) = \phi_t^N(h) - g_t(h), \quad \tilde{\Upsilon}_t^N = \sup_h |\Upsilon_t^N(h)|,$$

then for $m \le \lfloor NT\rfloor$,

$$\begin{aligned}
\left|\Upsilon_{(m+1)/N}^N(h)\right| &= |\tilde{\phi}_{(m+1)/N}^N(h) - \tilde{\phi}_{m/N}^N(h) - g_{(m+1)/N}(h) + g_{m/N}(h) + \Upsilon_{m/N}^N(h)| \\
&\le |\phi_{m+1}^N(h) - \phi_m^N(h) + g_{(m+1)/N}(h) - g_{m/N}(h)| + |\Upsilon_{m/N}^N(h)| \\
&\le \Big| -\frac{\alpha}{N}\int_{\mathcal{X}} (\phi_m^N(\varsigma(\mathsf{H})) - \mathsf{y})\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{x},d\mathsf{z},d\mathsf{y},d\mathsf{h}) \\
&\quad - \alpha\int_{m/N}^{(m+1)/N}\int_{\mathcal{X}} (g_s(\varsigma(\mathsf{H})) - \mathsf{y})\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{x},d\mathsf{z},d\mathsf{y},d\mathsf{h})\,dt\Big| + |\Upsilon_{m/N}^N(h)| \\
&\le \alpha\Big|\int_{m/N}^{(m+1)/N}\int_{\mathcal{X}} (\phi_m^N(\varsigma(\mathsf{H}) - g_{m/N}(\varsigma(\mathsf{H}) + g_{m/N}(\varsigma(\mathsf{H}) - g_s(\varsigma(\mathsf{H}))\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})\,\mu(d\mathsf{x},d\mathsf{z},d\mathsf{y},d\mathsf{h})\,dt\Big| \\
&\quad + |\Upsilon_{m/N}^N(h)| \\
&\le \alpha\int_{m/N}^{(m+1)/N}\int_{\mathcal{X}} \left[\left|\Upsilon_{m/N}^N(\mathsf{h})\right| + \left|g_{m/N}(\mathsf{h}) - g_s(\mathsf{h})\right|\,|\mathcal{K}_{\mathsf{x},\lambda}(h,\mathsf{h})|\right]\,\mu(d\mathsf{x},d\mathsf{z},d\mathsf{y},d\mathsf{h})\,dt + |\Upsilon_{m/N}^N(h)|,
\end{aligned}$$

The control of the operator norm of g_t by (6.18) suggests that for all $h \in \mathcal{H}$, $g_t(h) \le C_T\|h\|_{H^1(\lambda)} \le C_T$. As a result, for all $s \in [m/N, (m+1)/N)$ and $\mathsf{h} \in \mathcal{H}$,

$$|g_s(\varsigma(\mathsf{H})) - g_{m/N}(\varsigma(\mathsf{H}))| \le \int_{m/N}^s |[\mathcal{A}(g_\tau)](\varsigma(\mathsf{H})) + b(\varsigma(\mathsf{H}))|\,d\tau \le C_T(s - m/N) \le \frac{C_T}{N}.$$

Therefore,

$$\left|\Upsilon_{(m+1)/N}^N(h)\right| \le \frac{\alpha}{N}\left[\tilde{\Upsilon}_{m/N}^N + \frac{C_T}{N}\right] + |\Upsilon_{m/N}^N(h)|,$$

and that

$$\tilde{\Upsilon}_{(m+1)/N}^N(h) \le \left(1 + \frac{C_T}{N}\right)\tilde{\Upsilon}_{m/N}^N + \frac{C_T}{N^2}.$$

By Lemma (A.1), for all $m \le \lfloor TN\rfloor$ we have

$$\sup_{h\in\mathcal{H}} |\tilde{\phi}_{m/N}^N(h) - g_{m/N}(h)| =: \tilde{\Upsilon}_{m/N}^N(h) \le \left(1 + \frac{C_T}{N}\right)^m \frac{C_T}{N} \le \frac{C_T\exp(TC_T)}{N} = \frac{C_T}{N}.$$

Finally, for all $t \leq NT$, define $m = \lfloor t/N \rfloor$, then

$$\begin{aligned}\sup_{h\in\mathcal{H}} |\tilde{\phi}_s^N(h) - g_s(h)| &\leq \sup_{h\in\mathcal{H}} \underbrace{|\tilde{\phi}_s^N(h) - \tilde{\phi}_{m/N}^N(h)|}_{=0} + \sup_{h\in\mathcal{H}} |\tilde{\Upsilon}_{m/N}^N(h)| + \sup_{h\in\mathcal{H}} |g_{m/N}(h) - g_s(h)| \\ &\leq \frac{C_T}{N} + \frac{C_T}{N} = \frac{C_T}{N}.\end{aligned}$$

This completes the proof as the above bound is uniform in $h \in \mathcal{H}$. □

7.4 The final steps

Now, we are in position to prove the main convergence result of this paper.

Proof. (Proof of Theorem 4.2) As a summary, we collect

- Lemma 7.8:
$$\sup_{m\leq NT} \sup_{h\in\mathcal{H}} \mathbb{E}\left|g_m^N(h) - \varphi_m^N(h)\right| \leq \frac{C_T}{N^{(1-\beta-2\gamma)\wedge\gamma\wedge(\beta-1/2)}}.$$

- Lemma 7.16:
$$\sup_{m\leq NT} \sup_{h\in\mathcal{H}} \mathbb{E}[\varphi_m^N(h) - \phi_m^N(h)]^2 \leq \frac{C_T}{N} \implies \sup_{h\in\mathcal{H}} \mathbb{E}|\varphi_m^N(h) - \phi_m^N(h)| \leq \frac{C_T}{N^{1/2}}.$$

- Lemma 7.17:
$$\sup_{t\in[0,T]} \sup_{h\in\mathcal{H}} |\phi_t^N(h) - g_t(h)| \leq \frac{C_T}{N}.$$

Adding all of the error terms yields:

$$\begin{aligned}\sup_{t\in[0,T]} \sup_{h\in\mathcal{H}} \mathbb{E}|g_t^N(h) - g_t(h)| &\leq \sup_{t\in[0,T]} \sup_{h\in\mathcal{H}} \left[\mathbb{E}|g_t^N(h) - \varphi_t^N(h)| + \mathbb{E}|\varphi_t^N(h) - \phi_t^N(h)|\right] + \sup_{t\in[0,T]} \sup_{h\in\mathcal{H}} \left[\mathbb{E}|\phi_t^N(h) - g_t(h)|\right] \\ &\leq \sup_{m\leq NT} \sup_{h\in\mathcal{H}} \left[\mathbb{E}|g_m^N(h) - \varphi_m^N(h)| + \mathbb{E}|\varphi_m^N(h) - \phi_m^N(h)|\right] + \sup_{t\in[0,T]} \sup_{h\in\mathcal{H}} \left[\mathbb{E}|\phi_t^N(h) - g_t(h)|\right] \\ &\leq \frac{C_T}{N^{(1-\beta-2\gamma)\wedge\gamma\wedge(\beta-1/2)}} + \frac{C_T}{N^{1/2}} + \frac{C_T}{N} \\ &\leq \frac{C_T}{N^{\epsilon}}\end{aligned}$$

where $\epsilon = (1-\beta-2\gamma)\wedge\gamma\wedge(\beta-1/2)\wedge 1/2 > 0$. □

A Recursive Inequality

Many proofs of the technical lemmas involves the study of a sequence $(a_k)_{k\geq 0}$ that satisfies the following recursive inequality:

$$a_k \leq M_1 a_{k-1} + M_2,$$

where $M_1, M_2 \geq 0$. By recursion, we can prove that

Lemma A.1. *For $M_1 < 1$, we have*

$$a_k \leq M_1^k a_0 + \frac{1 - M_1^k}{1 - M_1} M_2 \leq M_1^k a_0 + \frac{1}{1 - M_1} M_2,$$

and for $M_1 > 1$, we have

$$a_k \leq M_1^k a_0 + \frac{M_1^k - 1}{M_1 - 1} M_2 \leq M_1^k \left(a_0 + \frac{M_2}{M_1 - 1} \right).$$

B Construction of a clipping function

We consider the function

$$f(x) = \begin{cases} \exp(-1/x) & x > 0 \\ 0 & x \leq 0, \end{cases}$$

which is known to be infinitely smooth (i.e. in $C^\infty(\mathbb{R})$). Therefore, the function

$$g(x) = \frac{f(x)}{f(x) + f(1 - x)}$$

is also infinitely smooth. In particular we have

$$g(x) = \begin{cases} = 0 & x < 0 \\ \in [0, 1] & x \in [0, 1] \\ = 1 & x > 1 \end{cases}.$$

Therefore for any $a < b$, the infinitely smooth function

$$g_{a,b}(x) = g\left(\frac{x - a}{b - a} \right) = \begin{cases} = 0 & x < a \\ \in [0, 1] & x \in [a, b] \\ = 1 & x > b \end{cases},$$

and that the infinitely smooth function

$$\rho_N(x) = g_{-2N^\gamma, -N^\gamma}(-x) g_{-2N^\gamma, -N^\gamma}(x) = \begin{cases} = 1 & |x| \leq N^\gamma \\ \in [0, 1] & N^\gamma < |x| \leq 2N^\gamma \\ = 0 & |x| > 2N^\gamma \end{cases}.$$

Finally, the function

$$\psi_N(x) = \int_0^x \rho_N(y)\, dy$$

satisfies all the requirement for being a smooth clipping function in definition 2.7: (2) follows by direct computation, and (3) is true by definition (that $\frac{d}{dx}\psi_N(x) = \rho_N(x)$). Finally, (1) follows by Fundamental theorem of calculus. By symmetry we could prove only for the case when $x > 0$, for which

$$|\psi_N(x)| \leq \int_0^{2N^\gamma} |\rho_N(y)|\, dy \leq 2N^\gamma.$$

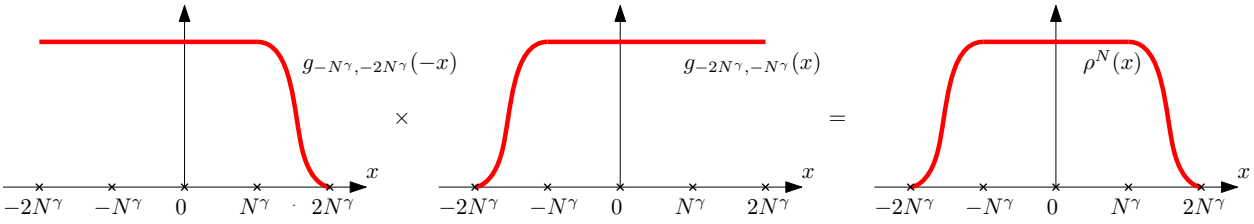

Figure 5: $\rho_N(x) = g_{-2N^\gamma,-N^\gamma}(-x) \times g_{-2N^\gamma,-N^\gamma}(x)$

Acknowledgement

The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility for completing the numerical simulations. Samuel Lam's fellowship is supported by the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1).

References

[1] Arik S, Chrzanowski M, Coates A, Diamos G, Gibiansky A, Kang Y, et al. Deep voice: Real-time neural text-to-speech. In: Proceedings of Machine Learning Research 70. ML Research Press; 2017. .

[2] Zhang Y, Chan W, Jaitly N. Very deep convolutional networks for end-to-end speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2017. p. 4845-9. Doi:10.1109/ICASSP.2017.7953077.

[3] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems 31. Curran Associates, Inc.; 2018. .

[4] Chizat L, Bach F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In: Inc CA, editor. Advances in Neural Information Processing Systems 31; 2018. p. 3040-50.

[5] Mei S, Montanari A, Nguyen P. A mean field view of the landscape of two-layer neural network. Proceedings of the National Academy of Sciences (PNAS). 2018;115(33):E7675-1. Edited by Peter J. Bickel.

[6] Rotskoff GM, Vanden-Eijnden E; 2018. Neural Networks as Interacting Particle Systems: Asymptotic Convexity of the Loss Landscape and Universal Scaling of the Approximation Error.

[7] Sirignano J, Spiliopoulos K. Mean Field Analysis of Neural Networks: A Law of Large Numbers. SIAM Journal on Applied Mathematics. 2020;80(2):725-52.

[8] Elman JL. Finding structure in time. Cognitive Science. 1990;14(2):179-211. Available from: `https://www.sciencedirect.com/science/article/pii/036402139090002E`.

[9] Sirignano J, Spiliopoulos K. Mean Field Analysis of Deep Neural Networks. Mathematics of Operations Research. 2021;47(1):120-52.

[10] Almeida LB. A Learning Rule for Asynchronous Perceptrons with Feedback in a Combinatorial Environment. In: Proceedings of the IEEE First International Conference on Neural Networks (San Diego, CA). vol. II. Piscataway, NJ: IEEE; 1987. p. 609-18.

[11] Pineda F. Generalization of Back propagation to Recurrent and Higher Order Neural Networks. In: Anderson D, editor. Neural Information Processing Systems. vol. 0. American Institute of Physics; 1987. Available from: `https://proceedings.neurips.cc/paper_files/paper/1987/file/735b90b4568125ed6c3f678819b6e058-Paper.pdf`.

[12] Liao R, Xiong Y, Fetaya E, Zhang L, Yoon K, Pitkow X, et al. Reviving and Improving Recurrent Back-Propagation. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning. vol. 80 of Proceedings of Machine Learning Research. PMLR; 2018. p. 3082-91. Available from: https://proceedings.mlr.press/v80/liao18c.html.

[13] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2016.

[14] Williams R, Peng J. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. Neural Computation. 1990;2(4):490-501.

[15] Sutskever I. Training Recurrent Neural Networks. Department of computer science, University of Toronto. 406, 413; 2012.

[16] Zhang J, He T, Sra S, Jadbabaie A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. arXiv preprint arXiv:190511881. 2019.

[17] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: Dasgupta S, McAllester D, editors. Proceedings of the 30th International Conference on Machine Learning. vol. 28(3) of Proceedings of Machine Learning Research. Atlanta, Georgia, USA: PMLR; 2013. p. 1310-8. Available from: https://proceedings.mlr.press/v28/pascanu13.html.

[18] Cohen S, Jiang D, Sirignano J. Neural Q-learning for solving elliptic PDEs.; 2022. Preprint.

[19] Nestruev J. Smooth manifolds and observables [electronic resource]. 2nd ed. Graduate texts in mathematics. Cham, Switzerland: Springer; 2020.

[20] Sirignano J, Spiliopoulos K. Asymptotics of Reinforcement Learning with Neural Networks. Stochastic Systems. 2022 March;12(1):2-29.

[21] Sirignano J, MacArt J, Spiliopoulos K. PDE-constrained Models with Neural Network Terms: Optimization and Global Convergence. Journal of Computational Physics. 2023 May;481.

[22] Spiliopoulos K, Yu J. Normalization effects on shallow neural networks and related asymptotic expansions. AIMS Journal on Foundations of Data Science. 2021;3(2):151-200.

[23] Yu J, Spiliopoulos K. Normalization effects on deep neural networks. Foundations of Data Science. 2023 September;5(3):389-465. Early access March 2023.

[24] Agazzi A, Lu J, Mukherjee S. Global optimality of Elman-type RNNs in the mean-field regime. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning. vol. 202 of Proceedings of Machine Learning Research. PMLR; 2023. p. 196-227. Available from: https://proceedings.mlr.press/v202/agazzi23a.html.

[25] Villani C. Optimal transport: old and new. Springer, Berlin; 2009.

[26] Pazy A. Semigroups of linear operators and applications to partial differential equations. Applied mathematical sciences (Springer-Verlag New York Inc.) ; v.44. New York: Springer-Verlag; 1983.

[27] Meyn SP, Tweedie RL. Markov Chains and Stochastic Stability. Springer, London; 2012.