Global Convergence in Training Large-Scale Transformers

Cheng Gao[1*] **Yuan Cao**[2*] **Zihao Li**[1] **Yihan He**[1] **Mengdi Wang**[1]
Han Liu[3] **Jason M. Klusowski**[1†] **Jianqing Fan**[1†]

[1]Princeton University [2]Hong Kong University [3]Northwestern University
{chenggao,zihaoli,yihan.he,mengdiw,jason.klusowski,jqfan}@princeton.edu
yuancao@hku.hk hanliu@northwestern.edu

Abstract

Despite the widespread success of Transformers across various domains, their optimization guarantees in large-scale model settings are not well-understood. This paper rigorously analyzes the convergence properties of gradient flow in training Transformers with weight decay regularization. First, we construct the mean-field limit of large-scale Transformers, showing that as the model width and depth go to infinity, gradient flow converges to the Wasserstein gradient flow, which is represented by a partial differential equation. Then, we demonstrate that the gradient flow reaches a global minimum consistent with the PDE solution when the weight decay regularization parameter is sufficiently small. Our analysis is based on a series of novel mean-field techniques that adapt to Transformers. Compared with existing tools for deep networks [41] that demand homogeneity and global Lipschitz smoothness, we utilize a refined analysis assuming only *partial homogeneity* and *local Lipschitz smoothness*. These new techniques are of independent interest.

1 Introduction

Transformers have revolutionized the field of deep learning since their introduction in [59]. These models are distinguished by their immense scales, often comprising billions of parameters to achieve state-of-the-art performance. Notably, this massive parameterization enables them to excel in a variety of domains, notably in natural language processing [18, 49, 58] and vision tasks [17, 31], where they have significantly advanced the frontiers of machine learning.

Despite the widespread adoption of Transformer models, our understanding of their optimization guarantees is still in its early stages. One particularly intriguing phenomenon is that as the size of model increases, training algorithms typically converges globally despite the highly nonconvex landscape of the training objective function. Remarkably, it remains somewhat enigmatic how gradient-based approaches can consistently succeed when training large-scale Transformers.

Notably, there have been several recent works showing the global convergence of training overparameterized neural networks [45, 15, 24, 13, 41, 19, 20, 30, 3, 21, 68]. In particular, several works [41, 19, 20] studied the setting with deep neural networks with skip connections. By studying the connections between the network with discretization in the parameter space and a corresponding ordinary differential equation system [63, 11, 37], these works demonstrated global convergence guarantees of wide and deep neural networks based on a mean-field analysis. However, these results are established based on certain homogeneity and/or global Lipschitz smoothness properties of the neural network, which are not applicable to Transformer models. Therefore, it remains an open question how gradient-based methods can effectively train large-scale Transformers.

Preprint. Under review.

1.1 Our contribution

In this work, we bridge the gap between Transformer theory and practice by demonstrating the global convergence of Transformer training optimization via gradient flow in a large-scale model regime. We analyze the mean-field limit of the Transformer model, which is characterized by the *distribution* of model parameters, shifting the focus from parameter space to distributional dynamics in the Wasserstein metric [15]. This approach yields two key theorems:

i. We show the closeness between practical discrete Transformers trained by gradient flow and continuous Transformers whose parameter distribution follows a partial differential equation of the Wasserstein gradient flow (Theorem 3.1). Our result demonstrates that large-scale discrete Transformers can be approximated by its mean-field limit and the approximation error can be expressed in terms of the width and depth of the Transformer models.

ii. This approximation facilitates our analysis of the global convergence (Theorem 4.1) of discrete Transformer models. By leveraging the universal approximation capabilities of either the self-attention or feed-forward layers, we demonstrate that a basic gradient flow method can reliably find a global optimum, despite the highly non-convex landscape of the training objective.

We also highlight our novel contributions to Transformer theory through the development of these two core results:

i. The assumption on activation regularity conditions (Assumption 2) is less stringent compared to those usually found in studies of two-layer neural networks [45, 15, 24, 13] or deep ResNet networks [19, 20, 41]. In particular, many existing approximation guarantees reply on a Lipschitz continuity property of the network gradients, which limits the mean-filed study to neural networks with smooth activation functions. In comparison, our analysis relaxes this assumption and only requires local Lipschitz continuity of the gradient in expectation. This relaxation broadens the applicability of our approach and ensures that our result can cover more practical Transformer architectures.

ii. Our model differs from the ResNet models in [41, 19, 20, 12], as those models incorporate only a single identical encoder within each evolutionary block. Unlike the typical theoretical configurations, our model employs two distinct encoders f and h that alternate throughout the network's depth. More importantly, despite the distinct encoders used, the continuous limit of our model uniformly interprets the encoder as an average of f and h, providing a rigorous validation of concepts proposed in [41] and [60] from a new perspective.

iii. Our global convergence guarantee for training Transformer models is also broadly applicable: our assumption (Assumption 4) ensures global convergence by relying on the universal approximation capabilities of *either* the self-attention or the feed-forward encoder. Additionally, we incorporate a more flexible framework by adopting partial 1-homogeneity for only a subset of the parameters, in contrast to the full parameter homogeneity required in studies such as [41]. This modification enables the use of softmax and sigmoid activation, expanding beyond the hardmax and ReLU restricted by full homogeneity.

Additional related works. See Appendix B for a detailed discussion.

Notations. For any $\alpha \in \mathbb{R}^d$, $\dim(\alpha)$ refers to its dimension d. For any $B \in \mathbb{R}^{d\times d}$, its trace is denoted by $\mathrm{Tr}(B)$. For any positive integer n, Let $[n] = \{1, 2, \dots, n\}$. Let 0_d denote the d-dimension vector of all zeros. Let $W_p(\mu, \nu)$ denote the Wasserstein-p distance between two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ for $p \geq 1$. For a matrix $A = (a_1, a_2, \dots, a_n)$, define its vectorization version as $\mathrm{vec}[A] := (a_1^T, a_2^T, \dots, a_n^T)^T$. Let $\delta(\cdot)$ denote the Dirac mass and $\mathbf{1}\{\cdot\}$ be the indicator function. Let $\mathrm{supp}(\cdot)$ denote the support of any distribution. Let $\|\cdot\| = \|\cdot\|_2$ denote the l_2 norm and $\|\cdot\|_{\max}$ denote the maximum norm. For any subsets D_1, D_2 in Euclidean space, define $\mathcal{C}(D_1, D_2)$ as the collection of functions that map D_1 to D_2 and are continuous over D_1. Define the Bounded Lipschitz norm for any measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ as $\|\mu\|_{\mathrm{BL}} := \sup\{\int f d\mu : f : \mathbb{R}^d \to \mathbb{R},\ \sup|f| \leq 1,\ f \text{ is } 1\text{−Lipschitz}\}$.

2 Transformer model

In this section, we describe the our deep Transformer model with each data input as a sequence, and the gradient flow algorithm used for training.

2.1 Data setting

In our paper, the data input is both general and straightforward: an input sequence $H \in \mathbb{R}^{D\times(N+1)}$ consisting of $N+1$ tokens, each with dimension D. We consider the setting where each input sequence H is associated with a label $y(H) \in \mathbb{R}$, where $y(H)$ is the target function we aim to learn. Furthermore, we assume that each instance H is i.i.d. drawn from a population distribution μ.

Relation to in-context learning (ICL) Our data setting is versatile and applicable to any task involving sequential input. It particularly suits the in-context learning (ICL) scenario [6, 9, 67], where models are capable of making accurate predictions on new data when prompted with training examples from the same pool. For clarity, consider the input sequence $H \in \mathbb{R}^{D\times(N+1)}$ formatted as follows:

$$H = [h_1, h_2, \ldots, h_{N+1}] = \begin{bmatrix} x_1 & x_2 & \ldots & x_N & x_{N+1} \\ y_1 & y_2 & \ldots & y_N & 0 \\ p_1 & p_2 & \ldots & p_N & p_{N+1} \end{bmatrix} \overset{i.i.d.}{\sim} \mu, \quad y_{N+1} = y(H).$$

Here, $\{x_i\}_{i\in[N]}$ are the input vectors, each associated with a corresponding label $\{y_i\}_{i\in[N]}$. The last token, x_{N+1} is the test input for which a prediction is made. The third row contains the customized and fixed positional encoding vectors $\{p_i\}_{i\in[N]}$, which typically include ones, zeros, and indicators denoting the token for prediction. The label for the query point x_{N+1} is then given by $y_{N+1} = y(H)$ in our terminology. ICL operates in a zero-shot fashion, without any updates to the model's parameters, highlighting a unique and powerful capability of these systems to adapt and generalize based on the provided context alone. In [6], the authors demonstrate that fixed Transformers can approximate in-context penalized generalized linear regression to any desired degree.

2.2 Model

We follow a common configuration of Transformer architectures [6, 33, 35, 42, 65] where each Transformer block consists of two distinct layers: a self-attention mechanism layer and a token-wise feed-forward neural network layer, both equipped with skip connections. We assume that both layers consist of the average of M heads, treated uniformly as the *width* across all blocks for simplicity. The formulation for a matrix input $Z \in \mathbb{R}^{D\times(N+1)}$ and a given residual step size $\eta > 0$ is as follows: Each residual self-attention layer is represented by

$$\mathrm{Attn}_{\theta_1,\theta_2,\ldots,\theta_M}(Z, \eta) = Z + \eta M^{-1} \sum_{j=1}^{M} f(Z, \theta_j), \tag{2.1}$$

and each residual feed-forward neural network layer is defined by

$$\mathrm{MLP}_{w_1,w_2,\ldots,w_M}(Z, \eta) = Z + \eta M^{-1} \sum_{j=1}^{M} h(Z, w_j) \tag{2.2}$$

for parameter vectors θ and w in the Euclidean space. The encoders for the self-attention and feed-forward layers are denoted as $f : \mathbb{R}^{D\times(N+1)} \to \mathbb{R}^{D\times(N+1)}$ and $h : \mathbb{R}^{D\times(N+1)} \to \mathbb{R}^{D\times(N+1)}$, respectively. The self-attention encoder f formulation, commonly adopting a *multiplicative* or *dot-product* approach as detailed in [8, 33, 42, 57, 59, 65], can be exemplified by

$$f(Z, \theta) = W_O W_V Z \sigma_{\mathrm{A}}\Big[(W_K Z)^T W_Q Z\Big],$$

where $W_V, W_K, W_Q \in \mathbb{R}^{s\times D}$, and $W_O \in \mathbb{R}^{D\times s}$. This formulation can be reparametrized to

$$f(Z, \theta) = V Z \sigma_{\mathrm{A}}\Big[Z^T W Z\Big], \tag{2.3}$$

where $V, W \in \mathbb{R}^{D\times D}$, $\theta = \mathrm{vec}[V, W]$. The activation σ_{A} typically uses column-wise softmax, but component-wise ReLU is also viable, as in [6]. For the feed-forward layer, an example of the encoder is $h(Z, w) = W_2\sigma_{\mathrm{M}}(W_1 H)$, as detailed in [6, 33, 65], where $w = \mathrm{vec}[W_1, W_2]$ and the activation σ_{M} is component-wise ReLU. Alternatively, setting $h \equiv 0$ results in a Transformer block that comprises only the self-attention layer, referred to as "attention-only" Transformers, as discussed in [6, 40, 43, 59].

Next, we analyze a Transformer network composed of L Transformer blocks, referring to L as the *depth* of the model. In this paper, we introduce an additional term, η, in (2.1) and (2.2) to simulate the model's evolution in a residual manner. We set the step size η as $\Delta t/2$, where $\Delta t = 1/L$. As L increases, Δt approaches zero, allowing Transformer blocks to incrementally contribute to the model's overall progression. The structure of the network is then defined as follows:

$$\begin{cases} \widehat{T}_\Theta(H, t+\Delta t/2) & = \mathrm{Attn}_{\theta_{t,1},\dots,\theta_{t,M}}(\widehat{T}_\Theta(H,t), \Delta t/2) \\ \widehat{T}_\Theta(H, t+\Delta t) & = \mathrm{MLP}_{w_{t,1},\dots,w_{t,M}}(\widehat{T}_\Theta(H,t+\Delta t/2), \Delta t/2) \end{cases} \tag{2.4}$$

for each $t = 0,\ \Delta t, \dots,\ (L-1)\Delta t$ with $\widehat{T}_\Theta(H, 0) = H$. We abbreviate the subscript $t = 0, \Delta t, \dots, (L-1)\Delta t$ by t and $j = 1, 2, \dots, M$ by j for simplicity. Here, $\Theta = \{\theta_{t,j}, w_{t,j}\}_{t,j}$ denotes all parameters in the Transformer model.

Throughout this paper, we treat D and N throughout the paper as bounded finite values, while M and L are treated as diverging, aligning with the setting of large-scale Transformers.

2.3 Gradient flow

For the l_2 regularization with $\lambda > 0$, we consider training the constructed Transformer model using the following λ-regularized risk objective:

$$\widehat{Q}(\Theta) = \widehat{R}(\Theta) + \frac{\lambda}{2ML}\sum_t \sum_{j=1}^{M} (\|\theta_{t,j}\|_2^2 + \|w_{t,j}\|_2^2), \tag{2.5}$$

with the population squared risk function defined as

$$\widehat{R}(\Theta) = \mathbb{E}_\mu\Big[\frac{1}{2}\Big(\mathrm{Read}[\widehat{T}_\Theta(H,1)] - y(H)\Big)^2\Big].$$

In Section 3.2, we will show that l_2-regularization on the parameter norms is essential for the well-posedness of the (Wasserstein) gradient flow to control parameter growth under our mild assumptions, even with a very small $\lambda > 0$. Similar strategies that consider necessary l_2 regularization are employed in [20] and [62]. Then, drawing on the methodologies in [6, 28, 40], our model processes the final output through a simple *read-out* function, $\mathrm{Read}[\cdot]$, extracting the $(d+1, N+1)$-th entry of its input. We propose that this read-out layer can be expanded to any linear mapping with bounded parameter norm without affecting the validity of our theoretical results.

To minimize the objective function (2.5), we implement the *standard gradient flow method* as follows:

Step 1. Initially, for each $t = 0, \Delta t, \dots, (L-1)\Delta t$, we sample M particles $\theta_{t,j}^{(0)}, w_{t,j}^{(0)}$ with $j \in [M]$ independently from $\rho_0(\theta, w|t)$, where ρ_0 is a pre-defined distribution with bounded support.

Step 2. Then, we update all parameters $\theta_{t,j}^{(\tau)}, w_{t,j}^{(\tau)}$ in the set $\Theta^{(\tau)} = \{\theta_{t,j}^{(\tau)}, w_{t,j}^{(\tau)}\}_{t,j}$ using gradient flow (scaled by ML), which is defined as follows:

$$\frac{d\theta_{t,j}^{(\tau)}}{d\tau} = -ML\nabla_{\theta_{t,j}}[\widehat{Q}(\Theta^{(\tau)})], \quad \frac{dw_{t,j}^{(\tau)}}{d\tau} = -ML\nabla_{w_{t,j}}[\widehat{Q}(\Theta^{(\tau)})]. \tag{2.6}$$

Define the function $\widehat{R}(H;\Theta) = \frac{1}{2}(\mathrm{Read}[\widehat{T}_\Theta(H,1)] - y(H))^2$, and the partial derivative $\widehat{p}_\Theta(H,t) = \partial\widehat{R}(H;\Theta)/\partial\widehat{T}_\Theta(H,t)^T$ for each $t = 0, \Delta t/2, \Delta t, \dots, (L-1)\Delta t, (L-1/2)\Delta t$. Refer to Appendix C.4 for the explicit formula of $\widehat{p}_\Theta(H,t)$. Using the chain rule, we derive the explicit form of the gradient flow as follows:

$$\frac{d\theta_{t,j}^{(\tau)}}{d\tau} = -\widehat{G}_f(\theta_{t,j}^{(\tau)}, \Theta^{(\tau)}, t), \quad \frac{dw_{t,j}^{(\tau)}}{d\tau} = -\widehat{G}_h(w_{t,j}^{(\tau)}, \Theta^{(\tau)}, t). \tag{2.7}$$

where

$$\widehat{G}_f(\theta, \Theta, t) = \frac{1}{2}\mathbb{E}_\mu\Big[\nabla_\theta \mathrm{Tr}\Big(f(\widehat{T}_\Theta(H,t),\theta)^T \widehat{p}_\Theta(H, t+\Delta t/2)\Big)\Big] + \lambda\theta,$$
$$\widehat{G}_h(w, \Theta, t) = \frac{1}{2}\mathbb{E}_\mu\Big[\nabla_w \mathrm{Tr}\Big(h(\widehat{T}_\Theta(H,t+\Delta/2),w)^T \widehat{p}_\Theta(H, t+\Delta t)\Big)\Big] + \lambda w$$

for $t = 0, \Delta t, \ldots, (L-1)\Delta t$.

3 Approximation by the mean-field limit

In this section, we present a rigorous approximation result that bridges Transformer models in (2.4) with their mean-field limit as continuous Transformers. Thus, the width M and depth L in our proposed model are treated as discretization of this continuous limit in the parameter space.

3.1 Assumptions

In addition, we introduce the norm $\|\cdot\|_{2-\mathrm{col}}$ as the maximum l_2 norm across all columns of a matrix. We proceed under several mild assumptions related to the data distribution and the encoders f and h.

Assumption 1 (Data regularity)**.** *There exists some constant $B > 0$ such that, for any $H \in \mathrm{supp}(\mu)$, we have $\max\{\|H\|_{2-\mathrm{col}}, y(H)\} \leq B$. In addition, a constant $K_y > 0$ ensures that $y(H)$ is K_y-Lipschitz continuous for $\|\cdot\|_F$ over $H \in \mathrm{supp}(\mu)$.*

Assumption 2 (Transformer particle growth bound)**.** *We assume that the gradient of $f(T, \theta)$ and $h(T, w)$ exists, and we define $\mathrm{ReLU}'(x) = \mathbf{1}\{x > 0\}$. Furthermore, we have*

i. $\|f(T,\theta)\|_{2-\mathrm{col}} \leq K\|T\|_{2-\mathrm{col}}(1 + \|\theta\| + \|\theta\|^2)$.

ii. For every $i \in [N+1]$, we have $\|\nabla_\theta f(T,\theta)_{:,i}\|_2 \leq \phi_P(\|T\|_{2-\mathrm{col}})(1 + \|\theta\|)$.

iii. $\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(T,\theta)]\|_2 \leq \phi_T(N, D, \|T\|_F)(1 + \|\theta\| + \|\theta\|^2)$.

for some continuous, monotonically increasing functions ψ_J, ϕ_T for every coordinate, and a constant $K > 0$. Similarly, if we replace f with h and θ with w, the same conditions apply.

Remark There are two key observations for Assumption 2. Firstly, it incorporates the $\|\cdot\|_{2-\mathrm{col}}$ norm, which is particularly useful for handling sequential inputs where each column represents a token. Secondly, as we consider higher-order multiplications between data and parameters, this assumption accommodates a broader range of self-attention encoders, such as the one in (2.3) with softmax or ReLU activation.

Assumption 3 (Locally Lipschitz continuous gradient in expectation)**.** *Besides Assumption 2, for any $L_T > 0$ and any L_T-Lipschitz continuous functions $T_1 = T_1(H)$ and $T_2 = T_2(H)$, for every $i \in [N+1]$, we have*

i. $\mathbb{E}_\mu\|\nabla_\theta f(T_1,\theta)_{:,i} - \nabla_\theta f(T_2,\theta)_{:,i}\|_2 \leq \phi_{PT}(\|\theta\|, K_T, L_T)\sup_H\|T_1 - T_2\|_{2-\mathrm{col}}$,

ii. $\mathbb{E}_\mu\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(T_1,\theta)] - \nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(T_1,\theta')]\|_2 \leq \phi_{TP}(N, D, \sup_H\|T_1\|_F, K_P, L_T)\|\theta - \theta'\|$

iii. $\mathbb{E}_\mu\|\nabla_\theta f(T_1,\theta)_{:,i} - \nabla_\theta f(T_1,\theta')_{:,i}\|_2 \leq \phi_{PP}(K_P, \sup_H\|T_1\|_{2-\mathrm{col}}, L_T)\|\theta - \theta'\|$,

iv. $\mathbb{E}_\mu\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(T_1,\theta)] - \nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(T_2,\theta)]\|_2 \leq \phi_{TT}(N, D, K_T, \|\theta\|, L_T)\sup_H\|T_1 - T_2\|_F$

for $K_T = \max\{\sup_H\|T_1\|_{2-\mathrm{col}}, \sup_H\|T_2\|_{2-\mathrm{col}}\}$, $K_P = \max\{\|\theta\|, \|\theta'\|\}$, and some continuous functions $\phi_{PT}, \phi_{TP}, \phi_{PP}, \phi_{TT}$ that are monotonically increasing for every coordinate. Similarly, if we replace f with h and θ with w, the same conditions apply.

Remark Assumption 3 states that functions are locally Lipschitz continuous in expectation, suitable for encoders that utilize ReLU functions and have second-order derivatives almost everywhere. This assumption is naturally satisfied if the activation has a locally Lipschitz continuous gradient.

Define $\mathcal{P}^2$ as the set of probability measures endowed with the Wasserstein-2 distance, where the Lipschitz continuity with respect to the depth holds, i.e. there exists some constant $C_\rho > 0$ such that $\|\rho(\cdot,t) - \rho(\cdot,t')\|_{\mathrm{BL}} \leq C_\rho|t - t'|$ for any $t, t' \in [0,1]$.

Choice of ρ_0 Suppose $\rho_0 \in \mathcal{P}^2$ satisfies that for any $t \in [0,1]$, the support of $\rho_0(\cdot,\cdot,t)$ is contained within the set $\{(\theta, w) : \|\theta\|^2 + \|w\|^2 \leq R^2\}$ for a constant R. Additionally, for each $t \in [0,1]$, it holds that $\int_{\theta,w} \rho_0(\theta, w, t) d(\theta, w) = 1$. This condition suits common bounded support distributions, and a natural choice is a uniform distribution across a disk with radius R for each $t \in [0,1]$.

3.2 Continuous Transformer and Wasserstein gradient flow

Drawing inspiration from [41] and [60], which suggest that deep residual networks behave like ensembles of residual networks locally, we apply a similar manipulation to formulate the continuous version of (2.4). Consider the following continuous version $T_\rho(H,t) \in \mathbb{R}^{D\times(N+1)}$, governed by the following continuous ODE that *averages the two encoders*:

$$\dot{T}_\rho(H,t) = \int_{\theta,w} \frac{f(T_\rho(H,t),\theta) + h(T_\rho(H,t),w)}{2} \rho(\theta,w,t) d(\theta,w), \quad T_\rho(H,0) = H \tag{3.1}$$

In (3.1), each encoder f or h is conceptualized as a particle, and we consider the distribution of these particles denoted as $\rho(\theta, w, t)$. For any $\rho \in \mathcal{P}^2$ that have a bounded support, the well-posedness of $T_\rho(H,t)$ that satisfies the Transformer ODE (3.1) is shown in Proposition C.1. Transitioning to the framework with continuous Transformers, our objective shifts to minimizing the l_2 risk function with regularization on the second moment of ρ as follows:

$$Q(\rho) = R(\rho) + \frac{\lambda}{2}\int_0^1 \int_{\theta,w} (\|\theta\|_2^2 + \|w\|_2^2)\rho(\theta,w,t)d(\theta,w)dt, \tag{3.2}$$

with

$$R(\rho) = \mathbb{E}_\mu\Big[\frac{1}{2}\Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big)^2\Big]. \tag{3.3}$$

Define $p_\rho(H,t) \in \mathbb{R}^{D\times(N+1)}$, the partial derivative of $R(\rho)$ relative to $T_\rho(H,t)$ at a local query point H, as the solution derived in Appendix C.4 using the classical adjoint sensitivity method [51]:

$$\mathrm{vec}[p_\rho(H,t)]^T = \Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big) \exp\Big(\int_t^1 \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,t),\beta)\rho(\beta,t)d\beta dt]\Big)_{DN+d+1,:}.$$

Using this, we can compute the functional derivative to ρ as follows:

$$\frac{\delta Q}{\delta \rho}(\theta,w,t) = \mathbb{E}_\mu\Big[\mathrm{Tr}\Big(\Big[\frac{f(T_\rho(H,t),\theta) + h(T_\rho(H,t),w)}{2}\Big]^T p_\rho(H,t)\Big)\Big] + \frac{\lambda}{2}(\|\theta\|_2^2 + \|w\|_2^2). \tag{3.4}$$

The following Proposition claims that $\frac{\delta Q}{\delta \rho}$ is indeed the derivative with respect to ρ (specifically, the Fréchet derivative [26]) for the functional $Q(\rho)$.

Proposition 3.1 (Functional derivative to ρ)**.** *Under Assumptions 1 and 2, for any pair $\rho, \nu \in \mathcal{P}^2$ that have bounded supports, we have*

$$Q(\rho + \eta(\nu - \rho)) = Q(\rho) + \eta\Big\langle \frac{\delta Q}{\delta \rho}, \nu - \rho\Big\rangle + o(\eta),$$

where $\frac{\delta Q}{\delta \rho}$ is defined in (3.4)*, and $\langle \frac{\delta Q}{\delta \rho}, \nu - \rho\rangle = \int_0^1 \int_{(\theta,w)} \frac{\delta Q}{\delta \rho} \cdot (\nu - \rho) d(\theta, w)dt \in \mathbb{R}$.*

Now, we are in a position to display the gradient flow of ρ in the Wasserstein metric [15], given by a McKean-Vlasov type equation [4, 32, 48, 50]. Specifically, we study the following partial differential equation of the distribution $\rho^{(\tau)}(\theta, w, t)$:

$$\begin{aligned}\frac{d\rho^{(\tau)}(\theta,w,t)}{d\tau} &= \mathrm{div}_{(\theta,w)}\Big(\rho^{(\tau)} \nabla_{(\theta,w)} \frac{\delta Q}{\delta \rho}\Big|_{\rho=\rho^{(\tau)}}\Big) \\ &= \mathrm{div}_\theta\Big(\rho^{(\tau)} G_f(\theta,\rho^{(\tau)},t)\Big) + \mathrm{div}_w\Big(\rho^{(\tau)} G_h(w,\rho^{(\tau)},t)\Big),\end{aligned} \tag{3.5}$$

where $\rho^{(0)} = \rho_0$, div is the divergence operator, and the gradient functions are defined as

$$G_f(\theta,\rho,t) = \frac{1}{2}\mathbb{E}_\mu\Big[\nabla_\theta \mathrm{Tr}\Big(f(T_\rho(H,t),\theta)^T p_\rho(H,t)\Big)\Big] + \lambda\theta,$$
$$G_h(w,\rho,t) = \frac{1}{2}\mathbb{E}_\mu\Big[\nabla_w \mathrm{Tr}\Big(h(T_\rho(H,t),w)^T p_\rho(H,t)\Big)\Big] + \lambda w.$$

Propositions D.1 and 3.2 provide the well-posedness of both gradient flow and Wasserstein gradient flow respectively. In both propositions, a $\lambda > 0$ is essential to stabilize the optimization process by controlling both the maximum and average norms across all parameters. If λ is set to 0, it is only possible to establish the well-posedness of (3.5) over a finite maximal interval [41]. Similar adjustments to regularize the risk function are also noted in [20].

Proposition 3.2 (Existence and uniqueness of Wasserstein gradient flow)**.** *Under Assumptions 1 and 2, there exists a unique solution $(\rho^{(\tau)})_{\tau\geq 0} \in \mathcal{P}^2 \times \mathbb{R}$ with $\rho^{(0)} = \rho_0$ for (3.5). Additionally, for any $\tau \geq 0$, we have*

i. $\rho^{(\tau)}$ has a bounded support $\{\theta, w : \|\theta\|^2 + \|w\|^2 \leq R_\tau\} \times [0,1]$, where $R_\tau = (R+1)\exp(C_R\tau) - 1$ for some universal constant C_R that only depends on N, D, λ and the assumptions.

ii. $\int_0^1 (\|\theta\|^2+\|w\|^2)\rho^{(\tau)}(\theta,w,t)d(\theta,w)dt \leq A_0^2$, where $A_0 := R^2+\lambda^{-1}(2B^2+2B^2\exp(K(1+R+R^2))^2)$.

iii. $\int_{(\theta,w)} \rho^{(\tau)}(\theta,w,t)d(\theta,w) = 1$ for any $t \in [0,1]$.

3.3 Approximation of large-scale Transformer

In this section, we discuss the general results associated with approximating our discrete Transformer model to its mean-field limit. First, we highlight that the minimization of the risk function with discretization, whether or not regularization is included, closely approximates the minimal risk achievable by continuous models.

Proposition 3.3 (Global minimum approximation of discretization)**.** *Under Assumptions 1 and 2, we define $\mathcal{P}^{2,r}$ as the set of distributions in $\mathcal{P}^2$ concentrated on $\{(\theta,w) : \|\theta\|^2 + \|w\|^2 \leq r^2\} \times [0,1]$. for any $r > 0$. Then there exists a constant C dependent on N, D, r and the parameters of the assumptions such that*

$$\inf_{\Theta} \widehat{R}(\Theta) \leq \inf_{\rho\in\mathcal{P}^{2,r}} R(\rho) + C\Big(L^{-1} + \sqrt{\frac{\log(L+1)}{M}}\Big),$$
$$\inf_{\Theta} \widehat{Q}(\Theta) \leq \inf_{\rho\in\mathcal{P}^{2,r}} Q(\rho) + C(1+\lambda)\Big(L^{-1} + \sqrt{\frac{\log(L+1)}{M}}\Big).$$

Proposition 3.3 specifies that the distributions under consideration must have bounded support. While it is typically challenging to confirm whether the minimal risk is indeed achieved on a distribution with bounded support, this assumption is justified as λ regulates parameter norms, implicitly encourages solutions residing in a compact region of the parameter space.

We now present the main theorem concerning the convergence of the gradient flow process to the Wasserstein gradient flow as outlined in (3.5). The proof with detailed explanation of the techniques used in Theorem 3.1 is provided in Appendix D.

Theorem 3.1 (Gradient flow approximation of discretization)**.** *Define the empirical distribution as $\hat{\rho}^{(\tau)} := \frac{1}{ML}\sum_t \sum_{j=1}^M \delta(\theta_{t,j}^{(\tau)}, w_{t,j}^{(\tau)}, t)$ for any $\tau \geq 0$. Under Assumptions 1-3, we have that $(\hat{\rho}^{(\tau)})_{\tau\geq 0}$ weakly converges to $(\rho^{(\tau)})_{\tau\geq 0}$ almost surely along any sequence such that $L \to \infty, M/\log L \to \infty$. Moreover, for any fixed $\tau > 0$ and any $\delta > 0$, with probability at least $1 - 3\exp(-\delta)$, we have*

i. $\sup_{s\in[0,\tau]}|\mathrm{Read}[\widehat{T}_{\Theta^{(s)}}(H,t)] - \mathrm{Read}[T_{\rho^{(s)}}(H,t)]| \leq C\Big(L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}\Big)$

ii. $\sup_{s\in[0,\tau]}|\widehat{R}(\Theta^{(s)}) - R(\rho^{(s)})| \leq C\Big(L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}\Big)$

iii. $\sup_{s\in[0,\tau]}|\widehat{Q}(\Theta^{(s)}) - Q(\rho^{(s)})| \leq C\Big(L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}\Big)$

for some universal constant C that depends on on N, D, τ, λ and the parameters of the assumptions.

Theorem 3.1 significantly advances our understanding by controlling the difference regarding both the Transformer output, the risk function, and the regularized risk function. It's noted that the difference bound in the model's approximation may increase, possibly exponentially [19, 20, 45], as the time horizon extends. As argued in [45], such behavior may be inherent to the systems being modeled and not readily improvable.

Additionally, the technical uniqueness and innovation of this theorem contrast sharply with previous results from overparametrized ResNet models. Our analysis distinguishes itself in two ways. First, our discrete Transformer model (2.4) uniquely splits the averaged encoder $(f + h)/2$ into two distinct blocks with encoders f and h. Second, we demonstrate uniform error control over any finite time interval $[0, \tau]$, enabling continuous monitoring of maximum error across the gradient flow's trajectory. In contrast, models in prior studies such as [19, 20] restricts the error analysis to a specific $s \in [0, \tau]$.

4 Global convergence of gradient flow

In this section, we explore the optimization problem for gradient flow in the context of the discrete Transformer model, focusing on our general global convergence results.

4.1 An additional assumption

To ensure the global convergence of gradient flow for our discrete Transformer model, we introduce the following assumption. While influenced by the work in [15, 19, 20, 41], our assumption is uniquely tailored to the context of Transformers:

Assumption 4. *There exists a pair $(g, \alpha) \in \{(f, \theta), (h, w)\}$ with a partition $\alpha = (\alpha_1, \alpha_2)$ such that*

i. (Partial 1-homogeneity) for any $T \in \mathbb{R}^{D\times(N+1)}$ and $c \in \mathbb{R}$, we have $g(T, c\alpha_1, \alpha_2) = cf(T, \alpha_1, \alpha_2)$.

ii. (Universal kernel) a compact set $\mathcal{K} \subset \mathbb{R}^{\dim(\alpha_2)}$ ensures that the span of $\Big\{g(\cdot, \alpha) : \alpha \in \mathbb{R}^{\dim(\alpha_1)} \times \mathcal{K}\Big\}$ is dense in $\mathcal{C}(\|T\|_{2-\mathrm{col}} \leq B, \mathbb{R}^{D\times(N+1)})$ for any $B > 0$.

We emphasize that the universal kernel property, as discussed in [46], closely relates to the universal approximation abilities. Under our assumption, we require the universal approximation capabilities of *either* the self-attention encoder or the feed-forward encoder.

The universal kernel property of the feed-forward layer encoder h is well-established, particularly in two-layer neural network contexts [66]. Conversely, the universal approximation abilities of self-attention layers is a frontier research area, which, while not extensively covered in this paper, holds significant potential. Often labeled as "memorization capacity", this area is recently explored across multiple studies [23, 27, 33, 34, 43, 56, 65]. The interconnection between approximation abilities and memorization capacities is established in [33]. Notably, [43] investigated the expressive capabilities of one single multi-head softmax self-attention layer, thereby potentially validating our assumptions.

Finally, we posit that the universal kernel applies to α_2 within a compact set, as the function's scale can be moderated by the homogeneous part α_1. In scenarios where α_2 and $\mathcal{K}$ are absent, our assumption simplifies to that in [41], characterized by complete homogeneity. Conversely, in the absence of the α_1 component, our framework aligns with [20] which necessitates a more stringent support condition for $\mathcal{K}$, as detailed later in Theorem 4.1.

4.2 Global convergence result

In this section, we establish the convergence properties of the optimization task for discrete Transformers through gradient flow dynamics.

Theorem 4.1 (Global convergence up to λ)**.** *Suppose that Assumptions 1-4 hold, and the Wasserstein gradient flow $(\rho^{(\tau)})_{\tau\geq 0}$ weakly converges to some $\rho_\infty \in \mathcal{P}^2$. If for some constant $R_\infty > 1$, the following two conditions hold:*

i. $(\rho^{(\tau)})_{\tau\geq 0}$ is concentrated on $\{\theta, w : \|\theta\|^2 + \|w\|^2 \leq R_\infty^2\} \times [0, 1]$ when τ is sufficiently large.

ii. If Assumption 4 holds with $(g,\alpha) = (f,\theta)$, we assume there exists a $t^* \in [0,1]$ such that the connected set $\mathrm{supp}(\rho_\infty(\cdot,t^*)) \supset \mathcal{D}\times\mathcal{K}\times\{w_0\}$, for some $w_0 \in \mathbb{R}^{\dim(w)}$ and $\mathcal{D} \subset \mathbb{R}^{\dim(\theta_1)}$ that separates $\{\theta_1 : \|\theta_1\| = 1/R_\infty\}$ and $\{\theta_1 : \|\theta_1\| = R_\infty\}$.

ii′. If Assumption 4 holds with $(g,\alpha) = (h,w)$, we assume there exists a $t^* \in [0,1]$ such that the connected set $\mathrm{supp}(\rho_\infty(\cdot,t^*)) \supset \times\{\theta_0\}\times\mathcal{K}\times\mathcal{D}$, for some $\theta_0 \in \mathbb{R}^{\dim(\theta)}$ and $\mathcal{D} \subset \mathbb{R}^{\dim(w_1)}$ that separates $\{w_1 : \|w_1\| = 1/R_\infty\}$ and $\{w_1 : \|w_1\| = R_\infty\}$.

Then, for any $\epsilon > 0$, there exists some $\tau_0 > 0$ such that

$$\sup_{\tau\geq\tau_0} \widehat{R}(\Theta^{(\tau)}) \leq \epsilon + C_1\Big(L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}\Big) + C_2\lambda$$

with probability at least $1-3\exp(-\delta)$ for any $\delta > 0$. Here, C_1 is some universal constant dependent only on N, D, τ_0, λ and the parameters of the assumptions, while C_2 depends only on N, D, R_∞ and the parameters of the assumptions.

Theorem 4.1 depicts the behavior of the risk function $\widehat{R}(\Theta^{(\tau)})$ as the training duration τ is sufficiently large. Specifically, $\widehat{R}(\Theta^{(\tau)})$ asymptotically approaches zero as both $L \to \infty$ and $M/\log L \to \infty$, with an additional term that scales with λ. This additional term attributes to the incorporation of a λ-weighted penalty on the norm of the parameters in our training objective $\widehat{Q}$. Consequently, by selecting an appropriately small $\lambda > 0$, the risk approximates zero, demonstrating global convergence to the minimum of $\widehat{R}$.

In addition, Theorem 4.1 posits some additional assumptions: the weak convergence of $\rho^{(\tau)}$, the long-time uniform boundedness, and the separation property for α_1 with the support expansion of α_2 to $\mathcal{K}$. Similar assumptions are made in the literature of deep model optimization theory [19, 20, 41]. While these types of assumptions are typically challenging to justify, we provide high-level justifications for them in Appendix C.5, deferring detailed verification to future research.

We then present a corollary that directly follows from Theorem 4.1:

Corollary 4.1. *Continuing with the notations and assumptions from Theorem 4.1, suppose $\lambda \leq C_\lambda\epsilon$ for some constant $C_\lambda > 0$. Then, for any $\delta > 0$, constants $\tau_0, L_0, K_0 > 0$ can be found such that:*

$$\sup_{\tau\geq\tau_0, L>L_0, M/\log L>K_0} \widehat{R}(\Theta^{(\tau)}) \leq (1/2 + C_2C_\lambda)\epsilon \tag{4.1}$$

with probability at least $1-3\exp(-\delta)$. Here, L_0 scales as $\Omega(\epsilon^{-1})$ and K_0 as $\Omega((1+\delta)\epsilon^{-2})$. Notably, if $C_\lambda \leq (2C_2)^{-1}$, the upper bound in (4.1) is less than or equal to ϵ.

Corollary 4.1 claims that with a fixed $\delta > 0$, to achieve an order of ϵ-close approximation, it suffices to set $L = \Theta(\epsilon^{-1})$ and $M = \Theta(\epsilon^{-2}\log(\epsilon^{-1}))$. This error rate could be directly derived from the approximation results in Theorem 3.1.

5 Conclusion

We conclude by summarizing our key contributions and suggesting future research directions. This paper establishes the global convergence of large-scale Transformer models through gradient flow dynamics, providing a thorough theoretical foundation. Our analysis, focused on the mean-field limit with infinite width and depth, shifts optimization from parameter space to distributional probability measures. We present two main theorems: one confirming the close approximation between discrete and continuous gradient flows, and another demonstrating global convergence, highlighting that basic optimization methods can successfully navigate complex landscapes to find optimal solutions. The techniques and results from this study lay the groundwork for further exploration into Transformer optimization. Future work could explore direct gradient descent with specific focus on step sizes, and expand on the in-context learning approximation capabilities of Transformers, as initiated by [6]. Additionally, it's crucial to rigorously assess under what conditions can self-attention layers serve as universal kernels to enhance our theoretical understanding, and to determine the generalization error bounds of Transformers trained on finite samples. These directions promise to deepen the theoretical and practical insights into Transformer models.

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2022.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, 2019.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. In metric spaces and in the space of probability measures. In *Gradient Flows*, 2005.

[5] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332, 2019.

[6] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023.

[7] Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. On global convergence of resnets: From finite to infinite width using linear parameterization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[8] James Bernhard. Alternatives to the scaled dot product for attention in the transformer neural network architecture, 2023.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[10] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

[11] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[12] Yihang Chen, Fanghui Liu, Yiping Lu, Grigorios Chrysos, and Volkan Cevher. Generalization of scaled deep resnets in the mean-field regime. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *arXiv: Learning*, 2020.

[14] Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation and controllability of deep neural networks. *arXiv preprint arXiv:2309.06015*, 2023.

[15] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3040–3050, Red Hook, NY, USA, 2018. Curran Associates Inc.

[16] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

[17] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7480–7512. PMLR, 23–29 Jul 2023.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[19] Zhiyan Ding, Shi Chen, Qin Li, and Stephen Wright. On the global convergence of gradient descent for multi-layer resnets in the mean-field regime, 2021.

[20] Zhiyan Ding, Shi Chen, Qin Li, and Stephen Wright. Overparameterization of deep resnet: Zero loss and mean-field analysis. *Journal of Machine Learning Research*, 23:48–1, 2022.

[21] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.

[22] Weinan E, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):1–41, 2019.

[23] Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms, 2022.

[24] Cong Fang, Hanze Dong, and Tong Zhang. Over parameterized two-level neural networks can learn near optimal feature representations. *ArXiv*, abs/1910.11508, 2019.

[25] Cong Fang, Yihong Gu, Weizhong Zhang, and Tong Zhang. Convex formulation of overparameterized deep neural networks. *IEEE Transactions on Information Theory*, 68(8):5340–5352, 2022.

[26] Bela A. Frigyik, Santosh Srivastava, and Maya R. Gupta. Introduction to functional derivatives. UWEE Tech Report 2008-0001, University of Washington Department of Electrical Engineering, 2008.

[27] Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[28] Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

[30] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[31] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 37822–37836. Curran Associates, Inc., 2022.

[32] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker-planck equation. *SIAM Journal on Mathematical Analysis*, 29, 04 2000.

[33] Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators?, 2024.

[34] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention, 2021.

[35] Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[36] Qianxiao Li, Long Chen, Cheng Tai, and Weinan E. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18(165):1–29, 2018.

[37] Qianxiao Li and Shuji Hao. An optimal control approach to deep learning and applications to discrete-weight neural networks. In *International Conference on Machine Learning*, pages 2985–2994. PMLR, 2018.

[38] Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

[39] Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pages 19689–19729. PMLR, 2023.

[40] Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024.

[41] Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6426–6436. PMLR, 13–18 Jul 2020.

[42] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[43] Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, 2019.

[45] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[46] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *J. Mach. Learn. Res.*, 7:2651–2667, 2006.

[47] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles, 2017.

[48] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Particle dual averaging: optimization of mean field neural network with global convergence rate analysis*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114010, nov 2022.

[49] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[50] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[51] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishechenko. The mathematical theory of optimal processes. *Zamm-zeitschrift Fur Angewandte Mathematik Und Mechanik*, 43:514–515, 1963.

[52] H. Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer, 1996.

[53] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[54] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2016.

[55] Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'Eté de Probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

[56] Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33416–33447. PMLR, 23–29 Jul 2023.

[57] Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *International Conference on Machine Learning*, 2020.

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[60] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

[61] C. Villani. *Optimal Transport, Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 2008.

[62] Colin Wei, J. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Neural Information Processing Systems*, 2018.

[63] E Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

[64] E Weinan, Chao Ma, and Lei Wu. Machine learning from a continuous viewpoint, i. *Science China Mathematics*, 63:2233 – 2266, 2019.

[65] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *ArXiv*, abs/1912.10077, 2019.

[66] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[67] Ruiqi Zhang, Spencer Frei, and Peter Bartlett. Trained transformers learn linear models in-context. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[68] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, Oct 2019.

A Overview of Appendix

The appendix is organized as follows:

- **Appendix B:** Additional related works are discussed.
- **Appendix C:**
  - In Appendix C.1, additional notations and preliminary details are introduced.
  - In Appendix C.2, we show the proof of Proposition C.1, concerning the existence and uniqueness of the continuous Transformer ODE (3.1).
  - In Appendix C.3, useful lemmas for the main proofs are detailed.
  - In Appendix C.4, the explicit formulas for $p_\rho(H,t)$ and $\widehat{p}_\Theta(H,t)$ are explored via the adjoint sensitivity method.
  - In Appendix C.5, high-level explanations are provided to substantiate the assumptions made in Theorem 4.1.
- **Appendix D:** Includes proofs of main results from Section 3.
- **Appendix E:** Includes proofs of main results from Section 4.
- **Appendix F:** Includes proofs of all auxiliary technical results mentioned in Appendix C-E.

B Additional related work

Theory of Transformers. Some very recent works have studied theoretical properties of Transformer models from different aspects. [67, 29] studied the in-context learning guarantees for single-layer Transformers to perform linear regression predictions after being trained with linear regression example tasks. [1, 6, 28] studied the in-context learning capability of Transformers through the function approximation point of view, and demonstrated that there exists Transformers with specific parameter configurations that can perform in-context learning. [31, 39] investigated how single-layer Transformers can be trained to learn simple image models and topic models respectively.

Global convergence of fully connected neural networks. A line of recent works have studied the global convergence of (stochastic) gradient descent in training overparameterized neural networks in the mean-field regime [15, 45, 44, 62, 24, 25]. They consider the limit of the neural network as the width of the network at each layer goes to infinity, and models the limit of the network as a functional of the distribution of network parameters. A separate line of works also established the global convergence guarantees for training overparameterized neural networks in the "neural tangent kernel" regime [30, 3, 21, 68, 16, 2, 5, 10], where the gradient descent training iterates are asymptotically equivalent to the training iterates of kernel regression based on the neural tangent kernel.

Connection between ordinary differential equation models and infinite-depth ResNets. Our work is also closely related to the recent literature aiming to understand ResNets by analyzing their connections to ordinary differential equations [63, 11, 37, 36, 64, 22, 19, 41, 20, 38, 7, 14, 12]. Specifically, [63, 11, 38, 14] studied the approximation of flow-based networks via discrete networks. [37, 36, 64, 22, 19, 41, 20, 38, 7] studied the optimization of the infinite-depth and infinite-width ResNets. [12] studied the generalization properties of the ResNet trained in the mean-field regime.

C Proof setup

C.1 Additional technical notations

Define

$$\beta := (\theta, w)^T, \quad g(T,\beta) := \frac{f(T,\theta) + h(T,w)}{2}.$$

Thus, $\frac{\delta Q}{\delta \rho}$ could be expressed as

$$\frac{\delta Q}{\delta \rho}(\beta,t) = \mathbb{E}_\mu\Big[\mathrm{Tr}\Big(\Big[g(T_\rho(H,t),\beta)\Big]^T p_\rho(H,t)\Big)\Big] + \frac{\lambda}{2}\|\beta\|_2^2. \tag{C.1}$$

Additionally, we can combine G_f with G_h, and $\widehat{G}_f$ with $\widehat{G}_h$ to reformulate as

$$G(\beta,\rho,t) = \mathbb{E}_\mu\Big[\nabla_\beta \mathrm{Tr}\Big(g(T_\rho(H,t),\beta)^T p_\rho(H,t)\Big)\Big] + \lambda\beta, \tag{C.2}$$

and

$$\widehat{G}(\beta,\Theta,t) = \mathbb{E}_\mu\Big[\nabla_\beta \mathrm{Tr}\Big(\begin{Bmatrix} f(\widehat{T}_\Theta(H,t),\theta)/2 \\ h(\widehat{T}_\Theta(H,t+\Delta t/2),w)/2 \end{Bmatrix}^T \begin{Bmatrix} \widehat{p}_\Theta(H,t+\Delta t/2) \\ \widehat{p}_\Theta(H,t+\Delta t) \end{Bmatrix}\Big)\Big] + \lambda\beta. \tag{C.3}$$

Remark 1. *To facilitate the proof, we restate Assumptions 2 and 3 for $g(T,\beta)$. Under Assumption 2, the gradient of $g(T,\beta)$ respect to T and β exists. Additionally, we have*

Under Assumption 2:

i. $\|g(T,\beta)\|_{2-\mathrm{col}} \le K\|T\|_{2-\mathrm{col}}(1+\|\beta\|+\|\beta\|^2)$

ii. For every $i\in[N+1]$, we have $\|\nabla_\beta g(T,\beta)_{:,i}\|_2 \le \phi_P(\|T\|_{2-\mathrm{col}})(1+\|\beta\|)$

iii. $\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T,\beta)]\|_2 \le \phi_T(N,D,\|T\|_F)(1+\|\beta\|+\|\beta\|^2)$

Under Assumption 3: For any $L_T>0$ and any L_T-Lipschitz continuous functions $T_1=T_1(H)$ and $T_2=T_2(H)$, for every $i\in[N+1]$, we have

$i.$ $\mathbb{E}_\mu\|\nabla_\theta g(T_1,\beta)_{:,i} - \nabla_\theta g(T_2,\beta)_{:,i}\|_2 \le \phi_{PT}(\|\theta\|,K_T,L_T)\sup_H\|T_1-T_2\|_{2-\mathrm{col}},$

$ii.$ $\mathbb{E}_\mu\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_1,\beta)] - \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_1,\beta')]\|_2 \le \phi_{TP}(N,D,\sup_H\|T_1\|_F,K_P,L_T)\|\theta-\theta'\|$

$iii.$ $\mathbb{E}_\mu\|\nabla_\theta g(T_1,\beta)_{:,i} - \nabla_\theta g(T_1,\beta')_{:,i}\|_2 \le \phi_{PP}(K_P,\sup_H\|T_1\|_{2-\mathrm{col}},L_T)\|\theta-\theta'\|,$

$iv.$ $\mathbb{E}_\mu\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_1,\beta)] - \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_2,\beta)]\|_2 \le \phi_{TT}(N,D,K_T,\|\theta\|,L_T)\sup_H\|T_1-T_2\|_F$

Verifying all these results above only needs the basic triangle inequality of general norms, so we omit the trivial proof. We will apply them directly throughout the proofs of results. Additionally, we omit writing L_T for simplicity in the proof, as all functions applied to Assumption 3 will be Lipschitz continuous with some universally bounded Lipschitz constant.

Next, we introduce some additional technical notations. Denote the identity matrix with d-dimension as I_d. Define the sample space $\Omega := \mathbb{R}^{\dim\beta}\times[0,1]$, and $\mathcal{P}(\Omega)$ as the probability measure space defined on Ω. For any $\Theta = \{\beta_{t,j}\}_{t/\Delta t+1\in[L],j\in[M]}$ and $\widetilde{\Theta} = \{\widetilde{\beta}_{t,j}\}_{t/\Delta t+1\in[L],j\in[M]}$, define $d(\Theta,\widetilde{\Theta}) = \sum_t\sum_{j=1}^M\|\beta_{t,j}-\widetilde{\beta}_{t,j}\|$. Define the local risk function as

$$R(H;\rho) = \frac{1}{2}\Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big)^2.$$

Define the nested family of compact subsets $(P_r)_{r>0}$ as

$$P_r := \{\beta : \|\beta\|\le r\}\times[0,1], \quad \forall r>0.$$

For any $\rho,\nu\in\mathcal{P}(\Omega)$ and $p\ge 1$, define the l_p distance $\|\rho-\nu\|_p$ as

$$\|\rho-\nu\|_p = \Big(\int_0^1\int_\beta |\rho(x)-\nu(x)|^p d\beta dt\Big)^{1/p}.$$

Specifically, when $p=1$, we have

$$\begin{aligned} W_1(\rho,\nu) &= \sup\Big\{\int_0^1\int_\beta f(\rho-\nu)d\beta dt : f \text{ is } 1\text{−Lipschitz}, \mathrm{f}(\mathbf{0},0)=0\Big\} \\ &\le \sup\Big\{\int_0^1\int_\beta |f||\rho-\nu|d\beta dt : f \text{ is } 1\text{−Lipschitz}, \mathrm{f}(\mathbf{0},0)=0\Big\} \\ &\le (r+1)\|\rho-\nu\|_1 \end{aligned}$$

for any $\rho,\nu\in\mathcal{P}^2$ concentrated on P_r. For simplicity, any H discussed throughout this paper is assumed to lie within $\mathrm{supp}(\mu)$.

C.2 Transformer ODE existence and uniqueness

In this section, we establish the existence and uniqueness of the solution $T_\rho(H,t)$ to the ODE presented in (3.1) for any H, given that $\rho \in \mathcal{P}^2$ is concentrated on a bounded support, specifically, P_r for some $r > 0$. This following proposition forms the cornerstone of the subsequent technical analyses:

Proposition C.1 (Existence and uniqueness of Transformer ODE)**.** *Under Assumptions 1 and 2, for any $\rho \in \mathcal{P}^2$ that has a bounded support, there exists a unique solution of* (3.1) *on $t \in [0,1]$ that is Lipschitz continuous with respect to (H,t).*

Initially, we demonstrate that the integral $\int_\beta \rho(\beta,t)d\beta$ is bounded. According to the definition $\mathcal{P}^2$, it follows that

$$|\int_\beta \rho(\beta,t)d\beta - \int_\beta \rho(\beta,t')d\beta| \leq C_\rho |t-t'| \tag{C.4}$$

for any $t, t' \in [0,1]$. Integrating (C.4) over $t_2 \in [0,1]$ obtains

$$\begin{aligned}\int_\beta \rho(\beta,t)d\beta =& 1 + \int_\beta \rho(\beta,t)d\beta - \int_0^1 \int_\beta \rho(\beta,t')d\beta dt' \\ \leq& 1 + C_\rho \int_0^1 |t-t'|dt' \\ \leq& 1 + C_\rho/2.\end{aligned} \tag{C.5}$$

For the remainder of the technical proof, we will employ (C.5) without additional elaboration.

Proof of Proposition C.1. Consider any vector β such that $\|\beta\| \leq r$. Define $F(T,t) := \int_\beta g(T,\beta)\rho(\beta,t)d\beta$. For T within the rectangle $\{T : \|T-H\|_{\max} \leq \delta\}$, where $\delta > 0$ is bounded, both $\|T\|_{2-\mathrm{col}}$ and $\|T\|_F$ are also bounded. Given Assumption 2 (i), $g(T,\beta)$ is universally bounded by some constant $K_{\delta,r}$. Moreover, under Assumption 2 (ii) and (iii), $g(T,\beta)$ is Lipschitz continuity with some constant $L_{\delta,r}$. Hence, within the rectangle $\{T : \|T-H\|_{\max} \leq \delta\} \times [0,1]$ the following properties hold:

$$|F(T,t_1) - F(T,t_2)| \leq \max\{K_{\delta,r}, L_{\delta,r}\}\|\rho(\cdot,t_1) - \rho(\cdot,t_2)\|_{\mathrm{BL}} \leq C_\rho \max\{K_{\delta,r}, L_{\delta,r}\}|t_1 - t_2|. \tag{C.6}$$

which indicates that $F(T,t)$ is continuous with respect to t within the rectangle $\{T : \|T-H\|_{\max} \leq \delta\} \times [0,1]$.

Moreover, within the bounded region $\{T : \|T-H\|_{\max} \leq \delta\}$, Assumption 2 (iii) ensures that $\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T,\beta)]\|_2$ bounded. Consequently, $g(T,\beta)$ is Lipschitz-continuous with respect to T for $\|\cdot\|_F$. Denote this Lipschitz constant by $L'_{\delta,r}$. Therefore, for any $T, T' \in \{T : \|T-H\|_{\max} \leq \delta\}$ and $t \in [0,1]$, it follows that

$$|F(T,t) - F(T',t)| \leq L'_{\delta,r}\|T-T'\|_F \int_\beta \rho(\beta,t)d\beta \leq L'_{\delta,r}(1 + C_\rho/2)\|T-T'\|_F, \tag{C.7}$$

which deduces the Lipschitz continuity of $F(T,t)$ with respect to T for $\|\cdot\|_F$. Invoking the Picard-Lindelöf Theorem, there exists some $\epsilon > 0$ such that the initial value problem

$$\dot{T}(H,t) = F(T,t), \quad T(H,0) = H$$

has a unique solution on $t \in [0,\epsilon]$. Given that this claim holds for any H, the standard ODE Extensibility Theorem guarantees a continuation of $T(t)$ to a maximal interval of existence, denoted as $[0, t_{\max}]$.

Assume by contradiction that $t_{\max} < 1$. From (3.1) and Assumption 2(i), for any $t \in [0, t_{\max}]$ we see that

$$\begin{aligned}\frac{d}{dt}\|T_\rho(H,t)\|_{2-\mathrm{col}} \leq \|\dot{T}_\rho(H,t)\|_{2-\mathrm{col}} &= \|\int_\beta g(T_\rho(H,t),\beta)\rho(\beta,t)d\beta\|_{2-\mathrm{col}} \\ &\leq \int_\beta \|g(T_\rho(H,t),\beta)\|_{2-\mathrm{col}}\rho(\beta,t)d\beta \\ &\leq \int_\beta K(1+||\beta||_2+||\beta||_2^2)\rho(\beta,t)\|T_\rho(H,t)\|_{2-\mathrm{col}}d\beta.\end{aligned} \tag{C.8}$$

Therefore, by the Grönwall's inequality, we have

$$\begin{aligned}\|T_\rho(H,t_{\max})\|_{2-\mathrm{col}} &\leq \|T_\rho(H,0)\|_{2-\mathrm{col}} \exp(\int_0^t \int_\beta K(1+||\beta||_2+||\beta||_2^2)\rho(\beta,s)d\beta ds) \\ &\leq \|H\|_{2-\mathrm{col}} \exp(K(1+C_\rho/2+r+r^2)t_{\max}) < \infty.\end{aligned}$$

This presents a contradiction to the notion that $t_{\max} < 1$. This is because, By reapplying the local Picard-Lindelöf Theorem using the state $T(H, t_{\max})$ as the new initial condition, we can extend the interval of existence beyond $t_{\max}$. Consequently, we must conclude that $t_{\max} = 1$, and the existence and uniqueness follows.

In the final part of our proof, we demonstrate that $T_\rho(H,t)$ is Lipschitz continuous with respect to (H,t) for $H \in \mathrm{supp}(\mu)$ and any $t \in [0,1]$. Given that $T_\rho(H,t)$ is universally bounded within $H \in \mathrm{supp}(\mu)$ and any $t \in [0,1]$, we only need to focus on establishing its Lipschitz continuity with respect to H and t separately. The Lipschitz continuity with respect to t is derived from

$$\|T_\rho(H,t_1)-T_\rho(H,t_2)\|_{2-\mathrm{col}} \leq \int_{t_1}^{t_2}\int_\beta \|g(T_\rho(H,t),\beta\|_{2-\mathrm{col}}\rho(\beta,t)d\beta dt \leq (1+C_\rho/2)KC(1+r+r^2)(t_2-t_1), \tag{C.9}$$

for any $t_1, t_2 \in [0,1]$. Given Assumption 2 (iii), we have

$$\begin{aligned}&\|T_\rho(H,t)-T_\rho(H',t)\|_F \leq \int_0^t\int_\beta \|g(T_\rho(H,s),\beta)-g(T_\rho(H',s),\beta)\|_F\rho(\beta,s)d\beta ds \\ &\leq \int_0^t\int_\beta \phi_T(N,D,B\exp(K(1+C_\rho/2+r+r^2)))(1+r+r^2)\|T_\rho(H,s)-T_\rho(H',s)\|_F\rho(\beta,s)d\beta ds\end{aligned} \tag{C.10}$$

for any $H, H' \in \mathrm{supp}(\mu)$. Define $L_H := \phi_T(N,D,B\exp(K(1+C_\rho/2+r+r^2)))(1+r+r^2)$. Utilizing Grönwall's inequality, we establish:

$$\|T_\rho(H,t)-T_\rho(H',t)\|_F \leq \|H-H'\|_F \exp(L_H\int_0^1\int_\beta \rho(\beta,s)d\beta ds) = \exp(L_H)\|H-H'\|_F$$

given that $T_\rho(\cdot,0)$ serves as the identity mapping. Consequently, $T_\rho(H,t)$ demonstrates Lipschitz continuity with respect to $H \in \mathrm{supp}(\mu)$. □

C.3 Useful technical lemmas

Lemma C.1 (Continuous Transformer output bound)**.** *Under Assumption 2, for any distribution $\rho \in \mathcal{P}(\Omega)$ where $\int_0^1\int_\beta ||\beta||_2^2\rho(\beta,t)d\beta dt \leq A^2$ for some universal constant $A > 0$ and for any $t \in [0,1]$, we have*

$$\|T_\rho(H,t)\|_{2-\mathrm{col}} \leq \|H\|_{2-\mathrm{col}}\exp(K(1+A+A^2)).$$

Lemma C.2 (Continuous Transformer difference bound)**.** *Under Assumption 2 and given H, for any $\rho, \nu \in \mathcal{P}^2$ that satisfy $\int_0^1\int_\beta ||\beta||_2^2\rho(\beta,t)d\beta dt \leq A^2$ and have bounded supports P_r for some constants $A, r > 0$, we have that*

$$\sup_{t\in[0,1]}\|T_\rho(H,t)-T_\nu(H,t)\|_F \leq C_r W_1(\rho,\nu).$$

Here, the universal constant C_r only depends on N, D, r, A, and the parameters of the assumptions.

Lemma C.3 (Continuous Transformer gradient component bound)**.** *Under Assumption 1 and 2, for any $\rho \in \mathcal{P}(\Omega)$ where $\operatorname{supp}(\rho) \subset P_r$ and $\int_0^1 \int_\beta \|\beta\|^2 \rho(\beta,t) d\beta dt \le A^2$, we have*

$$\sup_{(\beta,t)\in P_r} \|g(T_\rho(H,t),\beta)\|_F \le \sqrt{N+1} KB \exp(K(1+A+A^2))(1+r+r^2),$$

$$\sup_{(\beta,t)\in P_r} \|p_\rho(H,t)\|_F \le (B + B\exp(K(1+A+A^2))) \exp\Big(\phi_T(N, D, \sqrt{N+1}KB\exp(K(1+A+A^2))(1+A+A^2)\Big),$$

$$\sup_{(\beta,t)\in P_r} \Big|\frac{\delta Q}{\delta\rho}(\beta,t)\Big| \le \sqrt{N+1}KB\exp(K(1+A+A^2))(1+r+r^2)(B + B\exp(K(1+A+A^2)))$$
$$\exp\Big(\phi_T(N, D, \sqrt{N+1}KB\exp(K(1+A+A^2))(1+A+A^2)\Big) + \frac{\lambda}{2} r^2.$$

Lemma C.4 (Discrete Transformer bound)**.** *Under Assumptions 2, for any Θ where $\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta\|^2 \le A^2$ for some universal constant $A > 0$ and at any $t = 0, \Delta t/2, \Delta t, \ldots, (L-1/2)\Delta t, 1$, we have*

$$\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}} \le \|H\|_{2-\mathrm{col}} \exp(K(1+A+A^2)).$$

Lemma C.5 (Discrete Transformer difference bound)**.** *Under Assumption 1 and 2, for any H, let $\Theta = \{\beta_{t,j}\}_{t/\Delta t+1\in[L], j\in[M]}$, $\widetilde{\Theta} = \{\widetilde{\beta}_{t,j}\}_{t/\Delta t+1\in[L], j\in[M]}$ such that $\max\{\|\beta_{t,j}\|, \|\widetilde{\beta}_{t,j}\|\} \le r$ for any $t = 0, \ldots, (L-1)\Delta t,\ j = 1, \ldots, M$. Then, we have that*

$$\|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_F \le C_r \frac{1}{ML} d(\Theta, \widetilde{\Theta}).$$

Here the universal constant C_r only depends on N, D, r, and the parameters of the assumptions.

Lemma C.6 (Discrete Transformer gradient component bound)**.** *Under Assumption 1 and 2, for any Θ such that $\sup_{t,j}\|\beta_{t,j}\|^2 \le r^2$ and $\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta\|^2 \le A^2$, we have, for any $t = 0, \Delta t/2, \ldots, (L-1/2)\Delta t, 1$*

$$\sup_{(\theta,t)\in P_r} \max\{\|f(\widehat{T}_\Theta(H,t),\theta)\|_F, \|h(\widehat{T}_\Theta(H,t),w)\|_F, \|g(\widehat{T}_\Theta(H,t),\beta)\|_F\} \le \sqrt{N+1}KB_T(1+r+r^2),$$

$$\sup_{(\beta,t)\in P_r} \sup_{i\in[N+1]} \max\{\Big\|\nabla_\theta f(\widehat{T}_\Theta(H,t),\theta)_{:,i}\Big\|, \Big\|\nabla_w h(\widehat{T}_\Theta(H,t),w)_{:,i}\Big\|, \Big\|\nabla_\beta g(\widehat{T}_\Theta(H,t),\beta)_{:,i}\Big\|\} \le \phi_P(B_T)(1+r),$$

$$\sup_{(\beta,t)\in P_r} \|\widehat{p}_\Theta(H,t)\|_F \le (B + B_T)\exp\Big(\phi_T(N, D, \sqrt{N+1}KB_T)(1+A+A^2)\Big),$$

where $B_T = B\exp(K(1+A+A^2))$.

Lemma C.7 (Norm average concentration)**.** *Under Assumption 2, consider a parameter setting $\Theta = \{\beta_{t,j}\}_{t/\Delta t+1\in[L], j\in[M]}$ i.i.d. drawn from $\{\rho(\beta|t)\}_{t/\Delta t+1\in[L], j\in[M]}$ where $\rho \in \mathcal{P}^2$ is concentrated on P_r and satisfies $\int_\beta \rho(\beta,t) d\beta = 1$ for every $t \in [0,1]$. Then, with probability at least $1 - \exp(-\delta)$, we have*

$$\Big|\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta_{t,j}\|^2 - \int_0^1 \int_\beta \|\beta\|^2 \rho(\beta,t) d\beta dt\Big| \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}.$$

for any $\delta > 0$. Here, the $\lesssim$ notation hides the dependencies on r and the parameters specified in the assumption.

Lemma C.8 (Matrix product difference bound)**.** *Suppose that for some $d > 0$, the matrices $A_1, A_2, \ldots, A_L$ and $B_1, B_2, \ldots, B_L$ satisfy the following conditions:*

1. *For each $l = 1, \ldots, L$, the norms of the matrices are bounded as $\|A_l\| \le 1 + a_l$, $\|B_l\| \le 1 + b_l$, where $a_l, b_l > 0$.*
2. *The product of the increments for each matrix is bounded by $\prod_{l=1}^L 1 + \max\{a_l, b_l\} \le C$ for some constant $C > 0$.*

Under these conditions, it holds that

$$\Big\|\prod_{l=1}^L A_l - \prod_{l=1}^L B_l\Big\| \le C \sum_{l=1}^L \|A_l - B_l\|.$$

C.4 Solution of adjoint ODE

In this section, we define the partial derivative

$$p_\rho(H,t) := \frac{\partial R(H;\rho)}{\partial T_\rho(H,t)^T} \in \mathbb{R}^{D\times(N+1)}$$

without specifying its explicit formula. Denote the derivative of $T_\rho(H,1)$ to $T_\rho(H,t)$ (after vectorization) by the Jacobian $J_\rho(H,t) \in \mathbb{R}^{(N+1)D\times(N+1)D}$, and assume that $\dot{J}\rho(H,t)$ exists for any $t\in[0,1]$. Then [51] shows that $J_\rho(H,t)$ satisfies the adjoint equation of the ODE.

$$\dot{J}_\rho(H,t) = -J_\rho(H,t)\nabla_{\mathrm{vec}[T]}\Big\{\mathrm{vec}[\int_\beta g(T_\rho(H,t),\beta)\rho(\beta,t)d\beta]\Big\} \tag{C.11}$$

for any $t\in[0,1]$. By applying the chain rule and exchanging the order of the derivative and integral, we have, for any $t\in[0,1]$, that

$$\begin{aligned}\mathrm{vec}[p_\rho(H,t)]^T &= \frac{\partial R(H;\rho)}{\partial \mathrm{vec}[T_\rho(H,1)]}\frac{\partial \mathrm{vec}[T_\rho(H,1)]}{\partial \mathrm{vec}[T_\rho(H,t)]}\\ &= \mathrm{vec}[\frac{\partial R(H;\rho)}{\partial T_\rho(H,1)^T}]^T\frac{\partial \mathrm{vec}[T_\rho(H,1)]}{\partial \mathrm{vec}[T_\rho(H,t)]}\\ &= \mathrm{vec}[p_\rho(H,1)]^T J_\rho(H,t)\end{aligned}$$

Hence, by taking the derivative with respect to t, we obtain that

$$\mathrm{vec}[\dot{p}_\rho(H,t)]^T = -\mathrm{vec}[p_\rho(H,t)]^T\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t)d\beta$$

with the solution

$$\mathrm{vec}[p_\rho(H,t)]^T = \mathrm{vec}[p_\rho(H,1)]^T\exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta ds\Big). \tag{C.12}$$

On the other hand, we have

$$p_\rho(H,1) = \frac{\partial R(H;\rho)}{\partial T_\rho(H,1)} = (\mathrm{Read}[T_\rho(H,1)] - y(H))E_{\mathrm{read}}, \tag{C.13}$$

where E_{read} is a $D\times(N+1)$ zero matrix except 1 at the $(d+1,N+1)-$th entry. Moreover, from (C.12) and (C.13), we see that

$$\mathrm{vec}[p_\rho(H,t)]^T = (\mathrm{Read}[T_\rho(H,1)]-y(H))\cdot\exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta ds\Big)_{DN+d+1,:}. \tag{C.14}$$

Additionally, we could explicitly derive the formula for $\widehat{p}_\Theta(H,t) = \frac{\partial\widehat{R}(H;\Theta)}{\partial\widehat{T}_\Theta(H,t)}$ for the discrete Transformer. By applying the chain rule multiple times across each layer with the encoder either f or h, for any $t = 0,\Delta t,\ldots,(L-1)\Delta t, 1$, we have

$$\begin{aligned}&\mathrm{vec}[\widehat{p}_\Theta(H,t)] = (\mathrm{Read}[\widehat{T}_\Theta(H,1)] - y(H))\\ &\Big\{\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}}\Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^M\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\Big)\\ &\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}}\Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^M\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]\Big)\Big\}_{DN+d+1,:},\end{aligned} \tag{C.15}$$

and

$$\mathrm{vec}[\widehat{p}_\Theta(H,t+\Delta t/2)] = (\mathrm{Read}[\widehat{T}_\Theta(H,1)] - y(H))$$

$$\Big\{ \prod_{\substack{(s-t)/\Delta t+2\in[(1-t)/\Delta t]\\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\Big)$$

$$\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]\Big)\Big\}_{DN+d+1,:}. \tag{C.16}$$

C.5 Explanation for assumptions made in Theorem 4.1

The justification of the two assumptions outlined in Theorem 4 warrants careful consideration. While we provide only high-level justifications, they underpin significant aspects of our theoretical framework.

For the first assumption, we argue that the regularization parameter λ, which penalizes the magnitude of the parameter norms, implicitly promotes solutions that are confined to a compact subset of the parameter space. This rationale is conceptual and requires that regularization effectively constrains the growth of the parameter norms, thereby localizing the solutions.

The second assumption concerns the separation property. It is naturally satisfied as long as the origin $0_{\dim\theta+\dim w}$ remains an interior point of $\mathrm{supp}(\rho_\infty(\cdot,t))$. This condition is relatively mild and is generally satisfied. The challenge arises in verifying that $\mathrm{supp}(\rho_\infty(\cdot,t))$ for the α_2 component extends to encompass the entire space $\mathcal{K}$. While direct confirmation is elusive, it is suggested by [20] initially spans $\mathcal{K}$, this expansive support property is maintained at any finite time. Thus, we conjecture that the condition holds under these circumstances, providing a basis for this assumption.

D Proofs of main results in Section 3

This convergence is detailed in two parts. First, the finite time result, as stated in points (i)-(iii), utilizes a concept in probability theory known as *propagation of chaos* [55] to examine how differences evolve uniformly across a given time interval. In the context of our model, this involves comparing how parameter particles evolve under discrete versus continuous dynamics.

Secondly, the weak convergence of the empirical distribution process leverages optimal transport theory alongside abstract stability results for Wasserstein gradient flows [4]. This argument involves detailed analysis of the discretization of particle distributions in space, particularly focusing on obtaining the convergence of the sequence of *momentum fields* [4, 53].

D.1 Proof of Theorem 3.1

For any gradient flow parameter setting $\Theta^{(\tau)} = \{\beta_{t,j}^{(\tau)}\}_{t,j}$, from its definition (2.7), we could rewrite the dynamics as

$$\beta_{t,j}^{(\tau)} = \beta^{(0)} - \int_0^\tau \widehat{G}(\beta_{t,j}^{(s)}, \Theta^{(s)}, t)ds \tag{D.1}$$

for any gradient flow time $\tau > 0$, depth index $t = 0, \Delta t, \dots, (L-1)\Delta t$ and width index $j = 1, \dots, M$. For simplicity, any mentioned universal constant only depends on N, D, τ, λ and the parameters of the assumptions, and we abbreviate the subscript $t = 0, \Delta t, \dots, (L-1)\Delta t,\ j = 1, \dots, M$ by t, j throughout the proof.

Inspired by the "propagation of chaos" idea [55], we could define the "nonlinear dynamics" with the same initialization setting $\widetilde{\Theta}^{(\tau)} = \{\widetilde{\beta}_{t,j}^{(\tau)}\}_{t,j}$, i.e.

$$\begin{cases} \widetilde{\beta}_{t,j}^{(\tau)} = \widetilde{\beta}^{(0)} - \int_0^\tau G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)ds, \\ \widetilde{\beta}_{t,j}^{(0)} = \beta_{t,j}^{(0)} \end{cases} \tag{D.2}$$

for any t, j. Here, $(\rho^{(s)})_{s\geq 0}$ is the solution to the Wasserstein gradient flow (3.5), of which the uniqueness is implied by Proposition 3.2. Since (D.2) is just the particle flow of (3.5), its existence and uniqueness are guaranteed by Proposition 3.2.

Observing that $\{\beta_{t,j}\}_{t,j}$ are independent due to the dynamics only involving $(\rho^{(s)})_{s\geq 0}$ with i.i.d initialization over ρ_0, we can consider $\{\widetilde{\beta}_{t,j}^{(s)}\}_{t,j}$ as i.i.d. samples drawn from $\{\rho^{(s)}\}_{t,j}$. In addition, from Propositions 3.2 and D.1, for any t, j we have $\max\{\|\beta_{t,j}\|, \|\widetilde{\beta}_{t,j}\|\} \leq R_\tau$, where R_τ is defined as in these propositions and does not depend on M and L.

As the final preparatory step for the proof of Theorem 3.1, we present the following three lemmas that will be helpful:

Lemma D.1 (Continuous gradient difference bound)**.** *Suppose Assumptions 1-3 hold. If we have $\rho, \nu \in \mathcal{P}^2$ concentrated on P_r for some $r > 0$, and $\beta, \widetilde{\beta}$ such that $\max\{\|\beta\|, \|\widetilde{\beta}\|\} \leq r$, then*

$$\left\|G(\beta, \rho, t) - G(\widetilde{\beta}, \nu, t)\right\| \leq C_G\Big(\exp(C_G W_1(\rho, \nu)) - 1 + (1+\lambda)\|\beta - \widetilde{\beta}\|\Big).$$

for any $t \in [0, 1]$. Here, the universal constant C_G only depends on N, D, r, and the parameters of the assumptions.

Lemma D.2 (Discrete gradient difference bound)**.** *Under Assumption 1-3, for any $\Theta = \{\beta_{t,j}\}_{t/\Delta t+1\in[L], j\in[M]}, \widetilde{\Theta} = \{\widetilde{\beta}_{t,j}\}_{t/\Delta t+1\in[L], j\in[M]}$ such that $\max\{\|\beta_{t,j}\|, \|\widetilde{\beta}_{t,j}\|\} \leq r$ for any $t = 0, \ldots, (L-1)\Delta t,\ j = 1, \ldots, M$. Then we have that*

$$\|\widehat{G}(\beta, \Theta, t) - \widehat{G}(\widetilde{\beta}, \widetilde{\Theta}, t)\| \leq C_G\Big(\frac{1}{ML} d(\Theta, \widetilde{\Theta}) + (1+\lambda)\|\beta - \widetilde{\beta}\|\Big).$$

for any $\beta, \widetilde{\beta} \in \{\beta : \|\beta\| \leq r\}$ and $t = 0, \Delta t, \ldots, (L-1)\Delta t$. Here, the universal constant C_G only depends on N, D, r and the parameters of the assumptions and $d(\Theta, \widetilde{\Theta})$ is defined as $\sum_t \sum_{j=1}^M \|\beta_{t,j} - \widetilde{\beta}_{t,j}\|$.

Lemma D.3 (Oracle gradient approximation with discretization)**.** *Under Assumptions 1-3, suppose that the parameter setting Θ is i.i.d. drawn from $\{\rho(\beta|t)\}_{t/\Delta t+1\in[L], j\in[M]}$ for some $\rho \in \mathcal{P}^2$ concentrated on P_r and satisfies that $\int_\beta \rho(\beta, t) d\beta = 1$ for any $t \in [0, 1]$. Then with probability at least $1 - 4\exp(-\delta)$ we have*

$$\left\|\widehat{G}(\beta, \Theta, t) - G(\beta, \rho, t)\right\| \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}},$$
$$\|G(\beta, \hat{\rho}, t) - G(\beta, \rho, t)\| \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}},$$
$$\left\|\widehat{G}(\beta, \Theta, t) - G(\beta, \hat{\rho}, t)\right\| \lesssim L^{-1}$$

for any $\beta \in \{\beta : \|\beta\| \leq r\}$, $t = 0, \Delta t, \ldots, (L-1)\Delta t, 1$ and any $\delta > 0$. Here, $\lesssim$ hides the dependencies on N, D, r and the parameters of the assumptions.

Proof of Theorem 3.1. Our proof consists of several steps outlined below:

Step I: Show the W_2 continuity of parameter (sample) distributions:

Our analysis commences with the bound for $0 < s_1 < s_2 < \tau$, we have

$$W_2(\hat{\rho}^{(s_1)}, \hat{\rho}^{(s_2)})^2 \leq \frac{1}{ML}\sum_t \sum_{j=1}^M |\beta_{t,j}^{(s_1)} - \beta_{t,j}^{(s_2)}|^2 \leq \frac{(s_2 - s_1)}{ML}\sum_t \sum_{j=1}^M \int_{s_1}^{s_2} \|\widehat{G}(\beta_{t,j}^{(s)}, \Theta^{(s)}, t)\|^2 ds,$$

where each particle at time s_1 is paired with its position at time s_2, leveraging the Jensen's inequality. Recalling the identity

$$\frac{d\widehat{Q}(\Theta^{(s)})}{ds} = \frac{1}{ML}\sum_t \sum_{j=1}^M \|\widehat{G}(\beta_{t,j}^{(s)}, \Theta^{(s)}, t)\|^2$$

shown in Proposition D.1, it follows that

$$W_2(\hat{\rho}^{(s_1)}, \hat{\rho}^{(s_2)}) \leq (s_2 - s_1)^{1/2} \widehat{Q}^{1/2}(\Theta^{(0)}) \leq \frac{\lambda}{2} A_0^2,$$

where the last inequality uses (D.30). Since $A_0^2 \lesssim 1 + \lambda^{-1}$, we see that $W_2(\hat{\rho}^{(s_1)}, \hat{\rho}^{(s_2)}) \leq C(1+\lambda)(s_2 - s_1)^{1/2}$ for some universal constant C. Similarly, we have

$$W_2(\rho^{(s_1)}, \rho^{(s_2)})^2 \leq \mathbb{E}\left\|\widetilde{\beta}^{(s_2)} - \widetilde{\beta}^{(s_1)}\right\|^2 \leq (s_2 - s_1)\int_{s_1}^{s_2}\int_0^1\int_\beta \left\|G(\beta, \rho^{(s)}, t)\right\|^2 d\beta dt ds \leq (s_2 - s_1) Q(\rho_0) \lesssim (s_2 - s_1) C(1+\lambda$$

where $\widetilde{\beta}^{(s_1)} \sim \rho^{(s_1)}(\beta, t)$ is embedded with its future position at $\widetilde{\beta}^{(s_2)}$. The last step is feasible by setting C large enough, noticing that $Q(\rho_0) \leq \lambda A_0^2/2$. To summarize, we have

$$\max\left\{W_2(\hat{\rho}^{(s_1)}, \hat{\rho}^{(s_2)}), W_2(\rho^{(s_1)}, \rho^{(s_2)})\right\} \leq C(1+\lambda)\sqrt{s_2 - s_1} \tag{D.3}$$

for some universal constant $C > 0$.

Step II: Bound the difference between gradient flow dynamics and non-linear dynamics:

Next, we aim to bound $\Delta(s) := \sup_{s' \in [0,s]} \sup_{t,j} \|\beta_{t,j}^{(s')} - \widetilde{\beta}_{t,j}^{(s')}\|$ for any $s \in [0, \tau]$. Taking the difference of (D.1) and (D.2), we obtain that

$$\begin{aligned}
\|\beta_{t,j}^{(\tau)} - \widetilde{\beta}_{t,j}^{(\tau)}\| \leq & \int_0^\tau \left\|\widehat{G}(\beta_{t,j}^{(s)}, \Theta^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\| ds \\
\leq & \int_0^\tau \left\|\widehat{G}(\beta_{t,j}^{(s)}, \Theta^{(s)}, t) - \widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t)\right\| ds \\
+ & \int_0^\tau \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\| ds \\
\leq & C_G \int_0^\tau \Big(\frac{1}{ML} d(\Theta^{(s)}, \widetilde{\Theta}^{(s)}) + (1+\lambda)\|\beta^{(s)} - \widetilde{\beta}^{(s)}\|\Big) ds + \int_0^\tau \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\| ds \\
\leq & C_G \int_0^\tau \Big(\frac{1}{ML} d(\Theta^{(s)}, \widetilde{\Theta}^{(s)}) + (1+\lambda)\|\beta^{(s)} - \widetilde{\beta}^{(s)}\|\Big) ds + \sup_{s\in[0,\tau]} \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\|
\end{aligned} \tag{D.4}$$

where the final inequality stems from Lemma D.2, employing a universal constant C_G. By taking the supremacy over t, j in (D.4), and considering $\frac{1}{ML} d(\Theta^{(s)}, \widetilde{\Theta}^{(s)}) \leq \sup_{t,j} \|\beta_{t,j}^{(\tau)} - \widetilde{\beta}_{t,j}^{(\tau)}\|$ for any $s \geq 0$, we derive

$$\sup_{t,j} \|\beta_{t,j}^{(\tau)} - \widetilde{\beta}_{t,j}^{(\tau)}\| \leq C_G(2+\lambda)\int_0^\tau \sup_{t,j}\|\beta^{(s)} - \widetilde{\beta}^{(s)}\| ds + \sup_{t,j}\sup_{s\in[0,\tau]} \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\|$$

Further supremacy taken over $s \in [0, \tau]$ yields:

$$\Delta(\tau) \leq \widetilde{\Delta}(\tau) + C_G(2+\lambda)\int_0^\tau \Delta(s) ds, \tag{D.5}$$

where we define $\widetilde{\Delta}(\tau) := \sup_{t,j}\sup_{s\in[0,\tau]} \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\|$ for simplicity of notation. Apply the Grönwall's inequality to (D.5) yields

$$\Delta(\tau) \leq \exp\Big(C_G(2+\lambda)\tau\Big)\widetilde{\Delta}(\tau). \tag{D.6}$$

It remains to bound $\widetilde{\Delta}(\tau)$ to bound $\Delta(\tau)$. It's worth noting that by Lemmas D.1 and D.2, for any $s_1, s_2 \in [0, \tau]$ and t, j, we have

$$\begin{aligned}
& \left|\left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s_2)}, \widetilde{\Theta}^{(s_2)}, t) - G(\widetilde{\beta}_{t,j}^{(s_2)}, \rho^{(s_2)}, t)\right\| - \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s_1)}, \widetilde{\Theta}^{(s_1)}, t) - G(\widetilde{\beta}_{t,j}^{(s_1)}, \rho^{(s_1)}, t)\right\|\right| \\
\leq & \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s_1)}, \widetilde{\Theta}^{(s_1)}, t) - \widehat{G}(\widetilde{\beta}_{t,j}^{(s_2)}, \widetilde{\Theta}^{(s_2)}, t)\right\| + \left\|G(\widetilde{\beta}_{t,j}^{(s_1)}, \rho^{(s_1)}, t) - G(\widetilde{\beta}_{t,j}^{(s_2)}, \rho^{(s_2)}, t)\right\| \\
\leq & C_\Delta(\exp(C_\Delta W_1(\rho^{(s_1)}, \rho^{(s_2)})) - 1) + C_\Delta \frac{1}{ML} d(\widetilde{\Theta}^{(s_1)}, \widetilde{\Theta}^{(s_2)}) + C_\Delta(1+\lambda)\|\widetilde{\beta}_{t,j}^{(s_1)} - \widetilde{\beta}_{t,j}^{(s_2)}\| \\
\lesssim & \exp(C_\Delta C\sqrt{s_2 - s_1}) - 1 + (1+\lambda)\sup_{t,j}\|\widetilde{\beta}_{t,j}^{(s_1)} - \widetilde{\beta}_{t,j}^{(s_2)}\| \\
\lesssim & \sqrt{s_2 - s_1} + \sup_{t,j}\|\widetilde{\beta}_{t,j}^{(s_1)} - \widetilde{\beta}_{t,j}^{(s_2)}\|
\end{aligned} \tag{D.7}$$

for some universal constant C_Δ. Moreover, for any t, j, we have

$$\|\widetilde{\beta}_{t,j}^{(s_1)}-\widetilde{\beta}_{t,j}^{(s_2)}\| \leq \int_{s_1}^{s_2} \|G(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t)\| \leq \sqrt{s_2 - s_1} \int_{s_1}^{s_2} \|G(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t)\|^2 \lesssim (1+\lambda^2)\sqrt{s_2 - s_1}, \tag{D.8}$$

where the universal boundedness of $\|G(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - \lambda\widetilde{\beta}_{t,j}^{(s)}\|$ can be readily derived from the third result of Lemma C.3 alongside Assumption 2 (ii). Therefore, we obtain

$$\left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s_2)}, \widetilde{\Theta}^{(s_2)}, t) - G(\widetilde{\beta}_{t,j}^{(s_2)}, \rho^{(s_2)}, t)\right\| - \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s_1)}, \widetilde{\Theta}^{(s_1)}, t) - G(\widetilde{\beta}_{t,j}^{(s_1)}, \rho^{(s_1)}, t)\right\| \lesssim (1+\lambda+\lambda^2+\lambda^3)\sqrt{s_2 - s_1} \tag{D.9}$$

for any $s_1, s_2 \in [0, \tau]$.

Define $\tau_n = n\tau/L^2$ for $n = 1, 2, \ldots, L^2$. Leveraging Lemma D.3 and applying the union bound over $n \in [L^2]$ (updating the δ in the lemma to $\delta + 2\log L$), we obtain, with a probability of at least $1 - \exp(-\delta)$:

$$\begin{aligned}
\sup_{s\in\{\tau_n\}_{n\in[L^2]}} \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\| \lesssim& (1 + \lambda + \lambda^2 + \lambda^3)L^{-1} + \sqrt{\frac{\delta + \log L + \log(L + 1)}{M}} \\
\lesssim& L^{-1} + \sqrt{\frac{\delta + \log(L + 1)}{M}}
\end{aligned}$$

Furthermore, leveraging (D.9), we deduce

$$\begin{aligned}
\sup_{t,j} \sup_{s\in[0,\tau]} \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\| \lesssim& \sup_{t,j} \sup_{s\in\{\tau_n\}_{n\in[L^2]}} \left\|\widehat{G}(\widetilde{\beta}_{t,j}^{(s)}, \widetilde{\Theta}^{(s)}, t) - G(\widetilde{\beta}_{t,j}^{(s)}, \rho^{(s)}, t)\right\| \\
&+ \sup_{|s_1-s_2|\leq\tau/L^2} (1 + \lambda + \lambda^2 + \lambda^3)\sqrt{s_2 - s_1} \\
\lesssim& L^{-1} + \sqrt{\frac{\delta + \log(L + 1)}{M}}.
\end{aligned}$$

Returning to (D.6), we establish that with a probability of at least $1 - \exp(-\delta)$

$$\sup_{s\in[0,\tau]} \sup_{t,j} \|\beta_{t,j}^{(s)} - \widetilde{\beta}_{t,j}^{(s)}\| = \Delta(\tau) \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L + 1)}{M}}.$$

We denote the event where the above inequality holds as E_1, thus we have $\mathbb{P}(E_1) \geq 1 - \exp(-\delta)$.

Step III: Prove the finite time results:

Now, we are poised to demonstrate the results in Theorem 3.1 that concern supremacy over $s \in [0, \tau]$. The verification of Lemmas C.4 reveals the existence of a universal constant $B_\tau := B\exp(K(1 + R_\tau + R_\tau^2))$ such that

$$\max\{\|T_{\rho^{(\tau)}}(H, t)\|_{2-\mathrm{col}}, \|\widehat{T}_{\Theta^{(\tau)}}(H, t)\|_{2-\mathrm{col}}, \|\widehat{T}_{\widetilde{\Theta}^{(\tau)}}(H, t)\|_{2-\mathrm{col}}\} \leq B_\tau$$

for any H and $t \in [0, 1]$.

Utilizing Lemma D.6 and applying the union bound over $n \in [L^2]$, we observe

$$\sup_{s\in\{\tau_n\}_{n\in[L^2]}} \|\widehat{T}_{\widetilde{\Theta}^{(s)}}(H, t) - T_{\rho^{(s)}}(H, t)\|_F \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L + 1)}{M}}.$$

Additionally, note that for any $s_1, s_2 \in [0, \tau]$, we derive from Lemmas C.2 and C.5 that

$$\begin{aligned}
&\left|\|\widehat{T}_{\widetilde{\Theta}^{(s_1)}}(H, t) - T_{\rho^{(s_1)}}(H, t)\|_F - \|\widehat{T}_{\widetilde{\Theta}^{(s_2)}}(H, t) - T_{\rho^{(s_2)}}(H, t)\|_F\right| \\
\leq& \|\widehat{T}_{\widetilde{\Theta}^{(s_1)}}(H, t) - \widehat{T}_{\widetilde{\Theta}^{(s_2)}}(H, t)\|_F + \|T_{\rho^{(s_1)}}(H, t) - T_{\rho^{(s_2)}}(H, t)\|_F \\
\lesssim& \sup_{t,j} \|\widetilde{\beta}_{t,j}^{(s_1)} - \widetilde{\beta}_{t,j}^{(s_2)}\| + W_1(\rho^{(s_1)}, \rho^{(s_2)}) \\
\lesssim& \sqrt{s_2 - s_1}
\end{aligned} \tag{D.10}$$

where the last inequality is derived utilizing (D.3) and (D.8). Consequently, we have

$$
\begin{aligned}
\sup_{s\in[0,\tau]}\|\widehat{T}_{\widetilde{\Theta}^{(s)}}(H,t)-T_{\rho^{(s)}}(H,t)\|_F &\lesssim \sup_{s\in\{\tau_n\}_{n\in[L^2]}}\|\widehat{T}_{\widetilde{\Theta}^{(s)}}(H,t)-T_{\rho^{(s)}}(H,t)\|_F+\sup_{|s_2-s_1|\le\tau/L^2}\sqrt{s_2-s_1}\\
&\lesssim L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}}
\end{aligned}
\tag{D.11}
$$

with probability at by $1-\exp(-\delta)$. We denote the event where the above inequality holds as E_2, thus we have $\mathbb{P}(E_2)\ge 1-\exp(-\delta)$. Considering that $\{\widetilde{\beta}_{t,j}^{(s)}\}_{t,j}$ could be regarded as i.i.d. samples drawn from $\{\rho^{(s)}\}_{t,j}$, employing a similar method with the concentration guarantee from Lemma C.7, we can readily deduce the existence of an event E_3 with $\mathbb{P}(E_3)\ge 1-\exp(-\delta)$ such that, under E_3, we have

$$
\sup_{s\in[0,\tau]}\Big|\frac{1}{ML}\sum_t\sum_{j=1}^M\|\widetilde{\beta}_{t,j}^{(s)}\|^2-\int_0^1\int_\beta\|\beta\|^2\rho^{(s)}(\beta,t)d\beta dt\Big|\lesssim L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}}.
$$

Now, let's analyze the scenario under the probability event $E_1\cap E_2\cap E_3$ with $\mathbb{P}(E_1\cap E_2\cap E_3)\ge 1-3\exp(-\delta)$. Lemma C.5 demonstrates that

$$
\begin{aligned}
&\sup_{s\in[0,\tau]}|\mathrm{Read}[\widehat{T}_{\Theta^{(s)}}(H,t)]-\mathrm{Read}[\widehat{T}_{\widetilde{\Theta}^{(s)}}(H,t)]|\le\sup_{s\in[0,\tau]}\|\widehat{T}_{\Theta^{(s)}}(H,t)-\widehat{T}_{\widetilde{\Theta}^{(s)}}(H,t)\|_F\\
\lesssim&\sup_{s\in[0,\tau]}\sup_{t,j}\|\beta_{t,j}^{(s)}-\widetilde{\beta}_{t,j}^{(s)}\|\lesssim L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}}.
\end{aligned}
\tag{D.12}
$$

This further implies, by (D.11), that

$$
\begin{aligned}
&\sup_{s\in[0,\tau]}|\mathrm{Read}[\widehat{T}_{\Theta^{(s)}}(H,t)]-\mathrm{Read}[T_{\rho^{(s)}}(H,t)]|\\
\le&\sup_{s\in[0,\tau]}|\mathrm{Read}[\widehat{T}_{\Theta^{(s)}}(H,t)]-\mathrm{Read}[\widehat{T}_{\widetilde{\Theta}^{(s)}}(H,t)]|+\sup_{s\in[0,\tau]}\|\widehat{T}_{\widetilde{\Theta}^{(s)}}(H,t)-T_{\rho^{(s)}}(H,t)\|_F\\
\lesssim&L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}},
\end{aligned}
\tag{D.13}
$$

Since $\max\{\|T_{\rho^{(\tau)}}(H,t)\|_{2-\mathrm{col}},\|\widehat{T}_{\Theta^{(\tau)}}(H,t)\|_{2-\mathrm{col}},\|\widehat{T}_{\widetilde{\Theta}^{(\tau)}}(H,t)\|_{2-\mathrm{col}}\}$ is universally bounded, (D.13) immediately indicates that

$$
\sup_{s\in[0,\tau]}|\widehat{R}(\Theta^{(s)})-R(\rho^{(s)})|\lesssim\sup_{s\in[0,\tau]}|\mathrm{Read}[\widehat{T}_{\Theta^{(s)}}(H,t)]-\mathrm{Read}[T_{\rho^{(s)}}(H,t)]|\lesssim L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}},
\tag{D.14}
$$

and

$$
\begin{aligned}
\sup_{s\in[0,\tau]}|\widehat{Q}(\Theta^{(s)})-Q(\rho^{(s)})|\le&\sup_{s\in[0,\tau]}|\widehat{R}(\Theta^{(s)})-R(\rho^{(s)})|+\sup_{s\in[0,\tau]}\Big|\frac{1}{ML}\sum_t\sum_{j=1}^M\|\widetilde{\beta}_{t,j}^{(s)}\|^2-\int_0^1\int_\beta\|\beta\|^2\rho^{(s)}(\beta,t)d\beta dt\Big|\\
\lesssim&L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}}.
\end{aligned}
\tag{D.15}
$$

Step IV: Prove the weakly convergence:

For the remainder of the proof, we adopt a similar approach as in the proof of Theorem 2.6 in [15]. We denote $\hat{\rho}$ as $\hat{\rho}_{M,L}$ for any given M and L. It's essential to note that we treat $\hat{\rho}_{M,L}$ as probability measures in this step of the proof.

Recalling (D.3), for any $s_1,s_2\in[0,\tau]$, we have

$$
W_2(\hat{\rho}_{M,L}^{(s_1)},\hat{\rho}_{M,L}^{(s_2)})\le C(1+\lambda)\sqrt{s_2-s_1}
$$

for some universal constant C. We observe that the family of curves $(s\mapsto\hat{\rho}_{M,L}^{(s)})_{M,L}$ is equicontinuous in W_2 on $[0,\tau]$, uniformly in M,L. Additionally, the family $(\hat{\rho}_{M,L})_{M,L}$ lies within a W_2

ball, thus weakly precompact. As the weak topology is weaker than the topology induced by W_2, according to the Arzelà–Ascoli theorem, along any sequence where $L \to \infty$ and $\log L/M \to \infty$, we can identify a subsequence that converges weakly to a certain process $(\nu^{(s)})_{s \geq 0} \in \mathcal{P}^2 \times \mathbb{R}$, concentrated on P_{R_τ} at all times. In the subsequent analysis, we solely focus on this subsequence, still denoted as $(\hat{\rho}_{M,L})_{M,L}$.

For any $t \in [0,1]$, let's define the sequence $(E^t_{M,L})_{M,L}$ of momentum fields, which is a vector-valued measure on $[0,\tau] \times \Omega$, denoted by $E_{M,L} := \widehat{G}(\beta, \Theta^{(s)}_{M,L}, t)\hat{\rho}^{(s)}(\beta,t)ds$. We also define $E := G(\beta, \nu^{(s)}, t)\nu^{(s)}(\beta,t)ds$.

Considering that both $\hat{\rho}_{M,L}$ and ν are concentrated on P_{R_τ}, we also have uniform convergence in the Bounded Lipschitz metric. Hence, for any bounded and Lipschitz function $\varphi : [0,\tau] \times \mathbb{R}^{\dim\beta} \to \mathbb{R}^{\dim\beta}$, it holds

$$\|\hat{\rho}^{(s)} - \nu^s\|_{\mathrm{BL}} \to 0$$

uniformly among $s \in [0,\tau]$ along the sequence.

Note that

$$\begin{aligned}
\Big|\int_0^\tau \int_0^1 \int_\beta \varphi \cdot d(E_{M,L} - E)\Big| \leq & \|\varphi\|_{\max} \int_0^\tau \int_0^1 \int_\beta \left\|\widehat{G}(\beta, \Theta^{(s)}_{M,L}, t) - G(\beta, \nu^{(s)}, t)\right\| \hat{\rho}^{(s)}(\beta,t) d\beta dt ds \\
& + \Big|\int_0^\tau \int_0^1 \int_\beta \varphi \cdot (\hat{\rho}^{(s)} - \nu^s)(\beta,t) d\beta dt ds\Big| \\
\lesssim & \|\varphi\|_{\max} \int_0^\tau \int_0^1 \int_\beta \left\|\widehat{G}(\beta, \Theta^{(s)}_{M,L}, t) - G(\beta, \hat{\rho}^{(s)}_{M,L}, t)\right\| \hat{\rho}^{(s)}(\beta,t) d\beta dt ds \\
& + \|\varphi\|_{\max} \int_0^\tau \int_0^1 \int_\beta \left\|G(\beta, \hat{\rho}^{(s)}, t) - G(\beta, \nu^{(s)}, t)\right\| \hat{\rho}^{(s)}(\beta,t) d\beta dt ds \\
& + \sup_{s\in[0,\tau]} \|\hat{\rho}^{(s)} - \nu^s\|_{\mathrm{BL}} \\
\lesssim & L^{-1} + \sup_{s\in[0,\tau]} \|\hat{\rho}^{(s)} - \nu^s\|_{\mathrm{BL}} + \sup_{s\in[0,\tau], \|\beta\| \leq R_\tau} \left\|G(\beta, \hat{\rho}^{(s)}, t) - G(\beta, \nu^{(s)}, t)\right\| \\
\lesssim & L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}} + \sup_{s\in[0,\tau]} \|\hat{\rho}^{(s)} - \nu^s\|_{\mathrm{BL}}
\end{aligned} \tag{D.16}$$

for some universal constant C_E and with probability at least $1 - \exp(-\delta)$ for any $\delta > 0$. Here, the third inequality of (D.16) utilizes the third result of Lemma D.3, and the fourth inequality uses the second result of Lemma D.3, following a similar process in Step II to achieve supremacy over $s \in [0,\tau]$.

From (D.16), we infer that $\left|\int_0^\tau \int_0^1 \int_\beta \varphi \cdot d(E_{M,L} - E)\right| \to 0$ almost surely along the sequence. Hence, $E_{M,L}$ converges weakly to E almost surely along the sequence, and the particle gradient flow for $(\nu^{(\tau)})_{\tau \geq 0}$ almost surely satisfies (2.7) on $[0,\tau]$ for any arbitrarily given $\tau > 0$. According to the Fokker-Planck equation without noise involved [52], we conclude that $(\nu^{(\tau)})_{\tau \geq 0}$ almost surely satisfies (3.5). Consequently, the uniqueness stated in Proposition 3.2 ensures that $(\nu^{(\tau)})_{\tau \geq 0} = (\rho^{(\tau)})_{\tau \geq 0}$ almost surely. □

D.2 Proof of Proposition 3.1

Proof. Suppose that the Fréchet derivative $\frac{\delta R}{\delta \rho}$ indeed exists, we establish

$$\frac{\delta Q}{\delta \rho}(\theta, w, t) = \frac{\delta R}{\delta \rho}(\theta, w, t) + \frac{\lambda}{2}(\|\theta\|_2^2 + \|w\|_2^2).$$

Therefore, it suffices to show that the Fréchet derivative of L with respect to ρ is

$$\frac{\delta R}{\delta \rho}(\beta, t) = \mathbb{E}_\mu \Big[\mathrm{Tr}\Big(g(T_\rho(H,t),\beta)^T p_\rho(H,t)\Big)\Big]. \tag{D.17}$$

Denote $\rho_\eta = \rho + \eta(\nu - \rho)$. We provide the following lemma to bound $T_{\rho_\eta}(H,1) - T_\rho(H,1)$ by expanding the first-order derivative as follows

Lemma D.4 (First-order derivative of Transformer output)**.** *Under Assumption 2, for any H and $\rho, \nu \in \mathcal{P}^2$ that have bounded supports, we have*

$$
\begin{aligned}
\mathrm{vec}[T_{\rho_\eta}(H,1) - T_\rho(H,1)] =& \eta \int_0^1 \int_\beta \exp\Big(\int_t^1 \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta\Big) \\
&\cdot \mathrm{vec}[g(T_\rho(H,t),\beta)](\nu-\rho)(\beta,t)d\beta dt + o(\eta),
\end{aligned}
\tag{D.18}
$$

where $\rho_\eta := \rho + \eta(\nu - \rho)$.

Given (D.18), we observe from the solution of p_ρ (C.13) and (C.14) that

$$
\begin{aligned}
&\Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big)\mathrm{Read}[T_{\rho_\eta}(H,1) - T_\rho(H,1)] \\
=& \mathrm{vec}[p_\rho(H,1)]^T \mathrm{vec}[T_{\rho_\eta}(H,1) - T_\rho(H,1)] \\
=& \eta \int_0^1 \int_\beta \mathrm{vec}[p_\rho(H,t)]^T \mathrm{vec}[g(T_\rho(H,t),\beta)](\rho-\nu)(\beta,t)d\beta dt + o(\eta) \\
=& \eta \int_0^1 \int_\beta \mathrm{Tr}\Big(g(T_\rho(H,t),\beta)^T p_\rho(H,t)\Big)(\rho-\nu)(\beta,t)d\beta dt + o(\eta),
\end{aligned}
\tag{D.19}
$$

Hence, by applying (D.19) to the risk function, we obtain

$$
\begin{aligned}
R(\rho_\eta) - R(\rho) &= \frac{1}{2}\mathbb{E}_\mu\Big[\Big(\mathrm{Read}[T_{\rho_\eta}(H,1)] - y(H)\Big)^2 - \Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big)^2\Big] \\
&= \mathbb{E}_\mu\Big[\Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big)\mathrm{Read}[T_\rho(H,1) - T_{\rho_\eta}(H,1)]\Big] \\
&+ \mathrm{Read}[T_\rho(H,1) - T_{\rho_\eta}(H,1)]O(\mathrm{Read}[T_{\rho_\eta}(H,1) - T_\rho(H,1)]) \\
&= \eta\Big\langle \frac{\delta R}{\delta \rho}, \nu - \rho\Big\rangle + o(\eta),
\end{aligned}
$$

which indicates (D.17) and concludes the proof. □

D.3 Proof of well-posedness of Wasserstein gradient flow

Proof. Following a similar idea as Proposition 2.5 of [15], we leverage the general theory of Wasserstein gradient flow developed in [4]. Define the functional family $Q_r(\rho)$ as

$$
Q_r(\rho) = \begin{cases} Q(\rho) & \rho(P_r) = 1, \\ \infty & \text{otherwise.} \end{cases}
$$

For any $r > 0$, let's consider any *admissible transport* $\gamma \in \mathcal{P}^{\Omega\times\Omega}$ concentrated on P_r. By definition, both of its marginals, denoted by ρ_1 and ρ_2, are concentrated on P_r. We define the transport cost for γ as

$$
C_p(\gamma) := (\int |x-y|^p d\gamma(x,y))^{1/p}
$$

for $p \geq 1$. Additionally, we denote the transport interpolation as $\rho_\alpha^\gamma := ((1-\alpha)\rho_1 + \alpha\rho_2)_{\#\gamma}$. Our proof consists of several steps outlined below.

Step I: Show that Q_r is proper and continuous for W_2 on its closed domain:

Note that the parameters r, D, N, λ remain fixed throughout this proof step, so we hide the constant dependencies on them. For any $(\beta, t) \in P_r$, we have $Q_r(\delta_{(\beta,t)}) = \frac{1}{2}\mathbb{E}_\mu[(\mathrm{Read}[H + \Delta t g(H,\beta)] - y(H))^2] + \frac{\lambda}{2}\|\beta\|^2 < \infty$. This indicates that Q_r is proper. Moreover, for any $\rho, \nu \in \mathcal{P}^2$ whose bounded support belong to P_r, Lemma C.1 ensures that

$$
\|T_\rho(H,t) + T_\nu(H,t)\|_F = O(1),
$$

and Lemma C.2 guarantees that

$$
\|T_\rho(H,t) - T_\nu(H,t)\|_F = O(W_1(\rho,\nu)) = O(W_2(\rho,\nu))
$$

for any H. Therefore, we have

$$
\begin{aligned}
R(\nu) - R(\rho) &= \frac{1}{2}\mathbb{E}_\mu\Big[\Big(\mathrm{Read}[T_\nu(H,1)] - y(H)\Big)^2 - \Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big)^2\Big] \\
&= \mathbb{E}_\mu\Big[\Big(\mathrm{Read}[T_\rho(H,1)] - y(H)\Big)\mathrm{Read}[T_\rho(H,1) - T_\nu(H,1)]\Big] \\
&+ \mathrm{Read}[T_\rho(H,1) - T_\nu(H,1)]O(\mathrm{Read}[T_\nu(H,1) - T_\rho(H,1)]) \\
&= O(W_2(\rho,\nu)).
\end{aligned}
\tag{D.20}
$$

Furthermore, since both ρ and ν have bounded support, $\|\beta\|^2$ is Lipschitz continuous with respect to (β, t). Therefore, by the Kantorovich-Rubinstein Theorem (see Theorem 5.10 of [61], for example), we have

$$
\Big|\frac{\lambda}{2}\int_\beta \|\beta\|^2(\rho-\nu)d\beta dt\Big| = O(W_1(\rho,\nu)) = O(W_2(\rho,\nu)). \tag{D.21}
$$

Combining (D.20) and (D.21), we obtain that $Q(\rho) - Q(\nu) = O(W_2(\rho,\nu))$. Therefore, Q_r is continuous for W_2 on its closed domain.

Step II: Show that $\alpha \mapsto Q(\rho_\alpha^\gamma)/C_2^2(\gamma)$ is differentiable and has a Lipschitz continuous derivative

Let's denote $h(\alpha) := Q_r(\rho_\alpha^\gamma)$. Lemma C.3 ensures that for any $\rho \in \mathcal{P}^2$ with bounded support belonging to P_r, we have bounded $\|\frac{\delta Q}{\delta \rho}(\cdot,\cdot)\|_F$ on P_r. Therefore, $Q_r(\rho_\alpha^\gamma)$ is differentiable with respect to t, and the derivative reads

$$
\begin{aligned}
h'(\alpha) =& \langle \frac{\delta Q}{\delta\rho}\Big|_{\rho=\rho_\alpha^\gamma}, \frac{d}{d\alpha}\rho_\alpha^\gamma\rangle \\
=& \int d\frac{\delta Q}{\delta\rho}\Big|_{\rho=\rho_\alpha^\gamma}\Big((1-\alpha)(\beta_1,t_1)+\alpha(\beta_2,t_2)\Big]\Big[(\beta_1,t_1)-(\beta_2,t_2)\Big)d\gamma((\beta_1,t_1),(\beta_2,t_2))\Big\}.
\end{aligned}
\tag{D.22}
$$

Then, it suffices to show that $h'(\alpha)$ is Lipschitz continuous. To accomplish this, we first propose the following lemma for later use:

Lemma D.5 (Locally Lipschitz of ρ for the gradient)**.** *Under Assumptions 1 and 2, for any $\rho, \nu \in \mathcal{P}^2$ concentrated on P_r, there exists some constant L_r depending on r, N, D and parameters of the assumptions such that*

$$
\begin{aligned}
&\sup_{t\in[0,1]} \|p_\rho(H,t) - p_\nu(H,t)\|_F \le L_r\|\rho-\nu\|_1, \\
&\sup_{(\beta,t)\in P_r}\Big|\frac{\delta Q}{\delta\rho}\Big|_\rho(\beta,t) - \frac{\delta Q}{\delta\rho}\Big|_\nu(\beta,t)\Big| \le L_r\|\rho-\nu\|_1.
\end{aligned}
$$

Returning to the lemma proof, for $\alpha_1, \alpha_2 \in [0,1]$, by the triangle inequality we have $|h'(\alpha_1) - h'(\alpha_2)| \le J_1 + J_2$ where

$$
\begin{aligned}
J_1 :=& \Big|\int d\frac{\delta Q}{\delta\rho}\Big|_{\rho=\rho_{\alpha_1}^\gamma}\Big((1-\alpha_1)(\beta_1,t_1)+\alpha_1(\beta_2,t_2)\Big)\Big[(\beta_1,t_1)-(\beta_2,t_2)\Big]d\gamma((\beta_1,t_1),(\beta_2,t_2)) \\
&- \int d\frac{\delta Q}{\delta\rho}\Big|_{\rho=\rho_{\alpha_2}^\gamma}\Big((1-\alpha_1)(\beta_1,t_1)+\alpha_1(\beta_2,t_2)\Big)\Big[(\beta_1,t_1)-(\beta_2,t_2)\Big]d\gamma((\beta_1,t_1),(\beta_2,t_2))\Big| \\
\le& \sup_{(\beta,t)\in P_r}\Big|\frac{\delta Q}{\delta\rho}\Big|_\rho(\beta,t) - \frac{\delta Q}{\delta\rho}\Big|_\nu(\beta,t)\Big|\int \|(\beta_1,t_1)-(\beta_2,t_2)\|_1 d\gamma((\beta_1,t_1),(\beta_2,t_2)) \\
\le& L_r\|\rho_{\alpha_1}^\gamma - \rho_{\alpha_2}^\gamma\|_1 \cdot C_1(\gamma) \\
\le& L_r C_1^2(\gamma)|\alpha_1-\alpha_2| \\
\le& L_r C_2^2(\gamma)|\alpha_1-\alpha_2|
\end{aligned}
\tag{D.23}
$$

by Lemma D.5. The final inequality of (D.23) applies Hölder's inequality to obtain that $C_1^2(\gamma) \leq C_2^2(\gamma)$. Furthermore,

$$\begin{aligned}
J_2 :=& \Big| \int d\frac{\delta Q}{\delta \rho}\Big|_{\rho=\rho^{\gamma}_{\alpha_2}} \Big\{\Big[(1-\alpha_1)(\beta_1,t_1)+\alpha_1(\beta_2,t_2)\Big]\Big[(\beta_1,t_1)-(\beta_2,t_2)\Big] d\gamma((\beta_1,t_1),(\beta_2,t_2))\Big\} \\
&- \int d\frac{\delta Q}{\delta \rho}\Big|_{\rho=\rho^{\gamma}_{\alpha_2}} \Big\{\Big[(1-\alpha_2)(\beta_1,t_1)+\alpha_2(\beta_2,t_2)\Big]\Big[(\beta_1,t_1)-(\beta_2,t_2)\Big] d\gamma((\beta_1,t_1),(\beta_2,t_2))\Big\}\Big| \\
\leq& \sup_{(\beta,t)\in P_r} \left\| \frac{\delta Q}{\delta \rho}\Big|_{\rho=\rho^{\gamma}_{\alpha_2}}(\beta,t)\right\| \Big|\alpha_1-\alpha_2\Big| \int \|(\beta_1,t_1)-(\beta_2,t_2)\|^2 d\gamma((\beta_1,t_1),(\beta_2,t_2)) \\
\leq& L_r' C_2^2(\gamma)|\alpha_1-\alpha_2|
\end{aligned} \tag{D.24}$$

where $L_r' := \sup_{(\beta,t)\in P_r} \|\frac{\delta Q}{\delta \rho}|_{\rho=\rho^{\gamma}_{\alpha_2}}(\beta,t)\| < \infty$ from Lemma C.3. Combining (D.23) and (D.24) leads us to the result that $h'(\alpha)/C_2^2(\gamma)$ is Lipschitz continuous.

Step III: Show the well-posedness of Wasserstein gradient flow at some finite time

We follow a similar approach to the proof of Proposition 2.5 in [15]. Since $h'(\alpha)$ is $\lambda_h \times C_2^2(\gamma)$-Lipschitz continuous with respect to α for some λ_h, the well-posedness of the Wasserstein gradient flow for Q_r with the velocity field constrained on P_r is a corollary of Theorem 11.2.2 of [4]. Specifically, there exists a unique curve $(\rho_r^{(\tau)})_{\tau\geq 0}$ continuous in $\mathcal{P}^2$ such that:

$$\frac{d\rho_r^{(\tau)}}{d\tau} = \mathrm{div}_\beta(\rho_r^{(\tau)} v_r^{(\tau)})$$

where

$$v_r^{(\tau)}(\beta,t) = \begin{cases} G(\beta,\rho_r^{(\tau)},t), & (\beta,t)\in P_r, \\ 0, & \text{otherwise.} \end{cases}$$

for $\rho_r^{(\tau)}$-a.e. Given the initialization ρ_0 concentrated on P_R, for any $r > R$, the unique $\rho_r^{(\tau)}$ exhibits a first exit time denoted as

$$\tau_r := \inf\{\tau > 0 : \rho_r^{(\tau)}(P_r) < 1\}.$$

By defining this exit time, for any $\bar{r} > r$ and $\tau \in [0,\tau_r]$, we observe $v_r(\tau)(\beta,t) = G(\beta,\rho_r^{(\tau)},t)$ and $v_{\bar{r}}(\tau)(\beta,t) = G(\beta,\rho_{\bar{r}}^{(\tau)},t)$. Due to uniqueness, we infer $\rho_r^{(\tau)} = \rho_{\bar{r}}^{(\tau)}$ on $\tau \in [0,\tau_r]$. Considering $\rho_r^{(\tau)}$ as the solution to (3.5), we establish the existence and uniqueness of the Wasserstein gradient flow for Q over $[0,\tau_r]$.

Step IV: Show the well-posedness of Wasserstein gradient flow at all time

To establish the Wasserstein gradient flow's definition for $\tau \geq 0$, it's necessary to demonstrate that $\lim_{r\to\infty}\tau_r = \infty$. For any $r > R$, according to the energy identity in Theorem 11.2.1 of [4], on $[0,t_r]$, we observe that $\tau \mapsto Q(\rho^{(\tau)})$ is non-increasing. Specifically, this represents

$$\begin{aligned}
\frac{dQ(\rho^{(\tau)})}{d\tau} &= \int_0^1 \int_\beta \Big\langle \frac{dQ}{d\rho}\Big|_{\rho=\rho^{(\tau)}}, \mathrm{div}_\beta(\rho^{(\tau)} G(\beta,\rho^{(\tau)},t))\Big\rangle \\
&= \int_0^1 \int_\beta \Big\langle G(\beta,\rho^{(\tau)},t), \mathrm{div}_\beta(\rho^{(\tau)} G(\beta,\rho^{(\tau)},t))\Big\rangle d\beta dt \\
&= \int_0^1 \int_\beta \rho^{(\tau)} \|G(\beta,\rho^{(\tau)},t)\|_2^2 d\beta dt \leq 0.
\end{aligned} \tag{D.25}$$

Therefore, for any $\tau \in [0, \tau_r]$, utilizing Lemma C.1, we have

$$\begin{aligned}
Q(\rho^{(\tau)}) \le Q(\rho_0) =& \mathbb{E}_\mu\Big[\frac{1}{2}\Big(\mathrm{Read}[T_{\rho_0}(H,1)] - y(H)\Big)^2\Big] + \frac{\lambda}{2}\int_0^1 \int_\beta \|\beta\|^2 \rho_0(\beta,t) d\beta dt \\
\le& \frac{1}{2}\mathbb{E}_\mu[(\|T_{\rho_0}(H,1)\|_{2-\mathrm{col}} + B)^2] + \frac{\lambda R^2}{2} \\
\le& \mathbb{E}_\mu[(\|T_{\rho_0}(H,1)\|_{2-\mathrm{col}}^2 + B^2] + \frac{\lambda R^2}{2} \\
\le& B^2 + B^2 \exp\Big(K(1+R+R^2)\Big)^2 + \frac{\lambda R^2}{2}
\end{aligned} \tag{D.26}$$

Thus, we have $\int_0^1 \int_\beta \|\beta\|^2 \rho^{(\tau)}(\beta,t) \le \frac{2}{\lambda} Q(\rho^{(\tau)}) \le R^2 + \lambda^{-1}\Big(2B^2 + 2B^2\exp\Big(K(1+R+R^2)\Big)^2\Big) = A_0^2$. According to Assumption (ii) and Lemma C.3, for any $(\beta, t) \in P_r$, we have

$$\begin{aligned}
\|v_r^{(\tau)}(\beta,t) - \lambda\beta\| = \|G(\beta,\rho^{(\tau)},t) - \lambda\beta\| \le& \sum_{i=1}^{N+1} \Big\|\nabla_\beta\Big\{g(T_\rho(H,t),\beta)_{:,i}\Big\}p_\rho(H,t)_{:,i}\Big\| \\
\le& \sup_{i\in[N+1]} \|\nabla_\beta g(T_\rho(H,t),\beta)_{:,i}\| \sum_{i=1}^{N+1}\|p_\rho(H,t)_{:,i}\| \\
\le& \sqrt{N+1}\sup_{i\in[N+1]} \|\nabla_\beta g(T_\rho(H,t),\beta)_{:,i}\| \, \|p_\rho(H,t)\|_F \\
\le& \sqrt{N+1}\phi_P(\|T_\rho(H,t)\|_{2-\mathrm{col}})\|p_\rho(H,t)\|_F(1+\|\beta\|) \\
\le& C_R(1+\|\beta\|),
\end{aligned} \tag{D.27}$$

where $C_R := \sqrt{N+1}\phi_P(B\exp(K(1+A_0+A_0^2)))(B + B\exp(K(1+A_0+A_0^2)))\exp\Big(\phi_T(N,D,\sqrt{N+1}KB\exp(K(1+A_0+A_0^2))(1+A_0+A_0^2)\Big)$. Applying (D.27) to the gradient flow equation

$$\frac{d\beta^{(\tau)}}{d\tau} = -v_r^{(\tau)}(\beta,t), \quad \beta^{(0)} = \beta$$

for $\tau \ge 0$, we obtain $\frac{d\|\beta^{(\tau)}\|}{d\tau} = \frac{\langle -v_r^{(\tau)}(\beta,t), \|\beta^{(\tau)}\|\rangle}{\|\beta^{(\tau)}\|} \le \|v_r^{(\tau)}(\beta,t) - \lambda\beta\| \le C_R(1+\|\beta^{(\tau)}\|)$. This indicates

$$\|\beta^{(\tau)}\| \le (\|\beta\|+1)\exp(C_R\tau) - 1 \le (R+1)\exp(C_R\tau) - 1 \tag{D.28}$$

by the Grönwall's inequality. Therefore, for any $T > 0$, $\rho^{(T)}$ is concentrated on $P_{(R+1)\exp(C_R T)-1}$, implying that for $r > (R+1)\exp(C_R T)$, we have $\tau_r > T$. Hence, we conclude $\lim_{r\to\infty}\tau_r = \infty$, establishing the existence of a unique Wasserstein gradient flow from (3.5) over $\tau > 0$.

Eventually, we establish the three properties listed in Proposition 3.2 for $\rho^{(\tau)}$. By (3.5), for any $t \in [0,1]$, we have

$$\begin{aligned}
\int_\beta \rho^{(\tau)}(\beta,t)d\beta =& \int_\beta \rho^{(0)}(\beta,t)d\beta + \int_0^\tau \Big(\int_\beta \mathrm{div}_\beta(\rho^{(s)}G^{(s)}(\beta,\rho^{(s)},t))d\beta\Big)ds \\
=& 1 + \int_0^\tau 0\cdot ds = 1
\end{aligned}$$

indicated by the Divergence Theorem as $\rho^{(s)}$ has bounded support. Next, for any $\tau \ge 0$, (D.28) shows that $\rho^{(\tau)}$ is concentrated on P_{R_τ}. Moreover, (D.25) now holds for any $\tau > 0$, implying $\int_0^1 \int_\beta \|\beta\|^2 \rho^{(\tau)}(\beta,t) \le A_0^2$ for any $\tau \ge 0$. □

D.4 Proof of well-posedness of gradient flow

Proposition D.1 (Existence and uniqueness of gradient flow)**.** *Under Assumptions 1-3, for any initialization of* $\Theta^{(0)}$ *i.i.d. drawn from* $\{\rho_0(\theta, w|t)\}_{t,j}$*, there exists a unique solution* $(\Theta^{(\tau)})_{\tau\ge 0}$ *for* (2.7)*. Additionally, for any* $\tau \ge 0$*, we have* *i.* $\Theta^{(\tau)}$ *has a bounded support, meaning* $\sup_{t,j}(\|\theta_{t,j}^{(\tau)}\|_2^2 + \|w_{t,j}^{(\tau)}\|_2^2) \leq R_\tau$.

ii. $\frac{1}{ML}\sum_t \sum_{j=1}^M (\|\theta_{t,j}^{(\tau)}\|_2^2 + \|w_{t,j}^{(\tau)}\|_2^2) \leq A_0^2$.

Here, R_τ *and* A_0 *are defined as in Proposition 3.2.*

Proof. The local Lipschitz continuity established in Lemma D.2 directly implies the continuity of $\widehat{G}_\beta(\beta_{t,j}, \Theta, t)$ with respect to Θ. Since $\{ML \cdot \widehat{G}_\beta(\beta_{t,j}, \Theta, t)\}_{t/\Delta t+1\in[L], j\in[M]}$ serves as the gradient of $\widehat{Q}(\Theta)$, it follows that $\widehat{Q}(\Theta)$ is continuously differentiable, indicating the local semiconvexity of $\widehat{Q}(\Theta)$. Specifically, for any Θ, there exists some $\kappa > 0$ such that $\widehat{Q}(\Theta) + \kappa \sum_t \sum_{j=1}^M (\|\theta_{t,j}\|_2^2 + \|w_{t,j}\|_2^2)$ is convex within a small neighborhood of Θ. The existence and uniqueness of a gradient flow over the maximal interval $[0, \tau_{\max}]$ is a standard result (see Section 2.1 of [54]).

For any $\tau \in [0, \tau_{\max}]$, it holds that

$$
\begin{aligned}
\widehat{Q}(\Theta^{(0)}) \geq \widehat{Q}(\Theta^{(0)}) - \widehat{Q}(\Theta^{(\tau)}) &= \int_0^\tau \sum_t \sum_{j=1}^M \langle \frac{d\widehat{Q}(\Theta)}{d\beta_{t,j}}\Big|_{\Theta=\Theta^{(\tau)}}, ML\frac{d\widehat{Q}(\Theta)}{d\beta_{t,j}}\Big|_{\Theta=\Theta^{(\tau)}} \rangle d\tau \\
&= \int_0^\tau \frac{1}{ML} \sum_t \sum_{j=1}^M \|\widehat{G}_\beta(\beta_{t,j}^{(\tau)}, \Theta^{(\tau)}, t)\|^2 d\tau \\
&\geq \frac{1}{ML\tau} \sum_t \sum_{j=1}^M \Big(\int_0^\tau \|\widehat{G}_\beta(\beta_{t,j}^{(\tau)}, \Theta^{(\tau)}, t)\| d\tau \Big)^2,
\end{aligned} \tag{D.29}
$$

where the last inequality follows from Jensen's inequality. (D.29) establishes that $\widehat{Q}(\Theta^{(\tau)})$ is both upper and lower bounded, and $\Theta^{(\tau)}$ exhibits a bounded curve length over any the time interval $[0, \tau_{\max}]$. By compactness, if $\tau_{\max}$ is finite, then $\Theta^{(\tau_{\max})}$ exists and thus must exist beyond $\tau_{\max}$, which leads to contradiction. Therefore, $\tau_{\max} = \infty$, and the well-posedness of the gradient flow for $\tau \geq 0$ consequently follows. Additionally, (D.29) shows that for any $\tau \geq 0$,

$$
\begin{aligned}
\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta_{t,j}\|_2^2 \leq 2\lambda^{-1}\widehat{Q}(\Theta^{(\tau)}) \leq 2\lambda^{-1}\widehat{Q}(\Theta^{(0)}) &= \lambda^{-1}\mathbb{E}_\mu\Big[\Big(\mathrm{Read}[\widehat{T}_{\Theta^{(0)}}(H,1)] - y(H)\Big)^2\Big] \\
&\quad + \frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta_{t,j}^{(0)}\|_2^2 \\
&\leq \lambda^{-1}\mathbb{E}_\mu[(\|\widehat{T}_{\Theta^{(0)}}(H,1)\|_{2-\mathrm{col}} + B)^2] + R^2 \\
&\leq 2\lambda^{-1}\mathbb{E}_\mu[(\|\widehat{T}_{\Theta^{(0)}}(H,1)\|_{2-\mathrm{col}}^2 + B^2] + R^2 \\
&\leq R^2 + \lambda^{-1}\Big(2B^2 + 2B^2\exp\Big(K(1+R+R^2)\Big)^2\Big) \\
&= A_0^2
\end{aligned} \tag{D.30}
$$

The last inequality of (D.30) follows from Lemma C.4, thereby showing that $\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta_{t,j}\|_2^2 \leq A_0^2$ for any $\tau \geq 0$.

As the final part of our proof, we demonstrate that the norm of any entry of Θ is bounded at any given time $\tau \geq 0$. Note that

$$
\begin{aligned}
&\|\widehat{G}(\beta, \Theta^{(\tau)}, t) - \lambda\beta\| \\
\leq& \sum_{i=1}^{N+1} \left\|\nabla_\theta \Big\{ f(\widehat{T}_\Theta(H,t), \theta)_{:,i} \Big\} \widehat{p}_\Theta(H,t)_{:,i} \right\| /2 + \left\| \nabla_w \Big\{ h(\widehat{T}_\Theta(H, t+\Delta/2), w)_{:,i} \Big\} \widehat{p}_\Theta(H, t+\Delta t/2)_{:,i} \right\| /2 \\
\leq& \sup_{i\in[N+1]} (\|\nabla_\theta f(T_\rho(H,t),\theta)_{:,i}\| \sum_{i=1}^{N+1} \|p_\rho(H,t)_{:,i}\|/2 + \|\nabla_w h(T_\rho(H,t), w)_{:,i}\| \sum_{i=1}^{N+1} \|p_\rho(H, t+\Delta t/2)_{:,i}\|/2) \\
\leq& \sqrt{N+1} \sup_{i\in[N+1]} \Big(\|\nabla_\theta f(T_\rho(H,t),\theta)_{:,i}\| \, \|p_\rho(H,t)\|_F/2 \\
&+ \|\nabla_w h(T_\rho(H, t+\Delta t/2), w)_{:,i}\| \, \|p_\rho(H, t+\Delta t/2)\|_F/2 \Big) \\
\leq& \sqrt{N+1} \max\{\phi_P(\|T_\rho(H,t)\|_{2-\mathrm{col}}), \phi_P(\|T_\rho(H, t+\Delta t/2)\|_{2-\mathrm{col}})\} \\
&\max\{\|p_\rho(H,t)\|_F, \|p_\rho(H,t+\Delta t/2)\|_F\}(1+\|\beta\|) \\
\leq& C_R(1+\|\beta\|),
\end{aligned}
\tag{D.31}
$$

Applying (D.31) to the gradient flow

$$
\frac{d\beta_{t,j}^{(\tau)}}{d\tau} = -\widehat{G}(\beta_{t,j}^{(\tau)}, \Theta^{(\tau)}, t)
$$

for $\tau \geq 0$, we have $\frac{d\|\beta_{t,j}^{(\tau)}\|}{d\tau} = \frac{\langle -\widehat{G}(\beta_{t,j}^{(\tau)}, \Theta^{(\tau)}, t), \|\beta^{(\tau)}\| \rangle}{\|\beta^{(\tau)}\|} \leq \|\widehat{G}(\beta_{t,j}^{(\tau)}, \Theta^{(\tau)}, t) - \lambda\beta\| \leq C_R(1+\|\beta^{(\tau)}\|)$. This indicates

$$
\|\beta^{(\tau)}\| \leq (\|\beta\|+1)\exp(C_R\tau) - 1 \leq (R+1)\exp(C_R\tau) - 1 = R_\tau. \tag{D.32}
$$

□

D.5 Proof of Proposition 3.3

Before commencing the proof, we introduce two proxy Transformer procedures in addition to $\bar{T}_\Theta$ and $\widetilde{T}_\rho$. The first proxy, denoted as $\bar{T}_\Theta$, involves moving the layers with the encoder h slightly forward by $\Delta t/2$ in the depth index. This adjustment results in a discrete Transformer with only L layers, where each layer has a step size of Δt and an encoder of $(f+h)/2$, represented by g. Specifically, $\bar{T}_\Theta$ can be written as

$$
\begin{aligned}
\bar{T}_\Theta(H, t+\Delta t) =& \bar{T}_\Theta(H,t) + \Delta t M^{-1} \sum_{j=1}^{M} \Big(f(\bar{T}_\Theta(H,t), \theta_{t,j}) + \sum_{j=1}^{M} h(\bar{T}_\Theta(H,t), w_{t,j}) \Big) \\
=& \bar{T}_\Theta(H,t) + \Delta t M^{-1} \sum_{j=1}^{M} g(\bar{T}_\Theta(H,t), \beta_{t,j}).
\end{aligned}
\tag{D.33}
$$

The second proxy, denoted as $\widetilde{T}_\rho$, extends the width to infinity by letting $M \to \infty$, effectively replacing the average with an integral:

$$
\widetilde{T}_\rho(H, t+\Delta t) = \widetilde{T}_\rho(H,t) + \Delta t \int_\beta g(\widetilde{T}_\rho(H,t), \beta)\rho(\beta|t) d\beta. \tag{D.34}
$$

We let all four Transformers share the same initial state $\widehat{T}_\Theta(H,0) = \bar{T}_\Theta(H,0) = \widetilde{T}_\rho(H,0) = T_\rho(H,0) = H$.

We first present the following lemma, considering parameters i.i.d. drawn from some distribution $\rho \in \mathcal{P}^2$ with bounded support:

Lemma D.6 (Oracle approximation of discretization)**.** *Under Assumptions 1 and 2, suppose that the parameter setting Θ is i.i.d. drawn from $\{\rho(\theta, w|t)\}_{t,j}$ for some $\rho \in \mathcal{P}^2$ concentrated on*

$\{(\theta, w) : \|\theta\|^2 + \|w\|^2 \le r^2\} \times [0,1]$ *and satisfies that* $\int_{\theta,w} \rho(\theta, w, t) d(\theta, w) = 1$ *for any* $t \in [0,1]$. *Then with probability at least* $1 - \exp(-\delta)$ *we have*

$$\|\widehat{T}_\Theta(H,t) - T_\rho(H,t)\|_F \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}.$$

for any H*,* $t = 0, \Delta t, \ldots, (L-1)\Delta t, 1$ *and any* $\delta > 0$*. Here,* $\lesssim$ *hides the dependencies on* N, D, r *and the parameters of the assumptions.*

Proof of Proposition 3.3. Since $\rho \in \mathcal{P}^{2,r}$ has a bounded support for any ρ, there exists some $\rho^* \in \mathcal{P}^{2,r}$ such that $R(\rho^*) = \inf_{\rho \in \mathcal{P}^{2,r}} R(\rho)$. According to Lemma D.6, we can find a specific Θ such that

$$\|\widehat{T}_\Theta(H,t) - T_{\rho^*}(H,t)\|_F \le C\Big(L^{-1} + \sqrt{\frac{\log(L+1)}{M}}\Big),$$

where C depends on N, D, r, and the parameters of the assumptions. Moreover, from Lemma D.6, we ensure that each entry $\beta_{t,j}$ of Θ satisfies $\|\beta_{t,j}\| \le r$. Verification of Lemmas C.1 and C.4 on T_ρ and $\widehat{T}\Theta$ respectively leads to their uniform boundedness, i.e., $\sup_t T_\rho(H,t) \lesssim 1$ and $\sup_t \widehat{T}_\Theta(H,t) \lesssim 1$. Therefore, we have

$$\begin{aligned}
|\widehat{R}(\Theta) - R(\rho^*)| &\le \mathbb{E}_\mu[|\mathrm{Read}[\widehat{T}_\Theta(H,1) - T_\Theta(H,1)]| \cdot ||\mathrm{Read}[\widehat{T}_\Theta(H,1) + T_\Theta(H,1)] + 2y(H)|] \\
&\lesssim L^{-1} + \sqrt{\frac{\log(L+1)}{M}}.
\end{aligned}$$

Here, $\lesssim$ hides the dependencies on N, D, r and the parameters of the assumptions. The result then follows.

The proof for the energy functional Q (and $\widehat{Q}$) follows a similar approach. There exists some $\rho^* \in \mathcal{P}^{2,r}$ such that $Q(\rho^*) = \inf_{\rho \in \mathcal{P}^{2,r}} Q(\rho)$. From Lemmas D.6 and C.7, we can find a specific Θ such that

$$\begin{cases}
\|\widehat{T}_\Theta(H,t) - T_{\rho^*}(H,t)\|_F \le C\Big(L^{-1} + \sqrt{\frac{\log(L+1)}{M}}\Big), \\
|\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta_{t,j}\|^2 - \int_0^1 \int_\beta \|\beta\|^2 \rho(\beta,t) d\beta dt| \le C\Big(L^{-1} + \sqrt{\frac{\log(L+1)}{M}}\Big).
\end{cases}$$

by setting C large enough. Verification of Lemmas C.1 and C.4 on T_ρ and $\widehat{T}_\Theta$ respectively leads to their uniform boundedness. Hence, we have

$$|\widehat{Q}(\Theta) - Q(\rho^*)| \le |\widehat{R}(\Theta) - R(\rho^*)| + \lambda |\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta_{t,j}\|^2 - \int_0^1 \int_\beta \|\beta\|^2 \rho(\beta,t) d\beta dt| \le C(1+\lambda)\Big(L^{-1} + \sqrt{\frac{\log(L+1)}{M}}\Big).$$

The result thus follows. □

E Proofs of main results in Section 4

For simplicity, we assume that Assumption 4 holds with $(g, \alpha) = (f, \theta)$. The proof for the case of $(g, \alpha) = (h, w)$ is symmetric, involving a simple substitution of f with h and θ with w.

E.1 Proofs of Theorem 4.1 and Corollary 4.1

First, the following lemma suggests that as long as the risk $R(\rho)$ remains positive, a descent direction for $Q(\rho)$ can be constructed at any depth index, provided that λ is sufficiently small. This implies that by adjusting λ, one can influence the gradient flow to effectively reduce $Q(\rho)$.

Lemma E.1 (Landscape of $Q(\rho)$)**.** *Suppose that Assumptions 1-4 hold. For any* $\rho \in \mathcal{P}^2$ *concentrated on* P_r *with some* $r > 0$*, any* $w_0 \in \mathbb{R}^{\dim w}$*, and any* $t^* \in [0,1]$ *such that* $\int_\beta \rho(\beta, t^*) \ge 1/2$*, there exists a* $\nu \in \mathcal{P}(\mathbb{R}^{\dim \beta})$ *such that*

i. for any $\beta \in \mathrm{supp}(\nu)$*, we have* $1/B_r \le \|\beta\| \le B_r$

ii. for any (θ_1, θ_2, w)*, we have* $\theta_2 \in \mathcal{K}$ *and* $w = w_0$*.*

iii. $\int_\beta \frac{\delta Q}{\delta\rho}(\beta,t^*)\Big(\nu(\beta)-\rho(\beta|t^*)\Big)d\beta \le C_1\lambda - C_2R(\rho)$

Here, B_r, C_1, C_2 are universal constants that depends on N, d, r and the parameters of the assumptions.

Given the t^* and w_0 specified in the theorem, Lemma E.1 indicates that there exists some

$$\nu \in \mathcal{P}\Big(\Big(B_{\dim\theta_1}(0, B_{R_\infty})/B_{\dim\theta_1}(0, 1/B_{R_\infty})\Big)\times\mathcal{K}\times\{w_0\}\Big)$$

such that $\int_\beta \frac{\delta Q}{\delta\rho}|_{\rho_\infty}(\beta,t^*)\Big(\nu(\beta)-\rho_\infty(\beta|t^*)\Big)d\beta \le C_1\lambda - C_2R(\rho)$, where B_{R_∞}, C_1, C_2 are universal constants dependent on N, d, R_∞ and the parameters of the assumptions.

In addition, for any $\rho\in\mathcal{P}^2$, we define the following two functional derivatives:

$$\frac{\delta Q_f}{\delta\rho}(\theta,w,t) = \mathbb{E}_\mu\Big[\mathrm{Tr}\Big(\Big[f(T_\rho(H,t),\theta)\Big]^T p_\rho(H,t)\Big)\Big] + \lambda\|\theta\|_2^2$$
$$\frac{\delta Q_h}{\delta\rho}(\theta,w,t) = \mathbb{E}_\mu\Big[\mathrm{Tr}\Big(\Big[h(T_\rho(H,t),w)\Big]^T p_\rho(H,t)\Big)\Big] + \lambda\|w\|_2^2.$$

It is obvious that $\frac{\delta Q}{\delta\rho} \equiv (\frac{\delta Q_f}{\delta\rho} + \frac{\delta Q_h}{\delta\rho})/2$.

Proof of Theorem 4.1. Our proof consists of three steps, each aimed at bounding the differences related to the energy functional Q or $\widehat{Q}$.

Step I: Show that $\frac{\delta Q}{\delta\rho}|_{\rho_\infty}(\beta,t)$ is continuous with respect to (β,t)

In Step I of the proof of Lemma E.1, we establish that $\rho_\infty\in\mathcal{P}^2$, with a bounded support P_{R_∞}, implies $p_{\rho_\infty}(H,t)$ is C_p-Lipschitz continuous, where C_p is a universal constant dependent solely on N, D, R_∞, and the parameters in our assumptions.

Next, we would like to show that $\frac{\delta Q}{\delta\rho}{}_{\rho_\infty}(\beta,t)$ is continuous with respect to $(\beta,t)\in\mathbb{R}^{\mathrm{div}\beta}\times[0,1]$. Let's focus on the region $(\beta,t)\in P_r$ for any $r>R_\infty$, so that ρ_∞ is also concentrated on P_r. It's noteworthy that for any bounded support $(\beta,t)\in P_r$, Lemma C.1 and Assumption 2 (i) ensure the universal boundedness of $g(T_\rho(H,t),\beta)$, and Lemma C.3 ensures the universal boundedness of $p_\rho(H,t)$ for any $H\in\mathrm{supp}(\mu)$, with the constants depending solely on N, D, r, and the parameters of the assumptions.

Combining the Lipschitz continuity of $p_{\rho_\infty}(H,t)$ and $T_{\rho_\infty}(H,t)$ with respect to (H,t), as shown in Proposition C.2, along with the Lipschitz continuity of $g(T,\beta)$ with respect to (T,β) when $\|T\|_F$ is universally bounded (as guaranteed by Assumption 2 (ii) and (iii)), and their universal boundedness, we derive that $\mathrm{Tr}\Big(\Big[g(T_{\rho_\infty}(H,t),\beta)\Big]^T p_{\rho_\infty}(H,t)\Big)$ is C_G-Lipschitz continuous for $\|\cdot\|_2$ with respect to $(\beta,t)\in P_r$ for some universal constant C_G that depends only on N, D, r and the parameters of the assumptions. Since the Lipschitz constant C_G is independent of the choice of H, we see that $\mathrm{Tr}\Big(\Big[g(T_{\rho_\infty}(H,t),\beta)\Big]^T p_{\rho_\infty}(H,t)\Big)$ is uniformly continuous across all $H\in\mathrm{supp}(\mu)$ with respect to $(\beta,t)\in P_r$. Consequently, we have that

$$\frac{\delta Q}{\delta\rho}(\beta,t)|_{\rho_\infty} = \mathbb{E}_\mu\Big[\mathrm{Tr}\Big(\Big[g(T_{\rho_\infty}(H,t),\beta)\Big]^T p_{\rho_\infty}(H,t)\Big)\Big] + \frac{\lambda}{2}\|\beta\|_2^2$$

is continuous with respect to $(\beta,t)\in P_r$. Since the choice of r is arbitrary, we conclude that $\frac{\delta Q}{\delta\rho}|_{\rho_\infty}(\beta,t)$ is continuous with respect to $(\beta,t)\in\mathbb{R}^{\mathrm{div}\beta}\times[0,1]$.

Step II: Show that $Q(\rho_\infty)\lesssim\lambda$

In the first part of the proof, we will adopt a similar approach to Theorem 3.9 of [41] to demonstrate that the stationary point of the Wasserstein gradient flow, denoted ρ_∞, satisfies $Q(\rho_\infty)\lesssim\lambda$. It's worth noting that in [41], the authors assume $\lambda=0$ and conclude $R(\rho_\infty)=Q(\rho_\infty)=0$, but this claim relies on assuming the global existence of the Wasserstein gradient flow rather than proving it directly.

Based on the pivotal findings from [47] regarding the stationary points in the Wasserstein space, we infer that the stationary point ρ_∞ of the Wasserstein gradient flow (3.5), i.e.

$$\frac{d\rho(\beta,t)}{d\tau} = \mathrm{div}_\beta\Big(\rho\nabla_\beta\frac{\delta Q}{\delta\rho}\Big),$$

must satisfy $\nabla_\beta\frac{\delta Q}{\delta\rho}|_{\rho_\infty}= 0$ almost everywhere over $\mathrm{supp}(\rho_\infty)$. This further indicates that $\nabla_\beta\frac{\delta Q}{\delta\rho}|_{\rho_\infty}(\theta_1,\theta_2,w,t^*) = 0$ almost everywhere over $\mathrm{supp}(\rho_\infty(\cdot,t^*))$. The fact $\rho(\cdot,t^*)$ is a connected set, coupled with the continuity of the Frechét differential $\frac{\delta Q}{\delta\rho}|_{\rho_\infty}$ with respect to β, implies that, $\frac{\delta Q}{\delta\rho}|_{\rho_\infty}(\theta_1,\theta_2,w,t^*) = C$ for some constant C over $(\beta,t^*)\in\mathrm{supp}(\rho(\cdot,t^*))$.

Given the separation assumption on the support of $\rho_\infty(\cdot,t^*)$, we ensure that for any $(\theta_1,\theta_2)\in\Big(B_{\mathrm{dim}\theta_1}(0,B_{R_\infty})/B_{\mathrm{dim}\theta_1}(0,1/B_{R_\infty})\Big)\times\mathcal{K}$, there exists $c\in\mathbb{R}$, $1/R_\infty B_{R_\infty}\le|c|\le R_\infty B_{R_\infty}$ such that $(c\theta_1,\theta_2,w_0,t^*)\in\mathrm{supp}(\rho_\infty)$. Combined with Assumption 4 (i), which implies the 1-homogeneity of $f(T,\theta_1,\theta_2)$ with respect to θ_1, we have

$$\begin{aligned}
&\frac{\delta Q_f}{\delta\rho}\bigg|_{\rho_\infty}(c\theta_1,\theta_2,w_0,t^*) = c\frac{\delta Q_f}{\delta\rho}\bigg|_{\rho_\infty}(\theta_1,\theta_2,w_0,t^*) + (|c|-1)\lambda\|\theta_1\|^2.\\
&\frac{\delta Q_h}{\delta\rho}\bigg|_{\rho_\infty}(c\theta_1,\theta_2,w_0,t^*) = \frac{\delta Q_h}{\delta\rho}\bigg|_{\rho_\infty}(\theta_1,\theta_2,w_0,t^*).
\end{aligned}\tag{E.1}$$

Hence, given that $\nabla_\beta\frac{\delta Q}{\delta\rho}|_{\rho_\infty}(\cdot,t^*) = 0$ almost everywhere over $\mathrm{supp}(\rho_\infty(\cdot,t^*))$, it also holds that

$$\begin{aligned}
\nabla_{(\theta_1,\theta_2,w)}\frac{\delta Q}{\delta\rho}\bigg|_{\rho_\infty}(\theta_1,\theta_2,w_0,t^*) =&\Big(\nabla_{(\theta_1,\theta_2,w)}\frac{\delta Q_f}{\delta\rho}\bigg|_{\rho_\infty}(\theta_1,\theta_2,w_0,t^*) + \nabla_{(\theta_1,\theta_2,w)}\frac{\delta Q_h}{\delta\rho}\bigg|_{\rho_\infty}(\theta_1,\theta_2,w_0,t^*)\Big)/2\\
=&\Big(-\frac{|c|-1}{c}\lambda\theta_1,0_{\mathrm{dim}\theta_2},0_w\Big),
\end{aligned}$$

which implies

$$\bigg\|\nabla_{(\theta_1,\theta_2,w)}\frac{\delta Q}{\delta\rho}\bigg|_{\rho_\infty}(\theta_1,\theta_2,w_0,t^*)\bigg\| \le (R_\infty^2 B_{R_\infty}+R_\infty)\lambda.\tag{E.2}$$

Given the condition that $\|(c\theta_1,\theta_2,w_0,t^*)-(\theta_1,\theta_2,w_0,t^*)\|\le R_\infty^2 B_{R_\infty}$, and recalling that $\frac{\delta Q}{\delta\rho}|_{\rho_\infty}(\theta_1,\theta_2,w,t^*)\equiv C$ across $(\beta,t^*)\in\mathrm{supp}(\rho(\cdot,t^*))$, (E.2) further indicates that

$$\bigg|\frac{\delta Q}{\delta\rho}\bigg|_{\rho_\infty}(\theta_1,\theta_2,w_0,t^*) - C\bigg| \le (R_\infty^4 B_{R_\infty}^2 + R_\infty^3 B_{R_\infty})\lambda.\tag{E.3}$$

for any $(\theta_1,\theta_2)\in\Big(B_{\mathrm{dim}\theta_1}(0,B_{R_\infty})/B_{\mathrm{dim}\theta_1}(0,1/B_{R_\infty})\Big)\times\mathcal{K}$. Hence, we have

$$\begin{aligned}
C_1\lambda - C_2R(\rho_\infty) \ge& \int_\beta\frac{\delta Q}{\delta\rho}\bigg|_{\rho_\infty}(\beta,t^*)\Big(\nu(\beta)-\rho_\infty(\beta|t^*)\Big)d\beta\\
=&\int_\beta\Big(\frac{\delta Q}{\delta\rho}\bigg|_{\rho_\infty}(\beta,t^*)-C\Big)\Big(\nu(\beta)-\rho_\infty(\beta|t^*)\Big)d\beta\\
\ge&-\int_\beta(R_\infty^4 B_{R_\infty}^2 + R_\infty^3 B_{R_\infty})\lambda\Big(\nu(\beta)+\rho_\infty(\beta|t^*)\Big)d\beta\\
\ge&2(R_\infty^4 B_{R_\infty}^2 + R_\infty^3 B_{R_\infty})\lambda.
\end{aligned}\tag{E.4}$$

Therefore, we have $R(\rho_\infty)\le\frac{C_1+2(R_\infty^4 B_{R_\infty}^2+R_\infty^3 B_{R_\infty})}{C_2}\lambda$, and

$$Q(\rho_\infty)\le R(\rho_\infty)+\int_0^1\int_\beta\|\beta\|^2\rho(\beta,t)d\beta dt \le \Big(\frac{C_1+2(R_\infty^4 B_{R_\infty}^2+R_\infty^3 B_{R_\infty})}{C_2}+R_\infty^2\Big)\lambda,\tag{E.5}$$

which completes the first part of our proof.

Step III: Bound the difference between $Q(\rho^{(\tau)})$ and $Q(\rho_\infty)$ when τ is large

Proposition 3.2 establishes that the second moment for $\rho^{(\tau)}$ is uniformly bounded across all $\tau \geq 0$:

$$\int_0^1 \int_\beta \|\beta\|^2 \rho^{(\tau)}(\beta, t) d\beta dt \leq A_0^2,$$

where A_0 is defined as in Proposition 3.2. Therefore, the weak convergence of probability measures $(\rho^{(\tau)})_{\tau \geq 0}$ is equivalent to the convergence in the Wasserstein-2 distance, i.e.

$$\lim_{\tau \to \infty} W_2(\rho^{(\tau)}, \rho_\infty) = 0. \tag{E.6}$$

When τ is sufficiently large, $\rho^{(\tau)}$ concentrates on P_{R_∞}. Therefore, according to Lemma C.2, there exists a constant C_0 depending solely on N, D, R_∞, and the parameters of the assumptions that

$$\|T_{\rho_\infty}(H, t) - T_{\rho^{(\tau)}}(H, t)\|_F \leq C_0 W_2(\rho^{(\tau)}, \rho_\infty) \tag{E.7}$$

for any H and $t \in [0, 1]$ when τ is sufficiently large. Note that Lemma C.1 shows that

$$\max\{\|T_{\rho^{(\tau)}}(H, t)\|_{2-\mathrm{col}}, \|T_{\rho_\infty}(H, t)\|_{2-\mathrm{col}}\} \leq B \exp(K(1 + R_\infty + R_\infty^2)) =: B_T$$

for any H and $t \in [0, 1]$. Thus, from (E.7), we have

$$\begin{aligned}
|Q(\rho_\infty) - Q(\rho^{(\tau)})| \leq & \frac{1}{2}\mathbb{E}_\mu\Big[|\mathrm{Read}[|T_{\rho^{(\tau)}}(H, 1)]| + |T_{\rho_\infty}(H, 1)]| + 2|y(H)|\Big] \|T_{\rho_\infty}(H, 1) - T_{\rho^{(\tau)}}(H, 1)\|_F \\
& + \frac{\lambda}{2} \int_0^1 \int_\beta \|\beta\|^2 (\rho_\infty - \rho^{(\tau)})(\beta, t) d\beta dt \\
\leq & (B_T + B) C_0 W_2(\rho^{(\tau)}, \rho_\infty) + \frac{\lambda}{2} R_\infty W_1(\rho^{(\tau)}, \rho_\infty) \\
\leq & ((B_T + B) C_0 + \frac{\lambda}{2} R_\infty) W_2(\rho^{(\tau)}, \rho_\infty),
\end{aligned} \tag{E.8}$$

where the second inequality incorporates the Kantorovich-Rubinstein Theorem (see Theorem 5.10 of [61], for example) and the $2R_\infty$-Lipschitz continuity of $\|\beta\|^2$ over the region $(\beta, t) \in P_{R_\infty}$. Combining equations (E.6) and (E.8), we deduce that for any $\epsilon > 0$, there exists some $\tau_0 > 0$ such that $|Q(\rho_\infty) - Q(\rho^{(\tau_0)})| \leq \epsilon$.

Step IV: Complete the proof by bounding the difference between $\widehat{Q}(\Theta^{(\tau)})$ and $Q(\rho^{(\tau)})$ when τ is large

The final step can be seen as a direct corollary of the approximation result in Theorem 3.1. According to Theorem 3.1, there exists a constant C_1 dependent on N, D, τ_0, λ, and the parameters specified in the assumptions, such that

$$|\widehat{Q}(\Theta^{(\tau_0)}) - Q(\rho^{(\tau_0)})| \leq C_1 \Big(L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}\Big)$$

with probability at least $1 - 3\exp(-\delta)$ for any $\delta > 0$. Combining the outcomes from the preceding steps, we obtain

$$\widehat{Q}(\Theta^{(\tau_0)}) \leq \epsilon + C_1 \Big(L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}\Big) + C_2 \lambda. \tag{E.9}$$

Note that

$$\frac{d}{d\tau}\widehat{Q}(\Theta^{(\tau)}) = \sum_t \sum_{j=1}^M \langle \frac{d\widehat{Q}(\Theta)}{d\beta_{t,j}}\Big|_{\Theta = \Theta^{(\tau)}}, -ML \frac{d\widehat{Q}(\Theta)}{d\beta_{t,j}}\Big|_{\Theta = \Theta^{(\tau)}} \rangle = -\frac{1}{ML} \sum_t \sum_{j=1}^M \|\widehat{G}_\beta(\beta_{t,j}^{(\tau)}, \Theta^{(\tau)}, t)\|^2 \leq 0,$$

so the sequence $(\widehat{Q}(\Theta^{(\tau)}))_{\tau \geq 0}$ is non-decreasing. Hence, for any $\tau \geq \tau_0$,

$$\widehat{R}(\Theta^{(\tau)}) \leq \widehat{Q}(\Theta^{(\tau)}) \leq \widehat{Q}(\Theta^{(\tau_0)}) \leq \epsilon + C_1 \Big(L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}\Big) + C_2 \lambda,$$

which completes the proof, recalling that C_1 depends only on N, D, τ_0, λ and the parameters of the assumptions, and C_2 depends only on N, D, R_∞, and the parameters of the assumptions.

□

Proof of Corollary 4.1. Given the choice of λ By Theorem 4.1, there exists some $\tau_0 > 0$ such that

$$\begin{aligned}\sup_{\tau\geq\tau_0} \widehat{R}(\Theta^{(\tau)}) \leq& \epsilon/2 + C_1\Big(L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}\Big) + C_2 C_\lambda \lambda \\ \leq& (1/4 + C_2C_\lambda)\epsilon + C_1\Big(L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}\Big).\end{aligned}$$

The result holds by setting L and $M/\log L$ sufficiently large to ensure that

$$C_1\Big(L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}\Big) \leq C_1\Big(L^{-1} + \sqrt{\frac{2(1+\delta)\log(L+1)}{M}}\Big) \leq \epsilon/4.$$

□

F Proofs of auxiliary results

F.1 Proof of Lemma D.1

Proof. Lemma C.1 confirms that $\|T_\rho(H,t)\|_{2-\mathrm{col}}$ and $\|T_\nu(H,t)\|_{2-\mathrm{col}}$ are bounded universally by $B_T = B\exp(K(1+r+r^2))$. Considering the definition (C.2), it suffices to demonstrate that

$$\Big\|\mathbb{E}_\mu\Big[\nabla_\beta \mathrm{Tr}\Big(g(T_\rho(H,t),\beta)^T p_\rho(H,t)\Big) - \nabla_\beta \mathrm{Tr}\Big(g(T_\nu(H,t),\widetilde{\beta})^T p_\nu(H,t)\Big)\Big]\Big\| \leq J_1 + J_2,$$

where

$$\begin{aligned}J_1 :=& \Big\|\mathbb{E}_\mu\Big[\nabla_\beta \mathrm{Tr}\Big((g(T_\nu(H,t),\widetilde{\beta}) - g(T_\rho(H,t),\beta))^T p_\nu(H,t)\Big)\Big]\Big\| \lesssim \|\beta - \widetilde{\beta}\|, \\ J_2 :=& \Big\|\mathbb{E}_\mu\Big[\nabla_\beta \mathrm{Tr}\Big(g(T_\rho(H,t),\beta)^T (p_\nu(H,t) - p_\rho(H,t))\Big)\Big]\Big\| \lesssim \exp(C_G W_1(\rho,\nu)) - 1.\end{aligned}$$

Here, the symbol $\lesssim$ hides dependencies on N, D, r, and the parameters of the assumptions. To bound J_1, consider that

$$\begin{aligned}J_1 \leq& \sup_{i\in[N+1]} \mathbb{E}_\mu\Big[\Big\|g(T_\nu(H,t),\widetilde{\beta})_{:,i} - g(T_\rho(H,t),\beta)_{:,i}\Big\| \sum_{i=1}^{N+1} \|p_\nu(H,t)_{:,i}\|\Big] \\ \leq& \sqrt{N+1} \sup_{i\in[N+1]} \mathbb{E}_\mu\Big[\Big\|g(T_\nu(H,t),\widetilde{\beta})_{:,i} - g(T_\rho(H,t),\beta)_{:,i}\Big\| \|p_\nu(H,t)\|_F\Big] \\ \lesssim& \sup_{i\in[N+1]} \mathbb{E}_\mu\Big[\Big\|g(T_\nu(H,t),\widetilde{\beta})_{:,i} - g(T_\rho(H,t),\beta)_{:,i}\Big\|\Big] \\ \leq& \phi_{PT}(r,B_T)\|T_\rho(H,t) - T_\rho(H,t)\|_{2-\mathrm{col}} + \phi_P P(r,B_T)\|\beta - \widetilde{\beta}\| \\ \lesssim& W_1(\rho,\nu) + \|\beta - \widetilde{\beta}\|.\end{aligned} \tag{F.1}$$

The third inequality in Equation (F.1) is derived from Lemma C.3, while the fourth inequality relies on Assumption 3 (i) and (iii). Lastly, bounding $\|T_\rho(H,t) - T_\rho(H,t)\|_{2-\mathrm{col}}$ by $W_1(\rho,\nu)$ is achieved with Lemma C.2.

On the other hand, to bound J_2, we have

$$\begin{aligned}J_2 \leq& \sqrt{N+1} \sup_{i\in[N+1]} \mathbb{E}_\mu\Big[\|g(T_\rho(H,t),\beta)_{:,i}\| \, \|p_\nu(H,t) - p_\rho(H,t)\|_F\Big] \\ \leq& \sqrt{N+1}\|(T_\rho(H,t),\beta)\|_{2-\mathrm{col}}\|p_\nu(H,t) - p_\rho(H,t)\|_F \\ \lesssim& \|p_\nu(H,t) - p_\rho(H,t)\|_F.\end{aligned} \tag{F.2}$$

In (F.2), the third inequality relies on Assumption 2 (i). Consequently, to establish $J_2 \lesssim \exp(C_G W_1(\rho,\nu)) - 1$, it is adequate to demonstrate that $\|p_\nu(H,t) - p_\rho(H,t)\|_F \leq I_1 + I_2 \lesssim W_1(\rho,\nu)$, where

$$\begin{aligned}I_1 =& |\mathrm{Read}[T_\rho(H,1) - T_\nu(H,1)]| \Big\|\exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,s),\beta)]\rho(\beta,s)d\beta ds\Big)\Big\|, \\ I_2 =& |\mathrm{Read}[T_\rho(H,1)] - y(H)| \\ & \Big\|\exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta ds\Big) - \exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,s),\beta)]\rho(\beta,s)d\beta ds\Big)\Big\|.\end{aligned}$$

From Lemma C.2, it is trivial that $|\mathrm{Read}[T_\rho(H,1) - T_\nu(H,1)]| \leq \|T_\rho(H,1) - T_\nu(H,1)\|_F \lesssim W_1(\rho,\nu)$. Thus, $I_1 \lesssim W_1(\rho,\nu)$ given the boundedness of $\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,t),\beta)]\|$ as provided in Assumption 2 (iii). To bound I_2, we have

$$
\begin{aligned}
I_2 &\lesssim \Big\|\exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta ds\Big) - \exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,s),\beta)]\rho(\beta,s)d\beta ds\Big)\Big\| \\
&\lesssim \Big\|\exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta ds - \int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,s),\beta)]\rho(\beta,s)d\beta ds\Big) - I_{\mathrm{dimvec}[T]}\Big\| \\
&\lesssim \exp\Big(\int_t^1\int_\beta \big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,s),\beta)] - \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,s),\beta)]\big\|\,\rho(\beta,s)d\beta ds\Big) - 1 \\
&\lesssim \exp\Big(\phi_{TT}(N,D,B_T,r)\|g(T_\rho(H,t),\beta) - g(T_\nu(H,t),\beta)\|_F\Big) - 1 \\
&\lesssim \exp\Big(C_r\phi_{TT}(N,D,B_T,r)\phi_T(N,D,\sqrt{N+1}B_T)(1+r+r^2)W_1(\rho,\nu)\Big) - 1
\end{aligned}
\tag{F.3}
$$

for some universal constant C_r. Here, the first inequality in (F.3) stems from $|\mathrm{Read}[T_\rho(H,1)] - y(H)| \leq B_T + B$, the second inequality is ensured by the boundedness of $\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,t),\beta)]\|$ as stated in Assumption 2 (iii), and the fourth inequality is provided by Assumption 3 (iv). The last inequality in (F.3) arises from Assumption 2 (iii) and Lemma C.2. By combining Equation (F.3) with the bounds of J_1 and I_1, we deduce that $J_1 + J_2 \lesssim \exp(C_G W_1(\rho,\nu)) - 1 + \|\beta - \widetilde{\beta}\|$ for some universal constant C_G large enough, thereby completing the proof. □

F.2 Proof of Lemma D.2

Proof. Lemma C.4 demonstrates that $\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}}$ and $\|\widehat{T}_{\widetilde{\Theta}}(H,t)\|_{2-\mathrm{col}}$ are bounded by $B_T = B\exp(K(1+A+A^2))$ for any H and $t \in [0,1]$. We begin by bounding

$$
\begin{aligned}
\|\widehat{G}(\beta,\Theta,t) - \widehat{G}(\beta,\widetilde{\Theta},t)\| = \Big\{&\frac{1}{2}\mathbb{E}_\mu\Big[\nabla_\theta \mathrm{Tr}\Big(f(\widehat{T}_\Theta(H,t),\theta) - f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)\Big)^T\widehat{p}_\Theta(H,t+\Delta t/2)\Big]^T \\
&+\frac{1}{2}\mathbb{E}_\mu\Big[\nabla_\theta \mathrm{Tr} f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)^T\Big(\widehat{p}_\Theta(H,t+\Delta t/2) - \widehat{p}_{\widetilde{\Theta}}(H,t+\Delta t/2)\Big)\Big]^T, \\
&\frac{1}{2}\mathbb{E}_\mu\Big[\nabla_w \mathrm{Tr}\Big(h(\widehat{T}_\Theta(H,t+\Delta t/2),w) - h(\widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t/2),w)\Big)^T\widehat{p}_\Theta(H,t)\Big]^T \\
&+\frac{1}{2}\mathbb{E}_\mu\Big[\nabla_w \mathrm{Tr} h(\widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t/2),w)^T\Big(\widehat{p}_\Theta(H,t) - \widehat{p}_{\widetilde{\Theta}}(H,t)\Big)\Big]^T\Big\}^T.
\end{aligned}
$$

To demonstrate that $\|\widehat{G}(\beta,\Theta,t) - \widehat{G}(\beta,\widetilde{\Theta},t)\| \leq C_G\frac{1}{ML}d(\Theta,\widetilde{\Theta})$, it suffices to show

$$
J_1 := \Big\|\mathbb{E}_\mu\Big[\nabla_\theta \mathrm{Tr}\Big(f(\widehat{T}_\Theta(H,t),\theta) - f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)\Big)^T\widehat{p}_\Theta(H,t+\Delta t/2)\Big]\Big\| \leq C_G\frac{1}{ML}d(\Theta,\widetilde{\Theta}),
\tag{F.4}
$$

and

$$
J_2 := \Big\|\mathbb{E}_\mu\Big[\nabla_\theta \mathrm{Tr} f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)^T\Big(\widehat{p}_\Theta(H,t+\Delta t/2) - \widehat{p}_{\widetilde{\Theta}}(H,t+\Delta t/2)\Big)\Big]\Big\| \leq C_G\frac{1}{ML}d(\Theta,\widetilde{\Theta}),
\tag{F.5}
$$

as the other part for h and w follows a similar proof approach.

To bound J_1, by Assumption 3 (i) we have

$$
\begin{aligned}
J_1 \leq & \sum_{i=1}^{N+1} \mathbb{E}_\mu\Big[\Big\| \nabla_\theta f(\widehat{T}_\Theta(H,t),\theta)_{:,i} - f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)_{:,i} \Big\| \, \|\widehat{p}_\Theta(H,t+\Delta t/2)_{:,i}\| \Big] \\
\leq & \mathbb{E}_\mu\Big[\sup_{i\in[N+1]} \Big\| \nabla_\theta f(\widehat{T}_\Theta(H,t),\theta)_{:,i} - f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)_{:,i} \Big\| \sum_{i=1}^{N+1} \|\widehat{p}_\Theta(H,t+\Delta t/2)_{:,i}\| \Big] \\
\leq & \phi_T(r,B_T) \|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_{2-\mathrm{col}} \sqrt{N+1} \|\widehat{p}_\Theta(H,t+\Delta t/2)\|_F \\
\lesssim & \phi_T(r,B_T)\sqrt{N+1}\|\widehat{p}_\Theta(H,t+\Delta t/2)\|_F \frac{1}{ML} d(\Theta,\widetilde{\Theta}) \\
\lesssim & \frac{1}{ML} d(\Theta,\widetilde{\Theta}),
\end{aligned}
\tag{F.6}
$$

where the fourth inequality utilizes Lemma C.5 and the last inequality uses Lemma C.6.

To bound J_2, by Assumption 2 (ii), we have

$$
\begin{aligned}
J_2 \leq & \sqrt{N+1}\mathbb{E}_\mu\Big[\sup_{i\in[N+1]} \Big\| f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)_{:,i} \Big\| \, \|\widehat{p}_\Theta(H,t+\Delta t/2) - \widehat{p}_{\widetilde{\Theta}}(H,t+\Delta t/2)\|_F \Big] \\
= & \sqrt{N+1}\phi(B_T)(1+r)\mathbb{E}_\mu\Big[\|\widehat{p}_\Theta(H,t+\Delta t/2) - \widehat{p}_{\widetilde{\Theta}}(H,t+\Delta t/2)\|_F \Big].
\end{aligned}
\tag{F.7}
$$

Hence, it suffices to show that $\|\widehat{p}_\Theta(H,t+\Delta t/2) - \widehat{p}_{\widetilde{\Theta}}(H,t+\Delta t/2)\|_F \lesssim \frac{1}{ML} d(\Theta,\widetilde{\Theta})$ to establish $J_2 \leq \frac{1}{ML} d(\Theta,\widetilde{\Theta})$. Recalling the formula $\widehat{p}_\Theta$ in (C.16), we have $\|\widehat{p}_\Theta(H,t+\Delta t/2) - \widehat{p}_{\widetilde{\Theta}}(H,t+\Delta t/2)\|_F \leq I_1 + I_2$, where

$$
\begin{aligned}
I_1 = & (\mathrm{Read}[\widehat{T}_\Theta(H,1) - \widehat{T}_{\widetilde{\Theta}}(H,1)] \\
& \Big\{ \prod_{\substack{(s-t)/\Delta t+2 \in [(1-t)/\Delta t] \\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2) M^{-1} \sum_{j=1}^M \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})] \Big) \\
& \prod_{\substack{(s-t)/\Delta t+1 \in [(1-t)/\Delta t] \\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2) M^{-1} \sum_{j=1}^M \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})] \Big) \Big\}_{DN+d+1,:} \\
\leq & \frac{1}{ML} d(\Theta,\widetilde{\Theta}) \\
& \Big\| \prod_{\substack{(s-t)/\Delta t+2 \in [(1-t)/\Delta t] \\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2) M^{-1} \sum_{j=1}^M \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})] \Big) \\
& \prod_{\substack{(s-t)/\Delta t+1 \in [(1-t)/\Delta t] \\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2) M^{-1} \sum_{j=1}^M \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})] \Big) \Big\| \\
\leq & \frac{1}{ML} d(\Theta,\widetilde{\Theta}) \\
& \exp\Big((\Delta t/2) \sum_{(s-t)/\Delta t+1 \in [(1-t)/\Delta t]} M^{-1} \sum_{j=1}^M \Big\| \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})] \Big\| \\
+ & (\Delta t/2) \sum_{(s-t)/\Delta t+1 \in [(1-t)/\Delta t]} M^{-1} \sum_{j=1}^M \Big\| \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(\widehat{T}_\Theta(H,t),w_{s,j})] \Big\| \Big) \\
\leq & \frac{1}{ML} d(\Theta,\widetilde{\Theta}) \exp\Big(\phi_T(N,D,\sqrt{N+1}KB_T)(1+r+r^2) \Big) \\
\lesssim & \frac{1}{ML} d(\Theta,\widetilde{\Theta}),
\end{aligned}
$$

and

$$
\begin{aligned}
I_2 =&|\mathrm{Read}[\widehat{T}_{\widetilde{\Theta}}(H,1)] + y(H)| \\
&\Big\{ \prod_{\substack{(s-t)/\Delta t+2\in[(1-t)/\Delta t]\\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_{\Theta}(H,s),\theta_{s,j})]\Big) \\
&\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_{\Theta}(H,s+\Delta t/2),w_{s,j})]\Big) - \\
&\prod_{\substack{(s-t)/\Delta t+2\in[(1-t)/\Delta t]\\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta}_{s,j})]\Big) \\
&\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}} \Big(I_{\mathrm{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_{\widetilde{\Theta}}(H,t+\Delta/2),\widetilde{w}_{s,j})]\Big)\Big\}_{DN+d+1,:} \\
\le&(B+B_T) \\
&\cdot \exp\Big((\Delta t/2)\sum_{(s-t)/\Delta t+1\in[(1-t)/\Delta t]} M^{-1}\sum_{j=1}^{M}\max\{\Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_{\Theta}(H,s),\theta_{s,j})]\Big\|, \Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta}_{s,j})]\}\Big\| \\
&+(\Delta t/2)\sum_{(s-t)/\Delta t+1\in[(1-t)/\Delta t]} M^{-1}\sum_{j=1}^{M}\max\{\Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_{\Theta}(H,t),w_{s,j})]\Big\|, \Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{w}_{s,j})]\}\Big\|\Big) \\
&\cdot(\Delta t/2)\sum_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}}\Big(\Big\|M^{-1}\sum_{j=1}^{M}\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_{\Theta}(H,s),\theta_{s,j})] - M^{-1}\sum_{j=1}^{M}\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta}_{s,j})]\Big\| \\
&+\Big\|M^{-1}\sum_{j=1}^{M}\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_{\Theta}(H,t),w_{s,j})] - M^{-1}\sum_{j=1}^{M}\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{w}_{s,j})]\Big\|\Big) \\
\le&(B+B_T)\exp\Big(\phi_T(N,D,\sqrt{N+1}KB_T)(1+r+r^2)\Big)\phi_{TP}(N,D,\sqrt{N+1}B_T,r)\frac{1}{ML}d(\Theta,\widetilde{\Theta}) \\
\lesssim&\frac{1}{ML}d(\Theta,\widetilde{\Theta}).
\end{aligned}
$$

where the first inequality applies Lemma C.8, and the second inequality relies on Assumption 2 (iii) and Assumption 3 (ii). Therefore, we conclude that $I_2 \lesssim \frac{1}{ML}d(\Theta,\widetilde{\Theta})$. By combining the bounds of I_1 and I_2, we observe that Equation (F.5) holds, thereby establishing $\|\widehat{G}(\beta,\Theta,t)-\widehat{G}(\beta,\widetilde{\Theta},t)\| \le C_G\frac{1}{ML}d(\Theta,\widetilde{\Theta})$ for some C_G dependent on N, d, r, and the parameters of the assumptions.

It remains to prove that $\|\widehat{G}(\beta,\widetilde{\Theta},t)-\widehat{G}(\widetilde{\beta},\widetilde{\Theta},t)\| \le C_G(1+\lambda)\|\beta-\widetilde{\beta}\|$. Note that

$$
\begin{aligned}
\|\widehat{G}(\beta,\widetilde{\Theta},t)-\widehat{G}(\widetilde{\beta},\widetilde{\Theta},t)\| \le&\lambda\|\beta-\widetilde{\beta}\| \\
&+\Big\{\frac{1}{2}\mathbb{E}_\mu\Big[\nabla_\theta\mathrm{Tr}\Big(f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)-f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta})\Big)^T\widehat{p}_{\Theta}(H,t+\Delta t/2)\Big]^T, \\
&\frac{1}{2}\mathbb{E}_\mu\Big[\nabla_w\mathrm{Tr}\Big(h(\widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t/2),w)-h(\widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t/2),\widetilde{w})\Big)^T\widehat{p}_{\Theta}(H,t)\Big]^T\Big\}^T.
\end{aligned}
$$

Therefore, we only need to show that

$$
\Big\|\mathbb{E}_\mu\Big[\nabla_\theta\mathrm{Tr}\Big(f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)-f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta})\Big)^T\widehat{p}_{\Theta}(H,t+\Delta t/2)\Big]\Big\| \le C_G\|\theta-\widetilde{\theta}\|,
$$

and

$$
\Big\|\mathbb{E}_\mu\Big[\nabla_w\mathrm{Tr}\Big(h(\widehat{T}_{\widetilde{\Theta}}(H,t),w)-h(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{w})\Big)^T\widehat{p}_{\Theta}(H,t)\Big]\Big\| \le C_G\|w-\widetilde{w}\|,
$$

to obtain $\|\widehat{G}(\beta,\widetilde{\Theta},t)-\widehat{G}(\widetilde{\beta},\widetilde{\Theta},t)\| \le C_G(1+\lambda)\|\beta-\widetilde{\beta}\|$. Here, we only establish the inequality above for f and θ, as the proof of the other inequality follows a similar pattern. Note that by Assumption 3 (iii), we have

$$
\begin{aligned}
&\left\|\mathbb{E}_\mu\Big[\nabla_\theta \mathrm{Tr}\Big(f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)-f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta})\Big)^T \widehat{p}_\Theta(H,t+\Delta t/2)\Big]\right\| \\
\le& \sup_{i\in[N+1]} \mathbb{E}_\mu\Big[\left\|\nabla_\theta\Big(f(\widehat{T}_{\widetilde{\Theta}}(H,t),\theta)-f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta})\Big)_{:,i}\right\|\Big]\sqrt{N+1}\|\widehat{p}_\Theta(H,t+\Delta t/2)\|_F \\
\le&\phi_{PP}(r,B_T)\|\theta-\widetilde{\theta}\|\sqrt{N+1}\|\widehat{p}_\Theta(H,t+\Delta t/2)\|_F \\
\lesssim&\|\theta-\widetilde{\theta}\|.
\end{aligned}
$$

where the last inequality applies Lemma C.6. Therefore, we conclude that $\|\widehat{G}(\beta,\widetilde{\Theta},t)-\widehat{G}(\widetilde{\beta},\widetilde{\Theta},t)\| \le C_G(1+\lambda)\|\beta-\widetilde{\beta}\|$, completing the proof. $\square$

F.3 Proof of Lemma D.3

Proof. Lemmas C.1 and C.4 establish that $\max\|T_\rho(H,t)\|_{2-\mathrm{col}}, \|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}} \le B_T := B\exp(K(1+r+r^2))$. According to Lemma D.6, there exists an event E with $\mathbb{P}(E)\ge 1-\exp(-\delta)$ such that under E, we have

$$\|\widehat{T}_\Theta(H,t)-T_\rho(H,t)\|_F \lesssim L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}}. \tag{F.8}$$

for any H and $t=0,\Delta t,\dots,(L-1)\Delta t,1$. Following the same proof procedure as in Lemma D.6, with ρ replaced by $\hat{\rho}$, and bounding only J_1 and J_3 in the proof (as there is no need to utilize Hoeffding's inequality to bridge the difference due to a finite width M), we could obtain the bound

$$\|\widehat{T}_\Theta(H,t)-T_{\hat{\rho}}(H,t)\|_F \lesssim L^{-1}. \tag{F.9}$$

We present the proof only for the case involving ρ. The bounding of $\|\widehat{G}(\beta,\Theta,t)-G(\beta,\hat{\rho},t)\|_F$ can be derived analogously by substituting ρ with $\hat{\rho}$ using Equation (F.9), and skipping the process of bounding $\|D_1-D_2\|$ where D_1 and D_2 will be defined later. The bounding of $\|G(\beta,\hat{\rho},t)-G(\beta,\rho,t)\|$ can be straightforwardly achieved by combining the results obtained from the other two cases.

By the definitions of the gradients in Equations (C.2) and (C.3), we observe that $\left\|\widehat{G}(\beta,\Theta,t)-G(\beta,\rho,t)\right\| \lesssim \left\|\widehat{G}_f(\theta,\Theta,t)-G_f(\theta,\rho,t)\right\|+\left\|\widehat{G}_h(w,\Theta,t)-G_h(w,\rho,t)\right\|$. We will focus on showing that $\left\|2(\widehat{G}_f(\theta,\Theta,t)-G_f(\theta,\rho,t))\right\| \lesssim L^{-1}+\sqrt{\frac{\delta+\log(L+1)}{M}}$, as the other part of the proof follows a similar approach.

Let's define the following quantities A_1,A_2,B_1,B_2,C_1,C_2:

$$
\begin{aligned}
A_1 :=&\nabla_\theta \mathrm{vec}[f(\widehat{T}_\Theta(H,t),\theta)], \quad A_2 := \nabla_\theta \mathrm{vec}[f(T_\rho(H,t),\theta)] \\
B_1 :=&\mathrm{Read}[\widehat{T}_\Theta(H,1)-y(H)], \quad B_2 := \mathrm{Read}[T_\rho(H,1)-y(H)] \\
C_1 :=&\Big\{\prod_{\substack{(s-t)/\Delta t+2\in[(1-t)/\Delta t]\\ j\in[M]}}\Big(I_{\mathrm{dimvec}[T]}+(\Delta t/2)M^{-1}\sum_{j=1}^M \nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\Big) \\
&\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}}\Big(I_{\mathrm{dimvec}[T]}+(\Delta t/2)M^{-1}\sum_{j=1}^M \nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]\Big)\Big\}_{DN+d+1,:} \\
C_2 :=&\exp\Big(\int_t^1\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t)d\beta dt\Big)_{DN+d+1,:}
\end{aligned}
$$

The universal boundedness of $\|A_1\|$ and $\|A_2\|$ is implied by Assumption 2 (ii), while the universal boundedness of $|B_1|$ and $|B_2|$ is implied by Assumption 1. Additionally, $\|C_1\|$ and $\|C_2\|$ can be bounded via Assumption 2 (iii), and one can refer to the proofs of Lemmas C.3 and C.6 for detailed explanations.

From (C.14) and (C.16), we could rewrite $J := \left\|2(\widehat{G}_f(\theta, \Theta, t) - G_f(\theta, \rho, t))\right\|$ as

$$\begin{aligned} J = \|A_1B_1C_1 - A_2B_2C_2\| \leq & \|(A_1 - A_2)B_2C_2\| + \|A_1(B_1 - B_2)C_2\| + \|A_1B_1(C_1 - C_2)\| \\ \leq & \|A_1 - A_2\|\|B_2\|\|C_2\| + \|A_1\|\|B_1 - B_2\|\|C_2\| + \|A_1\|\|B_1\|\|C_1 - C_2\| \\ \lesssim & \|A_1 - A_2\| + \|B_1 - B_2\| + \|C_1 - C_2\|. \end{aligned}$$

We claim that to obtain the result, it suffices to show that

i. $\|A_1 - A_2\| \lesssim L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}$ under event E.

ii. $\|B_1 - B_2\| \lesssim L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}$ under event E.

iii. There exists some event E_2 with $\mathbb{P}(E_2) \geq 1 - \exp(-\delta)$ such that $\|C_1 - C_2\| \lesssim L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}$ under $E \cap E_2$.

This is because if we can establish the above statements, then under the event $E \cap E_2$ with $\mathbb{P}(E \cap E_2) \geq 1 - 2\exp(-\delta)$, we obtain $J = \left\|2(\widehat{G}_f(\theta, \Theta, t) - G_f(\theta, \rho, t))\right\| \lesssim L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}$. Given the similarity in proof for $\left\|2(\widehat{G}_h(w, \Theta, t) - G_h(w, \rho, t))\right\|$, we deduce that with probability at least $1 - 4\exp(-\delta)$, we have $\left\|\widehat{G}(\beta, \Theta, t) - G(\beta, \rho, t)\right\| \lesssim L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}$. The remainder of the proof focuses on bounding the quantities in statements (i)-(iii).

Proof for statement (i): By Assumption 3 (iv), we have $\|A_1 - A_2\| \leq \phi_{TT}(N, D, \sqrt{N+1}B_T, r)\|\widehat{T}_\Theta(H, t) - T_\rho(H, t)\|_F \lesssim L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}$ under event E.

Proof for statement (ii): Under event E, it is obvious to see $\|B_1 - B_2\| \leq \|\widehat{T}_\Theta(H, t) - T_\rho(H, t)\|_F \lesssim L^{-1} + \sqrt{\frac{\delta+\log(L+1)}{M}}$.

Proof for statement (iii): We further define the following quantities

$$\begin{aligned} D_1 := & \prod_{\substack{(s-t)/\Delta t+2 \in [(1-t)/\Delta t] \\ j \in [M]}} \Big(I_{\text{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H, s), \theta_{s,j})]\Big) \\ & \prod_{\substack{(s-t)/\Delta t+1 \in [(1-t)/\Delta t] \\ j \in [M]}} \Big(I_{\text{dimvec}[T]} + (\Delta t/2)M^{-1}\sum_{j=1}^{M} \nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H, s+\Delta t/2), w_{s,j})]\Big) \\ D_2 := & \prod_{(s-t)/\Delta t+2 \in [(1-t)/\Delta t]} \Big(I_{\text{dimvec}[T]} + (\Delta t/2)\int_\beta \nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H, s), \theta)]\rho(\beta|s)d\beta\Big) \\ & \prod_{(s-t)/\Delta t+1 \in [(1-t)/\Delta t]} \Big(I_{\text{dimvec}[T]} + (\Delta t/2)\int_\beta \nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H, s+\Delta t/2), w)]\rho(\beta|s)d\beta\Big) \\ D_3 := & \exp\Big(\sum_{(s-t)/\Delta t+2 \in [(1-t)/\Delta t]} (\Delta t/2)\int_\beta \nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H, s), \theta)]\rho(\beta, s)d\beta \\ & + \sum_{(s-t)/\Delta t+1 \in [(1-t)/\Delta t]} (\Delta t/2)\int_\beta \nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H, s+\Delta t/2), w)]\rho(\beta, s)d\beta\Big) \\ D_4 := & \exp\Big(\int_t^1 \int_\beta \nabla_{\text{vec}[T]}\text{vec}[g(T_\rho(H, t), \beta)]\rho(\beta, t)d\beta dt\Big) \end{aligned}$$

Note that $\|C_1 - C_2\| \leq \|D_1 - D_2\| + \|D_2 - D_3\| + \|D_3 - D_4\|$. Assumption 2 (iii) indicates that for any $s = 0, \Delta t, \ldots, (L-1)\Delta$ and $j = 1, \ldots, M$, we have

$$\max\{\|\nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\|, \|\nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]\|\} \leq B_J$$

for some universal constant B_J. This implies that each column of $\nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]$ or $\nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]$ has l_2 norm upper bounded by B_J as well. Applying Hoeffding's inequality to each column of $\nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]$ and $\nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]$, and subsequently calculating the union bound across all columns yields:

$$\begin{aligned}&\mathbb{P}\Big(\Big\|M^{-1}\sum_{j=1}^{M}\nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})] - \int_\beta \nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\rho(\beta|s)d\beta\Big\| \geq \sqrt{(N+1)D}z\Big)\\ \leq& 2(N+1)D\exp(-\frac{z^2}{2B_J^2}M)\end{aligned} \tag{F.10}$$

and

$$\begin{aligned}&\mathbb{P}\Big(\Big\|M^{-1}\sum_{j=1}^{M}\nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,t),w_{s,j})] - \int_\beta \nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,t),w_{s,j})]\rho(\beta|s)d\beta\Big\| \geq \sqrt{(N+1)D}z\Big)\\ \leq& 2(N+1)D\exp(-\frac{z^2}{2B_J^2}M)\end{aligned} \tag{F.11}$$

for any $z > 0$. For (F.10) and (F.11), we further the consider the union bound across all $s = 0, \Delta t, \ldots, (L-1)\Delta$, and let $z = B_J\sqrt{2M(\delta + \log(2(N+1)DL))}$, which implies that with probability at least $1 - \exp(-\delta)$, we have

$$\|M^{-1}\sum_{j=1}^{M}\nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})] - \int_\beta \nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\rho(\beta|s)d\beta\|$$

and

$$\|M^{-1}\sum_{j=1}^{M}\nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,t),w_{s,j})] - \int_\beta \nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,t),w_{s,j})]\rho(\beta|s)d\beta\|$$

bounded by $B_J\sqrt{2M(N+1)D(\delta + \log(2(N+1)DL))} \leq C_J\sqrt{\frac{\delta+\log(L+1)}{M}}$ for any $s = 0, \Delta t, \ldots, (L-1)\Delta$. Here, C_J is some constant that only depends on N, D, r and the parameters of the assumptions. Denote this probability event by E_2, and we have $\mathbb{P}(E_2) \geq 1 - \exp(-\delta)$. Under E_2, by Lemma C.8, we have

$$\begin{aligned}&\|D_1 - D_2\|\\ \leq&\frac{1}{2L}\Big(\sum_{(s-t)/\Delta t+2\in[(1-t)/\Delta t]}\Big\|M^{-1}\sum_{j=1}^{M}\nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})] - \int_\beta \nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta)]\rho(\beta,s)d\beta\Big\|\\ &+\sum_{(s-t)/\Delta t+1\in[(1-t)/\Delta t]}\Big\|M^{-1}\sum_{j=1}^{M}\nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,s),w_{s,j})] - \int_\beta \nabla_{\text{vec}[T]}\text{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w)]\rho(\beta,s)d\beta\Big\|\Big)\\ \leq& C_J\sqrt{\frac{\delta+\log(L+1)}{M}}.\end{aligned} \tag{F.12}$$

For any $s = 0, \Delta t, \ldots, (L-1)\Delta$, we define

$$A_{s,j} = (\Delta t/2)\int_\beta \nabla_{\text{vec}[T]}\text{vec}[f(\widehat{T}_\Theta(H,s),\theta)]\rho(\beta|s)d\beta$$

and

$$B_{s,j} = (\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(\widehat{T}_\Theta(H, s + \Delta t/2), w)] \rho(\beta|s) d\beta.$$

Since Assumption 2 (iii) indicates that $\max\{\|A_{s,j}\|, \|B_{s,j}\|\} \lesssim \Delta t = L^{-1}$, we have

$$\|\exp(A_{s,j}) - I - A_{s,j}\| \lesssim \|A_{s,j}\|^2, \quad \|\exp(B_{s,j}) - I - B_{s,j}\| \lesssim \|B_{s,j}\|^2.$$

Applying Lemma C.8 once more, we have

$$\|D_2 - D_3\| \leq \sum_{(s-t)/\Delta t + 2 \in [(1-t)/\Delta t]} \|A_{s,j}\|^2 + \sum_{(s-t)/\Delta t + 1 \in [(1-t)/\Delta t]} \|B_{s,j}\|^2 \lesssim L^{-1}. \tag{F.13}$$

Since Assumption 2 (iii) ensures the boundedness of $\|D_4\|$, we have

$$\begin{aligned}
\|D_3 - D_4\| \lesssim \Big\| \exp\Big(& \sum_{(s-t)/\Delta t + 2 \in [(1-t)/\Delta t]} (\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(\widehat{T}_\Theta(H, s), \theta)] \rho(\beta, s) d\beta \\
& + \sum_{(s-t)/\Delta t + 1 \in [(1-t)/\Delta t]} (\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(\widehat{T}_\Theta(H, s + \Delta t/2), w)] \rho(\beta, s) d\beta \\
& - \int_t^1 \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H, t), \beta)] \rho(\beta, t) d\beta dt \Big) - 1 \Big\|.
\end{aligned} \tag{F.14}$$

Therefore, to show that $\|D_3 - D_4\| \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}$, it suffices to show that

$$\begin{aligned}
J_{34} := \Big\| & \sum_{(s-t)/\Delta t + 2 \in [(1-t)/\Delta t]} (\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(\widehat{T}_\Theta(H, s), \theta)] \rho(\beta, s) d\beta \\
& + \sum_{(s-t)/\Delta t + 1 \in [(1-t)/\Delta t]} (\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(\widehat{T}_\Theta(H, s + \Delta t/2), w)] \rho(\beta, s) d\beta \\
& - \int_t^1 \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H, t), \beta)] \rho(\beta, t) d\beta dt \Big\| \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}.
\end{aligned}$$

By Assumption 3 (iv) and Lemma D.6, we have

$$\Big\| \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(\widehat{T}_\Theta(H, s), \theta)] - \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(T_\rho(H, s), \theta)] \Big\| \lesssim \|\widehat{T}_\Theta(H, s) - T_\rho(H, s)\|_F \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}},$$

and

$$\begin{aligned}
& \Big\| \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(\widehat{T}_\Theta(H, s + \Delta t/2), w)] - \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(T_\rho(H, s), w)] \Big\| \lesssim \|\widehat{T}_\Theta(H, s + \Delta t/2) - T_\rho(H, s)\|_F \\
\lesssim & \|\widehat{T}_\Theta(H, s + \Delta t/2) - \widehat{T}_\Theta(H, s)\|_F + \|\widehat{T}_\Theta(H, s) - T_\rho(H, s)\|_F \\
\lesssim & L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}} + L^{-1} \|M^{-1} \sum_{j=1}^M h(\widehat{T}_\Theta(H, s), w_{s,j})\|_F \\
\lesssim & L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}},
\end{aligned}$$

where the last inequality employs Assumption 2 (i). Therefore, we conclude that

$$
\begin{aligned}
\|D_3 - D_4\| \lesssim J_{34} \lesssim & L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}} + \|(\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(T_\rho(H,t),\theta)]\rho(\beta,t) d\beta\| \\
& \Big\| \sum_{(s-t)/\Delta t + 1 \in [(1-t)/\Delta t]} (\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[f(T_\rho(H,s),\theta)]\rho(\beta,s) d\beta \\
& + \sum_{(s-t)/\Delta t + 1 \in [(1-t)/\Delta t]} (\Delta t/2) \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[h(T_\rho(H,s)(H,s),w)]\rho(\beta,s) d\beta \\
& - \int_t^1 \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t) d\beta dt \Big\| \\
\lesssim & L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}} + \sup_{|s_1 - s_2| \le \Delta t} \big\| \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s_1),\beta)] - \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s_2),\beta)] \big\| \\
\lesssim & L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}} + \sup_{|s_1 - s_2| \le \Delta t} \|T_\rho(H,s_1) - T_\rho(H,s_2)\|_F \\
\lesssim & L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}
\end{aligned}
\tag{F.15}
$$

where the third inequality uses Assumption 3 (iv), and the last inequality relies on the Lipschitz continuity as demonstrated in Proposition C.1. Combining (F.12), (F.13), and (F.15) yields $\|C_1 - C_2\| \le \|D_1 - D_2\| + \|D_2 - D_3\| + \|D_3 - D_4\| \lesssim L^{-1} + \sqrt{\frac{\delta + \log(L+1)}{M}}$. □

F.4 Proof of Lemma D.4

Proof. Define $F_\rho(T_{\bar\rho}, t) = \int_\beta \mathrm{vec}[g(T_{\bar\rho}(H,t),\beta)]\rho(\beta,t) d\beta$ for any $\rho, \bar\rho \in \mathcal{P}^2$. From Taylor's expansion, we have

$$
\begin{aligned}
\mathrm{vec}[\dot T_{\rho_\eta}(H,t) - \dot T_\rho(H,t)] &= F_\rho(T_{\rho_\eta}, t) - F_\rho(T_\rho, t) + F_{\rho_\eta}(T_{\rho_\eta}, t) - F_\rho(T_{\rho_\eta}, t) \\
&= \Big(\int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t) d\beta \Big) \Big(\mathrm{vec}[T_{\rho_\eta}(H,t) - T_\rho(H,t)] \Big) \\
&+ \eta \int_\beta \mathrm{vec}[g(T_{\rho_\eta}(H,t),\beta)](\nu - \rho)(\beta,t) d\beta + o(\eta) \\
&= \Big(\int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t) d\beta \Big) \Big(\mathrm{vec}[T_{\rho_\eta}(H,t) - T_\rho(H,t)] \Big) \\
&+ \eta \int_\beta \mathrm{vec}[g(T_\rho(H,t),\beta)](\nu - \rho)(\beta,t) d\beta \\
&+ \eta \Big(\int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,t),\beta)](\nu - \rho)(\beta,t) d\beta \Big) \Big(\mathrm{vec}[T_{\rho_\eta}(H,t) - T_\rho(H,t)] \Big) + o(\eta) \\
&= \Big(\int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t) d\beta \Big) \Big(\mathrm{vec}[T_{\rho_\eta}(H,t) - T_\rho(H,t)] \Big) \\
&+ \eta \int_\beta \mathrm{vec}[g(T_\rho(H,t),\beta)](\nu - \rho)(\beta,t) d\beta + o(\eta),
\end{aligned}
$$

of which the last equality holds as Lemma C.2 shows that $\|T_{\rho_\eta}(H,t) - T_\rho(H,t)\|_F = O(W_2(\rho_\eta, \rho)) = O(\eta)$, where we hide the constant dependence on B, K, N, r. Therefore, we have

$$\frac{d}{dt}\Big\{ \exp\Big(-\int_0^t \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta\Big)\Big(\mathrm{vec}[T_{\rho_\eta}(H,t)-T_\rho(H,t)]\Big)\Big\}$$

$$= \exp\Big(-\int_0^t \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta\Big)$$

$$\Big\{\mathrm{vec}[\dot{T}_{\rho_\eta}(H,t)-\dot{T}_\rho(H,t)] - \Big(\int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t)d\beta\Big)\Big(\mathrm{vec}[T_{\rho_\eta}(H,t)-T_\rho(H,t)]\Big)\Big\}$$

$$= \exp\Big(-\int_0^t \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta\Big)\Big\{\eta\int_\beta \mathrm{vec}[g(T_\rho(H,t),\beta)](\nu-\rho)(\beta,t)d\beta + o(\eta)\Big\}, \tag{F.16}$$

which leads to (D.18). □

F.5 Proof of Lemma D.5

Proof. Fix $(\beta,t)\in P_r$. Lemma C.2 implies that
$\|p_\rho(H,1)-p_\nu(H,1)\|_F = |\mathrm{Read}(T_\rho(H,1)-\mathrm{Read}(T_\nu(H,1)|\le C_r W_1(\rho,\nu)\le C_r(r+1)\|\rho-\nu\|_1$
for some constant C_r. Our goal is to regulate the difference between $\dot{p}_\rho(H,t)$ and $\dot{p}_\nu(H,t)$ to control the the propagation of $\|p_\rho(H,\cdot)-p_\nu(H,\cdot)\|$. Note that by (C.14) and Assumption 2,

$$\begin{aligned}
\frac{d}{dt}\|p_\rho(H,t)-p_\nu(H,t)\|_F \le& \|\dot{p}_\rho(H,t)-\dot{p}_\nu(H,t)\|_F\\
=&\Big\|\mathrm{vec}[p_\rho(H,t)]^T\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t)d\beta\\
&-\mathrm{vec}[p_\nu(H,t)]^T\int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,t),\beta)]\nu(\beta,t)d\beta\Big\|\\
\le&\|p_\rho(H,t)-p_\nu(H,t)\|_F\int_\beta \big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,t),\beta)]\big\|\rho(\beta,t)d\beta\\
&+\|p_\nu(H,t)\|_F\int_\beta \big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\nu(H,t),\beta)]\big\|(\rho-\nu)(\beta,t)d\beta\\
\le&L_{r,1}\Big(\|p_\rho(H,t)-p_\nu(H,t)\|_F\int_\beta\rho(\beta,t)d\beta+\|p_\nu(H,t)\|_F\|\rho(\cdot,t)-\nu(\cdot,t)\|_1\Big)\\
\le&L_{r,1}\Big(\|p_\rho(H,t)-p_\nu(H,t)\|_F\int_\beta\rho(\beta,t)d\beta+L_{r,2}\|\rho(\cdot,t)-\nu(\cdot,t)\|_1\Big)
\end{aligned}\tag{F.17}$$

where

$$L_{r,1}=\phi_T(N,D,\sqrt{N+1}B\exp(K(1+r+r^2)))(1+r+r^2)$$

and

$$L_{r,2}=(B+B\exp(K(1+r+r^2)))\exp\Big(\phi_T(N,D,\sqrt{N+1}KB\exp(K(1+r+r^2))(1+r+r^2)\Big).$$

The third inequality of (F.17) uses Lemma C.1 to obtain

$$\|T_\rho(H,t)\|_F\le\sqrt{N+1}\|T_\rho(H,t)\|_{2-\mathrm{col}}\le\sqrt{N+1}B\exp(K(1+r+r^2))$$

and

$$\|T_\nu(H,t)\|_F\le\sqrt{N+1}\|T_\nu(H,t)\|_{2-\mathrm{col}}\le\sqrt{N+1}B\exp(K(1+r+r^2))$$

with Assumption 2(iii) to bound the norm of the Jacobian matrix with $L_{r,1}$. The last inequality of (F.17) employs Lemma C.3 to bound $\|p_\nu(H,t)\|_F$ with $L_{r,2}$. Applying the Grönwall's inequality to (F.17), we obtain

$$\begin{aligned}
\|p_\rho(H,t)-p_\nu(H,t)\|_F\le& C_r(r+1)\exp(L_{r,1})\|\rho-\nu\|_1+\int_0^1 L_{r,1}L_{r,2}\exp(L_{r,1}\int_t^1\int_\beta\rho(\beta,t)d\beta ds)\|\rho(\cdot,t)-\nu(\cdot,t)\|_1dt\\
\le& C_r(r+1)\exp(L_{r,1})\|\rho-\nu\|_1+L_{r,1}L_{r,2}\exp(L_{r,1})\int_0^1\|\rho(\cdot,t)-\nu(\cdot,t)\|_1dt\\
=&(C_r+L_{r,1}L_{r,2})\exp(L_{r,1}))\|\rho-\nu\|_1
\end{aligned}\tag{F.18}$$

Since

$$\int_0^1 \|\rho(\cdot,t)-\nu(\cdot,t)\|_1 dt = \int_0^1 \int_\beta |\rho(\beta,t)-\nu(\beta,t)| d\beta dt = \|\rho-\nu\|_1.$$

Thus, we complete the proof of the first result.

By Lemma C.3, under Assumption 1 we have

$$\|p_\rho(H,t)\|_F \le L_{r,3} := (B{+}B\exp(K(1{+}r{+}r^2)))\exp\Big(\phi_T(N,D,\sqrt{N+1}KB\exp(K(1{+}r{+}r^2))(1{+}r{+}r^2)\Big).$$

In addition, by Lemma C.1, under Assumption 2 (i) we have

$$\|g(T_\nu(H,t),\beta)\|_F \le \sqrt{N+1}\|g(T_\nu(H,t),\beta)\|_{2-\mathrm{col}} \le L_{r,4} := KB\exp(K(1{+}r{+}r^2))(1{+}r{+}r^2).$$

Therefore, for the gradient function $\frac{\delta Q}{\delta \rho}$, by Lemma C.3 we have

$$\begin{aligned}
\Big|\frac{\delta Q}{\delta \rho}\Big|_\rho(\beta,t) - \frac{\delta Q}{\delta \rho}\Big|_\nu(\beta,t)\Big| =& \mathbb{E}_\mu[\mathrm{Tr}(g(T_\rho(H,t),\beta)^T p_\rho(H,t)) - \mathrm{Tr}(g(T_\nu(H,t),\beta)^T p_\nu(H,t))] \\
&\le \mathbb{E}_\mu[\|g(T_\rho(H,t),\beta)-g(T_\nu(H,t),\beta)\|_F\|p_\rho(H,t)\|_F] \\
&+ \mathbb{E}_\mu[\|g(T_\nu(H,t),\beta)\|_F\|p_\rho(H,t)-p_\nu(H,t)\|_F] \\
&\le L_{r,3}\mathbb{E}_\mu[\|g(T_\rho(H,t),\beta)-g(T_\nu(H,t),\beta)\|_F] + L_{r,4}(C_r+L_{r,1}L_{r,2})\|\rho-\nu\|_1 \\
&\le \sqrt{N+1}L_{r,3}\mathbb{E}_\mu[\|g(T_\rho(H,t),\beta)-g(T_\nu(H,t),\beta)\|_{2-\mathrm{col}}] + L_{r,4}(C_r+L_{r,1}L_{r,2})\|\rho-\nu\|_1.
\end{aligned} \tag{F.19}$$

Hence, it suffices to show that for any H such that $\|H\|_{2-\mathrm{col}} \le B$, we have $\|g(T_\rho(H,t),\beta) - g(T_\nu(H,t),\beta)\|_{2-\mathrm{col}} \le L_{r,5}\|\rho-\nu\|_1$ for some $L_{r,5} > 0$ in order to obtain the second result of this lemma. By Assumption 2 (iii), we see that

$$\begin{aligned}
\|g(T_\rho(H,t),\beta)-g(T_\nu(H,t),\beta)\|_{2-\mathrm{col}} \le& \phi_T(N,D,\max\{\|T_\rho(H,t)\|_F,\|T_\nu(H,t)\|_F\})(1+r+r^2)\|T_\rho(H,t)-T_\rho(H,t)\|_2 \\
\le& \phi_T(N,D,\sqrt{N+1}KB\exp(K(1+r+r^2)))(1+r+r^2)C_rW_1(\rho,\nu) \\
\le& \phi_T(N,D,\sqrt{N+1}KB\exp(K(1+r+r^2)))(1+r+r^2)C_r(1+r)\|\rho-\nu\|_1,
\end{aligned} \tag{F.20}$$

where the second inequality again uses Lemma (C.2). Combining (F.19) and F.20 completes the proof of the second result. □

F.6 Proof of Lemma D.6

Proof. Denote the empirical distribution of $\bar{T}_\Theta$ and $\widetilde{T}_\rho$ by

$$\bar{\rho} = \frac{1}{ML}\sum_t \sum_{j=1}^M \delta(\beta_{t,j},t), \quad \widetilde{\rho} = \frac{1}{L}\sum_t \delta_t(t)\rho(\beta|t),$$

respectively. It's straightforward to verify that $\bar{\rho}$ and $\widetilde{\rho}$ meet the conditions outlined in Lemma C.1, and $T_{\bar{\rho}} = \bar{T}_\Theta$ and $T_{\widetilde{\rho}} = \widetilde{T}_\rho$. Hence, Lemmas C.1 and C.4 indicate that $\max\{\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}}, \|\bar{T}_\Theta(H,t)\|_{2-\mathrm{col}}, \|\widetilde{T}_\rho(H,t)\|_{2-\mathrm{col}}\} \le B\exp(K(1+r+r^2))$ for any H and $t \in [0,1]$. We then define $B_T := B\exp(K(1+r+r^2))$.

The following decomposition equation holds:

$$\|\widehat{T}_\Theta(H,t){-}T_\rho(H,t)\|_F \le \underbrace{\|\widehat{T}_\Theta(H,t)-\bar{T}_\Theta(H,t)\|_F}_{J_1} + \underbrace{\|\bar{T}_\Theta(H,t)-\widetilde{T}_\rho(H,t)\|_F}_{J_2} + \underbrace{\|\widetilde{T}_\rho(H,t)-T_\rho(H,t)\|_F}_{J_3}, \tag{F.21}$$

Our proof will bound J_1, J_2 and J_3, possibly in a probabilistic manner, to obtain the desired result.

Bounding J_1: Note that according to Assumption 2 (i),

$$
\begin{aligned}
&\|\widehat{T}_\Theta(H,t)+(\Delta t/2)M^{-1}\sum_{j=1}^{M} f(\widehat{T}_\Theta(H,t),\theta_{t,j})\|_{2-\mathrm{col}}\\
\le&\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}}+(\Delta t/2)M^{-1}\sum_{j=1}^{M}\|f(\widehat{T}_\Theta(H,t),\theta_{t,j})\|_{2-\mathrm{col}}\\
\le&\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}}+(K\Delta t/2)M^{-1}\sum_{j=1}^{M}\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}}(1+\|\theta_{t,j}\|+\|\theta_{t,j}\|^2)\\
\le&B_T(1+(K\Delta t/2)(1+r+r^2))\\
\le&B_T(1+(K/2)(1+r+r^2))
\end{aligned}
$$

Denote $B_T(1+(K/2)(1+r+r^2))$ by $\bar{B}_T$. Combining the two equations in (2.4) gives us

$$
\begin{aligned}
&\widehat{T}_\Theta(H,t+\Delta t)=\mathrm{MLP}_{w_{t,1},\ldots,w_{t,M}}\Big(\mathrm{Attn}_{\theta_{t,1},\ldots,\theta_{t,M}}(\widehat{T}_\Theta(H,t),\Delta t/2),\Delta t/2\Big)\\
=&\mathrm{Attn}_{\theta_{t,1},\ldots,\theta_{t,M}}(\widehat{T}_\Theta(H,t),\Delta t/2)+(\Delta t/2)M^{-1}\sum_{j=1}^{M} h\Big(\mathrm{Attn}_{\theta_{t,1},\ldots,\theta_{t,M}}(\widehat{T}_\Theta(H,t),\Delta t/2),w_{t,j}\Big)\\
=&\widehat{T}_\Theta(H,t)+(\Delta t/2)M^{-1}\sum_{j=1}^{M} f(\widehat{T}_\Theta(H,t),\theta_{t,j})+(\Delta t/2)M^{-1}\sum_{j=1}^{M} h\Big(\widehat{T}_\Theta(H,t)\\
+&(\Delta t/2)M^{-1}\sum_{j=1}^{M} f(\widehat{T}_\Theta(H,t),\theta_{t,j}),w_{t,j}\Big)
\end{aligned}
\tag{F.22}
$$

Then, from the formula of (F.22) and Assumption 2 (iii), we see that for any $t=0,\Delta t,\ldots,(L-1)\Delta$,

$$
\begin{aligned}
&\|\widehat{T}_\Theta(H,t+\Delta t)-\bar{T}_\Theta(H,t+\Delta t)\|_F\\
\le&\|\widehat{T}_\Theta(H,t)-\bar{T}_\Theta(H,t)\|_F\Big(1+(\Delta t/2)M^{-1}\sum_{j=1}^{M}\phi_T(N,D,\sqrt{N+1}B_T)(1+\|\theta_{t,j}\|+\|\theta_{t,j}\|^2)\Big)\\
+&(\Delta t/2)M^{-1}\sum_{j=1}^{M}\phi_T(N,D,\sqrt{N+1}\bar{B}_T)(1+\|w_{t,j}\|+\|w_{t,j}\|^2)\\
&\|\widehat{T}_\Theta(H,t)+(\Delta t/2)M^{-1}\sum_{j=1}^{M} f(\widehat{T}_\Theta(H,t),\theta_{t,j})-\bar{T}_\Theta(H,t)\|_F\\
\le&\|\widehat{T}_\Theta(H,t)-\bar{T}_\Theta(H,t)\|_F\Big(1+\Delta tM^{-1}\sum_{j=1}^{M}\phi_T(N,D,\sqrt{N+1}\bar{B}_T)(1+\|\beta_{t,j}\|+\|\beta_{t,j}\|^2)\Big)\\
+&\sqrt{N+1}B_T(K\Delta t/2)(1+r+r^2)(\Delta t/2)M^{-1}\sum_{j=1}^{M}\phi_T(N,D,\sqrt{N+1}\bar{B}_T)(1+\|w_{t,j}\|+\|w_{t,j}\|^2)\\
\le&\|\widehat{T}_\Theta(H,t)-\bar{T}_\Theta(H,t)\|_F\Big(1+\Delta t\phi_T(N,D,\sqrt{N+1}\bar{B}_T)(1+r+r^2)\Big)\\
+&\sqrt{N+1}\phi_T(N,D,\sqrt{N+1}\bar{B}_T)B_T(K\Delta t^2/4)(1+r+r^2)^2
\end{aligned}
\tag{F.23}
$$

Therefore, applying (F.23), we deduce that for any $t=0,\Delta t,\ldots,(L-1)\Delta t,1$, we have:

$$
\|\widehat{T}_\Theta(H,t)-\bar{T}_\Theta(H,t)\|_{2-\mathrm{col}}\le\sqrt{N+1}B_T(K\Delta t^2/4)(1+r+r^2)\frac{\exp(\phi_T(N,D,\sqrt{N+1}\bar{B}_T)(1+r+r^2))}{\Delta t}C_1L^{-1}
\tag{F.24}
$$

where $C_1:=\sqrt{N+1}B_TK(1+r+r^2)\exp(\phi_T(N,D,\sqrt{N+1}\bar{B}_T)(1+r+r^2))/4$.

Bounding J_2**:** For any $t = 0, \Delta t, \dots, (L-1)\Delta, i \in [N+1], j \in [M]$, we have $\|g(\bar{T}_\Theta(H,t), \beta_{t,j})_{:,i}\| \leq B_T$. Hence, by the Hoeffding's inequality, for any $z > 0$ we have

$$\mathbb{P}(\|M^{-1}\sum_{j=1}^{M} g(\bar{T}_\Theta(H,t), \beta_{t,j})_{:,i} - \int_\beta g(\bar{T}_\Theta(H,t), \beta_{t,j})_{:,i}\rho(\beta|t)d\beta\| \geq z) \leq 2\exp(-\frac{z^2}{2B_T^2}M).$$

By the union bound over $i \in [N+1]$ and $t = 0, \Delta t, \dots, (L-1)\Delta$, the above inequality implies

$$\mathbb{P}(\sup_t\|M^{-1}\sum_{j=1}^{M} g(\bar{T}_\Theta(H,t), \beta_{t,j}) - \int_\beta g(\bar{T}_\Theta(H,t), \beta_{t,j})\rho(\beta|t)d\beta\|_{2-\text{col}} \geq z) \leq 2(N+1)L\exp(-\frac{z^2}{2B_T^2}M). \tag{F.25}$$

We let $z = B_T\sqrt{2M^{-1}(\delta + \log((N+1)L))}$. Then, (F.25) turns into

$$\mathbb{P}(\sup_t\|M^{-1}\sum_{j=1}^{M} g(\bar{T}_\Theta(H,t), \beta_{t,j}) - \int_\beta g(\bar{T}_\Theta(H,t), \beta_{t,j})\rho(\beta|t)d\beta\|_{2-\text{col}} \geq B_T\sqrt{2M^{-1}(\delta + \log((N+1)L))} \leq \exp(-\delta). \tag{F.26}$$

Denote the event such that

$$\sup_t\|M^{-1}\sum_{j=1}^{M} g(\bar{T}_\Theta(H,t), \beta_{t,j}) - \int_\beta g(\bar{T}_\Theta(H,t), \beta_{t,j})\rho(\beta|t)d\beta\|_{2-\text{col}} \leq B_T\sqrt{2M^{-1}(\delta + \log((N+1)L))}$$

by E_δ. (F.26) directly indicates $\mathbb{P}(E_\delta) \geq 1 - \exp(-\delta)$.

Suppose that the high probability event E_δ occurs. Let's denote $B_T\sqrt{2M^{-1}(\delta + \log((N+1)L))}$ by B_δ for brevity. From Assumption 2 (iii), it follows that for any $t = 0, \Delta t, \dots, (L-1)\Delta$,

$$\begin{aligned}
\|\bar{T}_\Theta(H, t+\Delta t) - \widetilde{T}_\rho(H, t+\Delta t)\|_F \leq& \|\bar{T}_\Theta(H,t) - \widetilde{T}_\rho(H,t)\|_F \\
&+ \Delta t\|M^{-1}\sum_{j=1}^{M} g(\bar{T}_\Theta(H,t), \beta_{t,j}) - \int_\beta g(\bar{T}_\Theta(H,t), \beta_{t,j})\rho(\beta|t)d\beta\|_F \\
&+ \Delta t\|\int_\beta (g(\bar{T}_\Theta(H,t), \beta_{t,j}) - g(\widetilde{T}_\rho(H,t), \beta_{t,j}))\rho(\beta|t)d\beta\|_F \\
\leq& \|\bar{T}_\Theta(H,t) - \widetilde{T}_\rho(H,t)\|_F + \Delta t\sqrt{N+1}B_\delta + \\
&+ \Delta t\int_\beta \|g(\bar{T}_\Theta(H,t), \beta_{t,j}) - g(\widetilde{T}_\rho(H,t), \beta_{t,j})\|_F\rho(\beta|t)d\beta \\
\leq& \|\bar{T}_\Theta(H,t) - \widetilde{T}_\rho(H,t)\|_F(1 + \Delta t\phi_T(N, D, \sqrt{N+1}B_T)(1 + r + r^2)) \\
&+ \Delta t\sqrt{N+1}B_\delta.
\end{aligned} \tag{F.27}$$

Repeatedly applying Equation (F.27) yields

$$\begin{aligned}
\|\widehat{T}_\Theta(H,t) - \bar{T}_\Theta(H,t)\|_F \leq& \frac{\sqrt{N+1}B_\delta}{\phi_T(N, D, \sqrt{N+1}B_T)(1 + r + r^2)}\exp(\phi_T(N, D, \sqrt{N+1}B_T)(1 + r + r^2)) \\
=& C_2\sqrt{M^{-1}(\delta + \log((N+1)L))}.
\end{aligned} \tag{F.28}$$

for some universal constant $C_2 > 0$.

Bounding J_3**:** It's worth noting that the convergence proof with a convergence rate of $O(\Delta t) = O(\frac{1}{L})$ for $\widetilde{T}_\rho(H,t)$, the first-order Euler method for $T_\rho(H,t)$, is non-standard. This departure from convention arises because we do not assume the boundedness of the second-order derivative $\frac{d^2T_\rho(H,t)}{dt^2}$, instead relying on the continuity of $\rho(\cdot, t)$ with respect to the depth index t. In this proof, $O(\cdot)$ hides dependencies on N, D, r, and the parameters of the assumptions.

From the definition of $\widetilde{T}_\rho$ and T_ρ, we have

$$
\begin{aligned}
\|\widetilde{T}_\rho(H,t+\Delta t)-T_\rho(H,t+\Delta t)\|_F \leq& \|\widetilde{T}_\rho(H,t)-T_\rho(H,t)\|_F \\
&+\underbrace{\|T_\rho(H,t+\Delta t)-T_\rho(H,t)-\Delta t\int_\beta g(T_\rho(H,t),\beta)\rho(\beta|t)d\beta\|_F}_{I_1} \\
&+\Delta t\underbrace{\|\int_\beta (g(T_\rho(H,t),\beta)-g(\widetilde{T}_\rho(H,t),\beta))\rho(\beta|t)d\beta\|_F}_{I_2}.
\end{aligned}
\tag{F.29}
$$

To bound I_1, we use (3.1) to get

$$
\begin{aligned}
I_1 \leq& \|\int_t^{t+\Delta t}\int_\beta g(T_\rho(H,s),\beta)\rho(\beta|s)d\beta ds-\Delta t\int_\beta g(T_\rho(H,t),\beta)\rho(\beta|t)d\beta\|_F \\
\leq& \Delta t \sup_{s\in[t,t+\Delta t]}\|\int_\beta g(T_\rho(H,s),\beta)\rho(\beta|s)d\beta-\int_\beta g(T_\rho(H,t),\beta)\rho(\beta|t)d\beta\|_F \\
\leq& \Delta t \sup_{s\in[t,t+\Delta t]}\|\int_\beta (g(T_\rho(H,s),\beta)-g(T_\rho(H,t),\beta))\rho(\beta|s)d\beta\|_F \\
&+\Delta t \sup_{s\in[t,t+\Delta t]}\|\int_\beta g(T_\rho(H,t),\beta)(\rho(\beta|s)-\rho(\beta|t))d\beta\|_F \\
\leq& \Delta t \sup_{s\in[t,t+\Delta t]}\int_\beta \|g(T_\rho(H,s),\beta)-g(T_\rho(H,t),\beta)\|_F\rho(\beta|s)d\beta \\
&+\Delta t \sup_{s\in[t,t+\Delta t]}\|\int_\beta g(T_\rho(H,t),\beta)(\rho(\beta|s)-\rho(\beta|t))d\beta\|_F
\end{aligned}
\tag{F.30}
$$

Given that Proposition C.1 establishes the Lipschitz continuity of $T_\rho(H,t)$ with respect to t under the condition that $\rho\in\mathcal{P}^2$ has a bounded support, we can conclude:

$$
\sup_{s\in[t,t+\Delta t]}\|g(T_\rho(H,s),\beta)-g(T_\rho(H,t),\beta)\|_F \leq C_{3,1}|t-s|\leq C_{3,1}\Delta t \tag{F.31}
$$

for some constant $C_{3,1}>0$. Furthermore, Lemma C.1 and Proposition C.1 demonstrate that $g(T_\rho(H,t),\beta)$ is both bounded and Lipschitz continuous with respect to β. Thus, we have

$$
\begin{aligned}
\sup_{s\in[t,t+\Delta t]}\|\int_\beta g(T_\rho(H,t),\beta)(\rho(\beta|s)-\rho(\beta|t))d\beta\|_F =& \sup_{s\in[t,t+\Delta t]}\|\int_\beta g(T_\rho(H,t),\beta)(\rho(\beta,s)-\rho(\beta,t))d\beta\|_F \\
\leq& \sup_{s\in[t,t+\Delta t]} C_{3,2}\|\rho(\cdot,s)-\rho(\cdot,t)\|_{\mathrm{BL}} \\
\leq& C_{3,2}C_\rho\Delta t
\end{aligned}
\tag{F.32}
$$

for some constant $C_{3,2}>0$. Substituting (F.31) and (F.32) into (F.30), we find that there exists a constant $C_{3,5}$ such that $I_1\leq C_{3,5}\Delta t^2$.

Additionally, Assumption 2 (iii) implies that

$$
\begin{aligned}
I_2 \leq& \int_\beta \|g(T_\rho(H,t),\beta)-g(\widetilde{T}_\rho(H,t),\beta)\|_F\rho(\beta|t)d\beta \\
\leq& \phi_T(N,D,\sqrt{N+1}B_T)(1+r+r^2)\|T_\rho(H,t),\beta)-\widetilde{T}_\rho(H,t),\beta)\|_F
\end{aligned}
\tag{F.33}
$$

Therefore, by bounding I_1+I_2, we obtain the following inequality

$$
\|T_\rho(H,t+\Delta t),\beta)-\widetilde{T}_\rho(H,t+\Delta t),\beta)\|_F \leq C_{3,5}\Delta t^2+(1+C_{3,6}\Delta t)\|T_\rho(H,t),\beta)-\widetilde{T}_\rho(H,t),\beta)\|_F, \tag{F.34}
$$

which implies, after being used multiple times, that

$$
\|T_\rho(H,t),\beta)-\widetilde{T}_\rho(H,t),\beta)\|_F \leq C_{3,5}\Delta t^2\frac{(1+C_{3,6}\Delta t)^{L+1}-1}{C_{3,6}\Delta t}\leq C_3L^{-1} \tag{F.35}
$$

for any $t=0,\Delta t,\ldots,(L-1)\Delta t,1$ and some constant C_3. Combining (F.24), (F.28), and (F.35) yields the desired result. □

F.7 Proof of Lemma E.1

Proof. Let $\widetilde{\mu}(t)$ be the measure induced by $T_\rho(H,t)$ with $H \sim \mu$, and $\widetilde{\mu}$ be $\widetilde{\mu}(t^*)$. By verifying Lemma C.1, we establish that $\|T_\rho(H,t)\|_{2-\text{col}} \le B_T := B\exp(K(1+r+r^2))$ for any H and $t \in [0,1]$. Consequently, $\text{supp}(\widetilde{\mu}(t)) \subset \{T : \|T\|_{2-\text{col}} \le B_T\}$ for any $t \in [0,1]$. The remainder of the proof involves four steps:

Step I: Show that $p_\rho(H,t)$ is Lipschitz continuous with respect to (H,t)

In the proof of this proposition, when referring to the Lipschitz continuity of a function, we imply its Lipschitz continuity for H within the support of μ. Recall that we have shown in Lemma C.3 that $p_\rho(H,t)$ is universally bounded and Lipschitz continuous for $\|\cdot\|_F$ and any $t \in [0,1]$.

From the formula of p_ρ in (C.14), for any H, H', we have

$$
\begin{aligned}
&\|p_\rho(H,t) - p_\rho(H',t)\|_F = \|\text{vec}[p_\rho(H,t)] - \text{vec}[p_\rho(H',t)]\| \\
&\le \underbrace{|\text{Read}[T_\rho(H,1) - T_\rho(H',1)] - (y(H) - y(H'))|}_{I_1} \underbrace{\Big\|\exp\Big(\int_t^1 \int_\beta \nabla_{\text{vec}[T]}\text{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta ds\Big)\Big\|}_{I_2} \\
&+ \underbrace{|\text{Read}[T_\rho(H',1)] - y(H')|}_{I_3} \\
&\cdot \underbrace{\Big\|\exp\Big(\int_t^1 \int_\beta \nabla_{\text{vec}[T]}\text{vec}[g(T_\rho(H,s),\beta)]\rho(\beta,s)d\beta ds\Big) - \exp\Big(\int_t^1 \int_\beta \nabla_{\text{vec}[T]}\text{vec}[g(T_\rho(H',s),\beta)]\rho(\beta,s)d\beta ds\Big)\Big\|}_{I_4}.
\end{aligned}
\tag{F.36}
$$

From Proposition C.1, we observe that $p_\rho(H,t)$ is Lipschitz continuous. Since $y(\cdot)$ is K_y-Lipschitz continuous, as given in Assumption 4 (iii), we obtain that $I_1 \lesssim \|H - H'\|_F$. Moreover, from Assumption 1 and Lemma C.1, we see that I_3 is universally bounded. Therefore, to show that $p_\rho(H,t)$ is Lipschitz continuous for $\|\cdot\|_F$, it suffices to demonstrate that $I_2 \lesssim 1$ and $I_4 \lesssim \|H - H'\|_F$.

From Assumption 2 (iii), we have

$$
I_2 \le \exp\Big(\int_t^1 \int_\beta \big\|\nabla_{\text{vec}[T]}\text{vec}[g(T_\rho(H,s),\beta)]\big\|\,\rho(\beta,s)d\beta ds\Big) \le \exp\Big(\phi_T(N,D,\sqrt{N+1}B_T)(1+r+r^2)\Big).
$$

Thus, dividing I_4 by the uniformly bounded part I_2, we obtain

$$
\begin{aligned}
I_4 &\le \Big\|\exp\Big(\int_t^1 \int_\beta \nabla_{\text{vec}[T]}\Big(\text{vec}[g(T_\rho(H',s),\beta)] - \nabla_{\text{vec}[T]}\text{vec}[g(T_\rho(H,s),\beta)]\Big)\rho(\beta,s)d\beta ds\Big) - I_{\text{divvec}[T]}\Big\| \\
&\le \exp\Big(\int_t^1 \int_\beta \Big\|\nabla_{\text{vec}[T]}\Big(\text{vec}[g(T_\rho(H',s),\beta)] - \nabla_{\text{vec}[T]}\text{vec}[g(T_\rho(H,s),\beta)]\Big)\Big\|\,\rho(\beta,s)d\beta ds\Big) - 1 \\
&\le \exp\Big(\phi_{TT}(N,D,\sqrt{N+1}B_T,r)\sup_{t\in[0,1]}\|T_\rho(H,t) - T_\rho(H',t)\|_F\Big) - 1,
\end{aligned}
\tag{F.37}
$$

where the final inequality holds by Assumption 3 (iv).

By the Lipschitz continuity of $T_\rho(H,t)$ for $\|\cdot\|_F$ as stated in Proposition C.1, we have $\sup_{t\in[0,1]}\|T_\rho(H,t) - T_\rho(H',t)\|_F \lesssim \|H - H'\|_F$. Hence, utilizing the universal boundedness of the last equation in (F.37), we derive $I_4 \lesssim \|H - H'\|_F$.

Considering all assertions regarding I_1, I_2, I_3 and I_4, we conclude that $p_\rho(H,t)$ is C_p-Lipschitz continuous with respect to H for some universal constant C_p that is sufficiently large. The Lipschitz continuity of $p_\rho(H,t)$ with respect to t could be easily derived from the boundedness of the Jacobian matrix, as asserted in Assumption 2 (iii). Moreover, since $p_\rho(H,t)$ is universally bounded shown in Lemma C.3, we conclude that $p_\rho(H,t)$ is Lipschitz continuous with respect to (H,t) for some universal Lipschitz constant C_p that is sufficiently large.

Step II: Prepare bounds related to p_ρ for later use

(C.14) implies that p_ρ solves the adjoint equation

$$\mathrm{vec}[\dot{p}_\rho(H,t)]^T = -\mathrm{vec}[p_\rho(H,t)]^T \int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t)d\beta \tag{F.38}$$

with

$$\left\| \int_\beta \nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T_\rho(H,t),\beta)]\rho(\beta,t)d\beta \right\| \le (1+C_\rho/2)\phi_T(N,D,\sqrt{N+1}B_T)(1+r+r^2) =: C_0$$

implied by Assumption 2 (iii). Therefore, the Grönwall's inequality directly indicates that

$$\mathbb{E}_\mu\|p_\rho(H,t)\|_F^2 \ge \exp(-2C_0 t)\mathbb{E}_\mu\|p_\rho(H,1)\|_F^2 \ge \exp(-2C_0)\mathbb{E}_\mu[|\mathrm{Read}[T_\rho(H,1)] - y(H)|^2] \ge 2\exp(-2C_0)R(\rho),$$
$$\mathbb{E}_\mu\|p_\rho(H,t)\|_F^2 \le \exp(2C_0 t)\mathbb{E}_\mu\|p_\rho(H,1)\|_F^2 \le \exp(2C_0)\mathbb{E}_\mu[|\mathrm{Read}[T_\rho(H,1)] - y(H)|^2] \le 2\exp(2C_0)R(\rho),$$

for any $t \in [0,1]$.

Step III: Construct the descent direction ν

By the well-posedness of the ODE solution to (3.1) as shown in Proposition C.1, the solution map $T_\rho(H,\cdot)$ is invertible. Hence, for any $\rho \in \mathcal{P}^2$, there exists a continuous inverse map T_t^{-1} such that $T_t^{-1}(T_\rho(H,t)) = H$ for any H and $t \in [0,1]$. Let's define the following function $\bar{F}(T)$ to approximate

$$\bar{F}(T) = -p_\rho(T_{t^*}^{-1}(T),t^*) + \int_\beta g(T,\beta)\rho(\beta|t^*)d\beta - \frac{h(T,w_0)}{2},$$

The function $\bar{F}(T)$, arising from the composition of $p_\rho(\cdot,t)$ and $T_{t^*}^{-1}(\cdot)$, exhibits continuity over $T \in \mathrm{supp}(\widetilde{\mu})$ owing to the continuous nature of $p_\rho(\cdot,t)$ and $T_{t^*}^{-1}(\cdot)$. Therefore, Assumption 4 (ii) could be applied to $\bar{F}$. Since $f(\cdot,\theta)$ is a universal kernel constrained on $\theta \in \mathbb{R}^{\mathrm{dim}\theta_1} \times \mathcal{K}$ [46], and $\bar{F}(T)$ is continuous with respect to T, there exists a sequence $\{(c_k,\theta^k)\}_{k\ge 0} \subset (\mathbb{R} \times \mathbb{R}^{\mathrm{dim}\theta_1} \times \mathcal{K})^{\mathbb{N}}$ such that

$$\left\| \bar{F}(T) - \sum_{k=1}^{\infty} c_k f(T,\theta^k) \right\|_{\max} \le \epsilon/3 \tag{F.39}$$

given some $\epsilon > 0$ and any T such that $\|T\|_{2-\mathrm{col}} \le B_T$. Notably, since $f(T,\theta_1^k,\theta_2^k) = -f(T,-\theta_1^k,\theta_2^k)$, we could assume without loss of generality that $c^k \ge 0$ for any k. Furthermore, there exists a constant k_ϵ such that

$$\left\| \bar{F}(T) - \sum_{k=1}^{k_\epsilon} c_k f(T,\theta^k) \right\|_{\max} \le 2\epsilon/3. \tag{F.40}$$

We define $C(\epsilon) := \sum_{k=1}^{k_\epsilon} c_k$, and $\bar{\nu}(\beta) \in \mathcal{P}(\mathbb{R}^{\mathrm{dim}\beta})$ as the probability distribution such that, given $\bar{\beta} \sim \bar{\nu}$ and for any $k \ge 0$, $\bar{\beta}$ has probability $c^k/C(\epsilon)$ of being $(2C(\epsilon)\theta_1^k, \theta_2, 0)$. Then, (F.40) transforms into

$$\left\| F(T) - \int_\beta g(T,\beta)\bar{\nu}(\beta)d\beta \right\|_{\max} \le 2\epsilon/3, \tag{F.41}$$

where

$$F(T) == -p_\rho(T_{t^*}^{-1}(T),t^*) + \int_\beta g(T,\beta)\rho(\beta|t^*)d\beta.$$

From (F.41), we claim that there exists some $R(\epsilon)$ such that

- $\bar{\nu}(\{\beta : 1/R(\epsilon) \le \|\beta\| \le R(\epsilon)\}) \ge 1/2$,
- $\left\| \int_{\|\beta\|>R(\epsilon)} g(T,\beta)\bar{\nu}(\beta)d\beta \right\|_{\max} \le \epsilon/3$.

We are now in the position to define the descent direction ν. By defining $\nu \in \mathcal{P}(\mathbb{R}^{\mathrm{dim}\beta})$ as the measure obtained by truncating any part outside $1/R(\epsilon) \le \|\beta\| \le R(\epsilon)$ from $\bar{\nu}$, and scaling the measure function by $1/\bar{\nu}(\{\beta : 1/R(\epsilon) \le \|\beta\| \le R(\epsilon)\}) \le 2$, we can establish that

$$\left\| F(T) - \int_\beta g(T,\beta)\nu(\beta)d\beta \right\|_{\max} \le \epsilon. \tag{F.42}$$

for any T such that $\|T\|_{2-\mathrm{col}} \leq B_T$. A straightforward deduction from (F.42) is that for any T such that $\|T\|_{2-\mathrm{col}} \leq B_T$, we have $\|F(T) - \int_\beta g(T,\beta)\nu(\beta)d\beta\|_F \leq \epsilon\sqrt{(N+1)D}$. It is clear that ν has a bounded support as $\{\beta : 1/R(\epsilon) \leq \|\beta\| \leq R(\epsilon)\}$. We will determine the value of ϵ later, ensuring it based only on N, D, r, and the parameters of the assumptions.

Step IV: Upper bound $\int_\beta \frac{\delta Q}{\delta \rho}(\beta, t^*)\Big(\nu(\beta) - \rho(\beta|t^*)\Big)d\beta$ to complete the proof

Utilizing the gradient definition in (3.4) and $\int_\beta g(T,\beta)\rho(\beta|t^*)d\beta = F(T) + p_\rho(T_{t^*}^{-1}(T), t^*)$, we obtain

$$
\begin{aligned}
&\int_\beta \frac{\delta Q}{\delta \rho}(\beta, t^*)\Big(\nu(\beta) - \rho(\beta|t^*)\Big)d\beta \\
=&\mathbb{E}_\mu \int_\beta \mathrm{Tr}\Big[g(T_\rho(H,t^*),\beta)^T p_\rho(H,t^*)\Big]\Big(\nu(\beta) - \rho(\beta|t^*)\Big)d\beta + \frac{\lambda}{2}\int_\beta \|\beta\|^2\Big(\nu(\beta) - \rho(\beta|t^*)\Big)d\beta \\
\leq&\mathbb{E}_{T\sim\widetilde{\mu}(t)} \mathrm{Tr}\Big[g(T,\beta)^T\Big(\widetilde{v}(\beta) - \rho(\beta|t^*)\Big)p_\rho(T_{t^*}^{-1}(T), t^*)\Big] + \frac{\lambda}{2}(R(\epsilon) + r) \\
=&\underbrace{\mathbb{E}_{T\sim\widetilde{\mu}(t)} \mathrm{Tr}\Big[\Big(F(T) - \int_\beta g(T,\beta)\nu(\beta)d\beta\Big)^T p_\rho(T_{t^*}^{-1}(T), t^*)\Big]}_{J_1} \\
-&\underbrace{\mathbb{E}_{T\sim\widetilde{\mu}(t)} \mathrm{Tr}\Big[p_\rho(T_{t^*}^{-1}(T), t^*)^T p_\rho(T_{t^*}^{-1}(T), t^*)\Big]}_{J_2} + \frac{\lambda}{2}(R(\epsilon) + r)
\end{aligned}
$$

For $\rho \in \mathcal{P}^2$ concentrated on P_r, we observe

$$R(\rho) \leq \frac{1}{2}\mathbb{E}_\mu[\Big(|\mathrm{Read}[T_\rho(H,1)]| + |y(H)|\Big)^2] \leq \frac{1}{2}(B_T + B)^2.$$

Hence, to bound J_1, we have

$$
\begin{aligned}
J_1 \leq&\mathbb{E}_{T\sim\widetilde{\mu}(t)}\|F(T) - \int_\beta g(T,\beta)\nu(\beta)d\beta\|_F \cdot \|p_\rho(T_{t^*}^{-1}(T), t^*)\|_F dt \\
\leq&\sqrt{(N+1)D}\epsilon\mathbb{E}_{T\sim\widetilde{\mu}(t)}\|p_\rho(T_{t^*}^{-1}(T), t^*)\|_F dt \\
\leq&(N+1)D\epsilon\Big(\mathbb{E}_{T\sim\widetilde{\mu}(t)}\|p_\rho(T_{t^*}^{-1}(T), t^*)\|_F^2\Big)^{1/2} dt \\
=&(N+1)D\epsilon\Big(\mathbb{E}_\mu\|p_\rho(H,t)\|_F^2\Big)^{1/2} dt \\
\leq&\sqrt{2}(N+1)D\exp(C_0)R(\rho)^{1/2}\epsilon \\
\leq&(N+1)(B_T + B)(N+1)D\exp(C_0)R(\rho)\epsilon \\
=&C_3 R(\rho)\epsilon,
\end{aligned}
\tag{F.43}
$$

where $C_3 := (N+1)(B_T + B)(N+1)D\exp(C_0)$.

To bound J_2, we have

$$
\begin{aligned}
J_2 =&\mathbb{E}_{T\sim\widetilde{\mu}(t)}\|p_\rho(T_t^{-1}(T), t)\|_F^2 dt \\
=&\mathbb{E}_\mu\|p_\rho(H,t)\|_F^2 dt \\
\geq&\mathbb{E}_\mu\exp(-2C_0)\|p_\rho(H,1)\|_F^2 dt \\
\geq&\frac{1}{2}\exp(-2C_0)R(\rho).
\end{aligned}
\tag{F.44}
$$

Combining (F.43) and (F.44), by choosing $\epsilon = \frac{1}{4}\exp(-2C_0)/C_3$, which only depends on N, D, r, and the parameters of the assumptions, we have

$$\int_\beta \frac{\delta Q}{\delta \rho}(\beta, t^*)\Big(\nu(\beta) - \rho(\beta|t^*)\Big)d\beta \leq -\frac{1}{4}\exp(-2C_0)R(\rho) + \frac{\lambda}{2}\Big(R(\frac{1}{4}\exp(-2C_0)/C_3) + r\Big).$$

Setting $B_r = R(\frac{1}{4}\exp(-2C_0)/C_3)$, $C_1 = \Big(R(\frac{1}{4}\exp(-2C_0)/C_3) + r\Big)/2$, and $C_2 = \frac{1}{4}\exp(-2C_0)$ completes the proof. □

F.8 Proof of Lemma C.1

Proof. By applying the Cauchy-Schwarz inequality, we trivially obtain $\int_0^1 \int_\beta ||\beta||_2 \rho(\beta, t) d\beta dt \leq A$. Thus, leveraging (3.1) and Assumption 2 (i), we can infer

$$\begin{aligned}\frac{d}{dt}\|T_\rho(H,t)\|_{2-\mathrm{col}} \leq \|\dot{T}_\rho(H,t)\|_{2-\mathrm{col}} &= \|\int_\beta g(T_\rho(H,t),\beta)\rho(\beta,t)d\beta\|_{2-\mathrm{col}} \\ &\leq \int_\beta \|g(T_\rho(H,t),\beta)\|_{2-\mathrm{col}}\rho(\beta,t)d\beta \\ &\leq \int_\beta K(1+||\beta||_2+||\beta||_2^2)\rho(\beta,t)\|T_\rho(H,t)\|_{2-\mathrm{col}}d\beta.\end{aligned} \tag{F.45}$$

Therefore, by the Grönwall's inequality, we have

$$\begin{aligned}\|T_\rho(H,t)\|_{2-\mathrm{col}} &\leq \|T_\rho(H,0)\|_{2-\mathrm{col}} \exp(\int_0^1 \int_\beta K(1+||\beta||_2+||\beta||_2^2)\rho(\beta,t)d\beta dt) \\ &\leq \|H\|_{2-\mathrm{col}} \exp(K(1+A+A^2)).\end{aligned}$$

□

F.9 Proof of Lemma C.2

Proof. As per Lemma C.1, the boundedness of $\|T_\nu(H,t)\|_{2-\mathrm{col}}$ and $\|T_\rho(H,t)\|_{2-\mathrm{col}}$ is established by a constant $C := \|H\|_{2-\mathrm{col}} \exp(K(1+A+A^2)) > 0$ for all $t \in [0,1]$. Consequently, from (3.1), this implies

$$\begin{aligned}\|T_\nu(H,t_1)-T_\nu(H,t_2)\|_{2-\mathrm{col}} &\leq \int_{t_1}^{t_2} \|\dot{T}_\nu(H,t)\|_{2-\mathrm{col}} \\ &\leq \int_{t_1}^{t_2}\int_\beta \|g(T_\nu(H,t),\beta\|_{2-\mathrm{col}}\nu(\beta,t)d\beta dt \\ &\leq (1+C_\rho/2)KC(1+r+r^2)(t_2-t_1).\end{aligned} \tag{F.46}$$

for any $t_1, t_2 \in [0,1]$. Therefore, $T_\nu(H,t)$ is $(1+C_\rho/2)KC(1+r+r^2)$-Lipschitz with respect to t for $\|\cdot\|_{2-\mathrm{col}}$, and thus $\sqrt{N+1}(1+C_\rho/2)KC(1+r+r^2)$-Lipschitz with respect to t for $\|\cdot\|_F$. Note that by (3.1),

$$\begin{aligned}\Delta(H,t) :=& \|T_\rho(H,t)-T_\nu(H,t)\|_F \\ =& \|\int_0^t \dot{T}_\rho(H,s)-\dot{T}_\nu(H,s)ds\|_F \\ =& \|\int_0^t\int_\beta g(T_\rho(H,s),\beta)\rho(\beta,t)d\beta ds - \int_0^t\int_\beta g(T_\nu(H,s),\beta)\nu(\beta,s)d\beta ds\|_F \\ \leq& \underbrace{\int_0^t\int_\beta \|g(T_\rho(H,s),\beta)-g(T_\nu(H,s),\beta)\|_F\rho(\beta,s)d\beta ds}_{J_1} \\ &+ \underbrace{\|\int_0^t\int_\beta g(T_\nu(H,s),\beta)\Big(\rho-\nu\Big)(\beta,s)d\beta ds\|_F}_{J_2}\end{aligned} \tag{F.47}$$

We then bound J_1 and J_2 using the following two lemmas separately. Firstly, since $\|T_\rho\|_F \leq \sqrt{N+1}C$ and $\|T_\nu\|_F \leq \sqrt{N+1}C$, we have by Assumption 2 that

$$\begin{aligned}\|g(T_\rho(H,s),\beta)-g(T_\nu(H,s),\beta)\|_F &\leq \Big(\sup_{\|T\|_F\leq\sqrt{N+1}C}\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[g(T,\beta)]\|_2\Big)\|T_\rho(H,s)-T_\nu(H,s)\|_F \\ &\leq \phi_T(N,D,\sqrt{N+1}C)(1+r+r^2)\Delta(H,s)\end{aligned} \tag{F.48}$$

Therefore, by (F.48) we have

$$J_1 \leq \phi_T(N, D, \sqrt{N+1}C)(1 + C_\rho/2)(1 + r + r^2) \int_0^t \Delta(H, s) ds. \tag{F.49}$$

Secondly, we aim to bound the integral J_2 given Assumption 2 and $\|T_\nu(H,t)\|_{2-\mathrm{col}} \leq C$ on $t \in [0,1]$. Again by Assumption 2 we have

$$\|\nabla_\beta \mathrm{vec}[g(T_\nu(H,s),\beta)]\|_F \leq \sqrt{\sum_{i=1}^{N+1} \|\nabla_\beta g(T_\nu(H,s),\beta)_{:,i}\|_{2-\mathrm{col}}^2} \leq \sqrt{N+1}\phi_P(C)(1+r) \tag{F.50}$$

$$\|\nabla_T \mathrm{vec}[g(T_\nu(H,s),\beta)]\|_F \leq \phi_T(N, D, \sqrt{N+1}C)(1 + r + r^2) \tag{F.51}$$

Since $T_\nu(H,s)$ is $\sqrt{N+1}(1+C_\rho/2)KC(1+r+r^2)$-Lipschitz with respect to s for $\|\cdot\|_F$ (as shown in (F.46)), by (F.51), we obtain that $g(T_\nu(H,s),\beta)$ is $\sqrt{N+1}(1+C_\rho/2)KC\phi_T(N, D, \sqrt{N+1}C)(1+r+r^2)^2$-Lipschitz with respect to s. Thus, $g(T_\nu(H,s),\beta)$ is C'-Lipschitz with respect to (s,β), where $C' = \sqrt{N+1}(1+C_\rho/2)KC\phi_T(N, D, \sqrt{N+1}C)(1+r+r^2)^2 + \sqrt{N+1}\phi_P(C)(1+r)$. This indicates, by the Kantorovich-Rubinstein Theorem (see Theorem 5.10 of [61], for example), that

$$J_2 = \|\int_0^1 \int_D g(T_\nu(H,s),\beta)\Big(\rho - \nu\Big)(\beta,s) d\beta ds\|_F \leq C' W_1(\rho,\nu). \tag{F.52}$$

Define $C^* := \max\{C', \phi_T(N, D, \sqrt{N+1}C)(1+r+r^2)\}$. By combining (F.49) and (F.52), we have

$$\Delta(H,t) \leq C^* \int_0^t \Delta(H,s) ds + C^* W_1(\rho,\nu). \tag{F.53}$$

Applying the Grönwall's inequality then shows

$$\Delta(H,t) \leq C^* \exp(C^* t) W_1(\rho,\nu). \tag{F.54}$$

Specifically, we have $\|T_\rho(H,1) - T_\nu(H,1)\|_F \leq C^* \exp(C^*) W_1(\rho,\nu)$. $\square$

F.10 Proof of Lemma C.3

Proof. Let's consider a fixed (β,t) pair within P_r. Given Lemma C.1, we establish $\|T_\rho(H,t)\|_{2-\mathrm{col}} \leq B\exp(K(1+A+A^2))$. Consequently, under Assumption 1, we deduce

$$\begin{aligned}\|g(T_\rho(H,t),\beta)\|_F &\leq \sqrt{N+1}\|g(T_\rho(H,t),\beta)\|_{2-\mathrm{col}} \\ &\leq \sqrt{N+1}K\|T_\rho(H,t)\|_{2-\mathrm{col}}(1+r+r^2) \\ &\leq \sqrt{N+1}KB\exp(K(1+A+A^2))(1+r+r^2).\end{aligned} \tag{F.55}$$

On the other hand, from (C.14), we have

$$\begin{aligned}\|p_\rho(H,t)\|_F \leq& |\mathrm{Read}(T_\rho(H,1)) - y(H)| \cdot \left\|\exp\Big(\int_t^1 \int_\beta \nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s),\beta)\rho(\beta,s) d\beta ds]\Big)\right\|_2 \\ &\leq (B + B\exp(K(1+A+A^2)))\exp\Big(\int_t^1 \int_\beta \left\|\nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s),\beta)]\right\| \rho(\beta,s) d\beta ds\Big) \\ &\leq (B + B\exp(K(1+A+A^2)))\exp\Big(\int_t^1 \int_\beta \left\|\nabla_{\mathrm{vec}[T]} \mathrm{vec}[g(T_\rho(H,s),\beta)]\right\| \rho(\beta,s) d\beta ds\Big) \\ &\leq (B + B\exp(K(1+A+A^2)))\exp\Big(\int_0^1 \int_\beta \phi_T(N, D, \|T_\rho(H,s)\|_F)(1 + \|\beta\| + \|\beta\|^2)\rho(\beta,s) d\beta ds\Big) \\ &\leq (B + B\exp(K(1+A+A^2)))\exp\Big(\phi_T(N, D, \sqrt{N+1}KB\exp(K(1+A+A^2))(1+A+A^2)\Big)\end{aligned} \tag{F.56}$$

Combining (F.55) and (F.56), we could obtain

$$\begin{aligned}\left\|\frac{\delta Q}{\delta \rho}(\beta,t)\right\| \leq& \mathbb{E}_\mu \|g(T_\rho(H,t),\beta)\|_F \|p_\rho(H,t)\|_F + \frac{\lambda}{2}r^2 \\ \leq& \sqrt{N+1}KB\exp(K(1+A+A^2))(1+r+r^2)(B + B\exp(K(1+A+A^2))) \\ & \exp\Big(\phi_T(N, D, \sqrt{N+1}KB\exp(K(1+A+A^2))(1+A+A^2)\Big) + \frac{\lambda}{2}r^2.\end{aligned}$$

$\square$

F.11 Proof of Lemma C.4

Proof. By employing the Cauchy-Schwarz inequality, we readily observe that $\frac{1}{ML}\sum_t \sum_{j=1}^M \|\beta\| \leq A$. Consequently, leveraging (2.4) and Assumption 2, we ascertain that for any $t = 0, \Delta t, \dots, (L-1)\Delta t$, we obtain:

$$\|\widehat{T}_\Theta(H, t+\Delta t/2)\|_{2-\mathrm{col}} \leq \|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}}\Big\{1 + \frac{1}{2ML}\sum_{j=1}^M K(1+\|\theta_{t,j}\| + \|\theta_{t,j}\|^2)\Big\}$$
$$\|\widehat{T}_\Theta(H, t+\Delta t)\|_{2-\mathrm{col}} \leq \|\widehat{T}_\Theta(H,t+\Delta t/2)\|_{2-\mathrm{col}}\Big\{1 + \frac{1}{2ML}\sum_{j=1}^M K(1+\|w_{t,j}\| + \|w_{t,j}\|^2)\Big\}. \tag{F.57}$$

Therefore, by applying (F.57) multiple times, we obtain that for any $t = 0, \Delta t/2, \dots, (L-1/2)\Delta t, 1$,

$$\begin{aligned}\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}} \leq & \|\widehat{T}_\Theta(H,0)\|_{2-\mathrm{col}} \prod_t \Big\{1 + \frac{1}{2ML}\sum_{j=1}^M K(1+\|\theta_{t,j}\| + \|\theta_{t,j}\|^2)\Big\}\Big\{1 + \frac{1}{2ML}\sum_{j=1}^M K(1+\|w_{t,j}\| + \|w_{t,j}\|^2)\\ \leq & \|H\|_{2-\mathrm{col}} \exp\Big(\frac{K}{ML}\sum_t\sum_{j=1}^M 1 + \|\beta_{t,j}\| + \|\beta_{t,j}\|^2\Big)\\ \leq & \|H\|_{2-\mathrm{col}} \exp\Big(K(1+A+A^2)\Big).\end{aligned}$$

□

F.12 Proof of Lemma C.5

Proof. Lemma C.4 shows that $\|\widetilde{T}_\Theta(H,t)\|_{2-\mathrm{col}}$ and $\|\widetilde{T}_{\widetilde{\Theta}}(H,t)\|_{2-\mathrm{col}}$ are bounded by $B_T := B\exp(K(1+A+A^2))$ for any H and $t \in [0,1]$. From (2.4), for any $t = 0, \dots, (L-1)\Delta t$, from Assumption 2 (ii) and (iii) we have

$$\begin{aligned}&\|\widehat{T}_\Theta(H,t+\Delta t/2) - \widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t/2)\|_F\\ \leq&\|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_F + \frac{\Delta t/2}{M}\sum_{j=1}^M \|f(\widehat{T}_\Theta(H,t),\theta_{t,j}) - f(\widehat{T}_{\widetilde{\Theta}}(H,t),\widetilde{\theta}_{t,j})\|_F\\ \leq&\|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_F + (\Delta t/2)\sqrt{N+1}\phi_P(B_T)(1+r)M^{-1}\sum_{j=1}^M\|\theta_{t,j} - \widetilde{\theta}_{t,j}\|\\ &+(\Delta t/2)\phi_T(N,D,\sqrt{N+1}B_T)(1+r+r^2)\|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_F\\ \leq&\|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_F(1+C_1(\Delta t/2)) + (\Delta t/2)C_2M^{-1}\sum_{j=1}^M\|\theta_{t,j} - \widetilde{\theta}_{t,j}\|.\end{aligned} \tag{F.58}$$

for some constant C_1 and C_2 depending only N, D, r and assumptions. Similarly, we have

$$\begin{aligned}\|\widehat{T}_\Theta(H,t+\Delta t) - \widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t)\|_F \leq&\|\widehat{T}_\Theta(H,t+\Delta t/2) - \widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t/2)\|_F(1+C_1(\Delta t/2))\\ &+(\Delta t/2)C_2M^{-1}\sum_{j=1}^M\|w_{t,j} - \widetilde{w}_{t,j}\|.\end{aligned} \tag{F.59}$$

Combining (F.58) and (F.59), we derive

$$\|\widehat{T}_\Theta(H,t+\Delta t) - \widehat{T}_{\widetilde{\Theta}}(H,t+\Delta t)\|_F \leq \|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_F(1+C_3\Delta t) + \Delta t C_3 M^{-1}\sum_{j=1}^M\|\beta_{t,j} - \widetilde{\beta}_{t,j}\|. \tag{F.60}$$

where C_3 is a constant depending solely on N, D, r, and the parameters of the assumptions. Iterating (F.60) multiple times yields

$$\|\widehat{T}_\Theta(H,t) - \widehat{T}_{\widetilde{\Theta}}(H,t)\|_F \leq \exp(C_3)\frac{1}{ML}d(\Theta,\widetilde{\Theta})$$

for any $t = 0, \Delta t, \ldots, 1$. □

F.13 Proof of Lemma C.6

Proof. By verifying that Θ satisfies the conditions outlined in Lemma C.4, we establish $\|\widehat{T}_\Theta(H,t)\|_{2-\mathrm{col}} \le B\exp(K(1+A+A^2))$. The first two results stem from Assumption 2 (i) and (ii), with recognition that $\|T\|_F \le \sqrt{N+1}\|T\|_{2-\mathrm{col}}$ for any $T \in \mathbb{R}^{D\times(N+1)}$. As for the third result, consider $t = 0, \Delta t, \ldots, (L-1)\Delta t, 1$, where

$$\begin{aligned}
&\max\{\|\widehat{p}_\Theta(H,t)\|_F, \|\widehat{p}_\Theta(H,t+\Delta t/2)\|_F\} \\
=&\max\{\|\mathrm{vec}[\widehat{p}_\Theta(H,t)]\|, \|\mathrm{vec}[\widehat{p}_\Theta(H,t+\Delta t/2)]\|\} \\
\le&|\mathrm{Read}[\widehat{T}_\Theta(H,1)] - y(H)| \\
&\Big\{\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}} \Big\|I_{\dim\theta} + (\Delta t/2)\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\Big\| \\
&\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}} \Big\|I_{\dim w} + (\Delta t/2)\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]\Big\|\Big\} \\
\le&|\mathrm{Read}[\widehat{T}_\Theta(H,1)] - y(H)| \\
&\prod_{\substack{(s-t)/\Delta t+1\in[(1-t)/\Delta t]\\ j\in[M]}} \Big(1 + (\Delta t/2)\Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\Big\|\Big) \\
&\Big(1 + (\Delta t/2)\Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_\Theta(H,s+\Delta t/2),w_{s,j})]\Big\|\Big) \\
\le&(B+B_T) \\
&\exp\Big((\Delta t/2)\sum_{(s-t)/\Delta t+1\in[(1-t)/\Delta t]} M^{-1}\sum_{j=1}^{M}\Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[f(\widehat{T}_\Theta(H,s),\theta_{s,j})]\Big\| \\
+&(\Delta t/2)\sum_{(s-t)/\Delta t+1\in[(1-t)/\Delta t]} M^{-1}\sum_{j=1}^{M}\Big\|\nabla_{\mathrm{vec}[T]}\mathrm{vec}[h(\widehat{T}_\Theta(H,t),w_{s,j})]\Big\|\Big) \\
\le&(B+B_T)\exp\Big(\phi_T(N,D,\sqrt{N+1}B_T)\frac{1}{ML}\sum_t\sum_{j=1}^{M}(1+\|\beta\|+\|\beta\|^2)\Big) \\
\le&(B+B_T)\exp\Big(\phi_T(N,D,\sqrt{N+1}KB_T)(1+A+A^2)\Big),
\end{aligned} \tag{F.61}$$

where the first inequality arises from the fact that the matrix 2-norm is greater equal than the norm of any of its columns, and the fourth inequality follows from Assumption 2 (iii). □

F.14 Proof of Lemma C.7

Proof. Note that $\|\beta_{t,j}\| \le r$ for any $t = 0, \Delta t, \ldots, (L-1)\Delta t$ and $j \in [M]$ with its expectation denoted as $\int_\beta \|\beta_{t,j}\|\rho(\beta,t)d\beta$. Applying Hoeffding's inequality yields, for any $z > 0$ and $t = 0, \Delta t, \ldots, (L-1)\Delta t$

$$\mathbb{P}(|M^{-1}\sum_{j=1}^{M}\|\beta_{t,j}\|^2 - \int_\beta \|\beta\|^2\rho(\beta,t)d\beta| \ge z) \le 2\exp(-\frac{z^2}{2r^2}M).$$

By applying the union bound over $t = 0, \Delta t, \ldots, (L-1)\Delta t$, the inequality above implies

$$\begin{aligned}
\mathbb{P}(|\frac{1}{ML}\sum_t\sum_{j=1}^{M}\|\beta_{t,j}\|^2 - \frac{1}{L}\sum_t\int_\beta\|\beta\|^2\rho(\beta,t)d\beta| \ge z) \le&\mathbb{P}(\sup_t|M^{-1}\sum_{j=1}^{M}\|\beta_{t,j}\|^2 - \int_\beta\|\beta\|^2\rho(\beta,t)d\beta| \ge z) \\
\le&2L\exp(-\frac{z^2}{2r^2}M).
\end{aligned} \tag{F.62}$$

In addition, we have

$$
\begin{aligned}
|\frac{1}{L}\sum_t \int_\beta \|\beta\|^2 \rho(\beta,t)d\beta - \int_0^1 \int_\beta \|\beta\|^2\rho(\beta,t)d\beta dt| \le & \sup_{|t-s|\le \Delta t} |\int_\beta \|\beta\|^2\rho(\beta,t)d\beta - \int_\beta \|\beta\|^2 \rho(\beta,s)d\beta| \\
\le & r^2 L^{-1} \sup_{|t-s|\le\Delta t} \|\rho(\cdot,t)-\rho(\cdot,s)\|_{\mathrm{BL}} \\
\le & r^2 C_\rho L^{-1}.
\end{aligned}
\tag{F.63}
$$

Combining (F.62) and (F.63), and setting $z = r\sqrt{2M^{-1}(\delta+\log(2L))}$ completes the proof. □

F.15 Proof of Lemma C.8

Proof. The proof will be trivial by noting the equality

$$
\prod_{l=1}^L A_l - \prod_{l=1}^L B_l = \sum_{l=1}^L \Big(\prod_{s=1}^{l-1} B_s (A_l - B_l) \prod_{s=l+1}^L A_s\Big).
$$

Hence, we have

$$
\|\prod_{l=1}^L A_l - \prod_{l=1}^L B_l\| \le \sum_{l=1}^L \|\prod_{s=1}^{l-1} B_s\| \cdot \|A_l - B_l\| \cdot \|\prod_{s=l+1}^L A_s\| \le C\sum_{l=1}^L \|A_l - B_l\|.
$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the main contributions and scope of the paper (Section 1.1), clearly outlining the theoretical advancements while aligning with the stated assumptions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We operate under noiseless label settings, as noisy conditions typically preclude achieving zero risk. The limitations of Assumption 4 and the assumptions in Theorem 4.1 are discussed in the respective sections. Although some assumptions are currently untestable, we provide high-level justifications for their validity.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All necessary assumptions (Assumptions 1-4) are stated in the main content prior to presenting the theoretical results, with complete and correct proofs included in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper does not include experimental results; all results are theoretical and backed by rigorous proofs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This paper does not include experimental results and, consequently, does not feature experiments that require code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper fully complies with the NeurIPS Code of Ethics. All ethical guidelines pertinent to the theoretical nature of the work have been adhered to rigorously.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper does not discuss societal impacts as it strictly addresses theoretical optimization guarantees for Transformer models, focusing solely on providing rigorous proofs without exploring practical applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper is purely theoretical and does not involve the release of data or models. Therefore, there are no risks for misuse or dual-use associated with this research.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper is purely theoretical and does not involve the use of existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper is purely theoretical and does not involve the use of existing assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper is purely theoretical and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper is purely theoretical and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.