Propagation of Chaos in Contextual Flow Maps

Shi Chen **Zhengjiang Lin** **Kaizhao Liu** **Philippe Rigollet**

MIT

Abstract

We develop a quantitative statistical theory of transformers in the long-context regime by adopting the abstraction of *contextual flow maps* (CFMs): dynamical systems that evolve a distinguished token in the presence of a contextual measure across a stack of attention blocks. Within this framework, the finite-context model approximates an idealized infinite-context system in which the contextual measure is replaced by its underlying population, so that the context length n becomes a statistical resource. Exploiting the McKean–Vlasov structure of the dynamics and the classical machinery of propagation of chaos, we establish a *forward* bound controlling the deviation between the finite- and infinite-context CFMs uniformly along depth, and a *backward* bound controlling the deviation between the corresponding training trajectories uniformly across iterations of online gradient descent. Both bounds achieve the optimal Wasserstein rate $n^{-1/d}$ for general CFMs and parametric rate $n^{-1/2}$ for a restricted class of CFMs that includes transformers as a special case. The analysis rests on a new Eulerian adjoint formulation of the loss gradient and stability estimates for the resulting forward–adjoint system, both of which may be of independent interest.

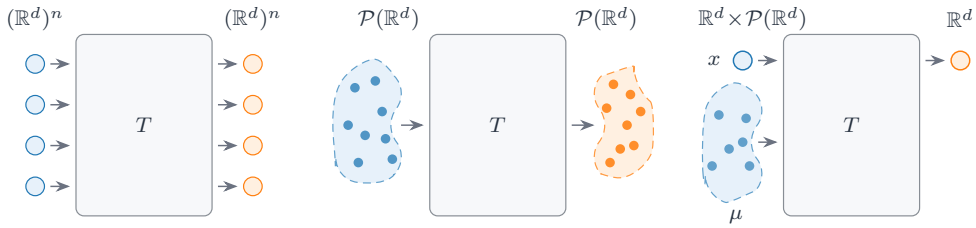
1 Introduction

The transformer architecture, introduced by Vaswani et al. [VSP$^+$17], is the foundation of modern machine learning systems, from large language models [DCLT19] to Vision Transformers [DBK$^+$21]. A convenient way to view inputs to these models is through a distinguished token $x \in \mathbb{R}^d$ that is updated in the presence of a context of n tokens $z^{(1)}, \dots, z^{(n)} \in \mathbb{R}^d$ encoded by the *contextual measure* $\mu = \frac{1}{n}\sum_i \delta_{z^{(i)}}$. In encoder-style models, x is typically a prepended special token such as `[CLS]`; in decoder-style models, it is the current final token whose representation is used to predict the next one. After passing through a deep stack of attention blocks, the model returns an output token y whose value depends on both x and its context. We refer to such maps as *contextual maps*.

This viewpoint on transformers was developed in [FdHP24], where transformers were shown to have universal approximation properties within the class of contextual maps. At the same time, the compositional structure of transformers, like that of deep neural networks more broadly, can be interpreted as a dynamical system, with the network realizing an input-to-output *flow map* [E17, HR18, CRBD18]. Combining these perspectives suggests a natural abstraction: transformers as *contextual flow maps*, namely dynamical systems that evolve an input token-context pair (x_0, μ_0) across layers and produce an output token x_1. We consider a class of general contextual flow maps $(x_0, \mu_0) \mapsto x_1$ that include transformers and take the form

$$\begin{cases} \dot{x}_s = \mathcal{V}(x_s, \mu_s; \theta(s)), \quad s \in [0,1], & \text{(1.1a)} \\ \partial_s \mu_s + \nabla \cdot \big(\mu_s \, \mathcal{V}(\cdot, \mu_s; \theta(s))\big) = 0, \quad s \in [0,1], & \text{(1.1b)} \\ x_s|_{s=0} = x_0, & \text{(1.1c)} \\ \mu_s|_{s=0} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \delta_{z^{(i)}}. & \text{(1.1d)} \end{cases}$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) sequence-to-sequence (b) measure-to-measure (c) contextual map

Figure 1: Three formalizations of transformers.

The final model output is given by the position of the distinguished particle at the terminal time,

$$x_1 = x_1(x_0, \mu_0; \{\theta(s)\}_s) \in \mathbb{R}^d.$$

Here, $\theta(s)$ denotes the set of parameters at layer s. Equation (1.1a) describes the distinguished token dynamics, (1.1b) the contextual measure dynamics, (1.1c) the input distinguished token, and (1.1d) the input contextual measure. For transformers, the self-attention vector field introduced in [SABP22] is given by

$$\mathcal{V}(x, \mu; \theta) = \frac{1}{Z(x)} \int_{\mathbb{R}^d} \exp\big(\langle Qx, Ky\rangle\big) Vy \, \mathrm{d}\mu(y), \qquad Z(x) = \int_{\mathbb{R}^d} \exp\big(\langle Qx, Kz\rangle\big) \, \mathrm{d}\mu(z)\,. \tag{1.2}$$

Here the parameter path $\theta(s) = (Q(s), K(s), V(s))$, $s \in [0, 1]$, encodes the query, key, and value matrices across layers of Transformer blocks and is learned from data.

The contextual measure $\mu_0 = \frac{1}{n}\sum_i \delta_{z^{(i)}}$ can be regarded as an n-sample empirical estimator of an underlying population measure μ_0^∞. The contextual flow map (1.1) then takes this empirical measure as input and produces an output x_1 that approximates an idealized *infinite-context* system, obtained by replacing μ_0 with μ_0^∞ in (1.1). The infinite-context system is a natural reference because an ideal transformer should be able to operate on arbitrarily long context: as the context grows, the contextual measure μ_0 resolves the underlying population μ_0^∞ ever more faithfully, and a model that fully exploits its context should reflect this in its output. The infinite-context system is exactly the object that does so, and the finite-context model should be understood as approximating it. From this perspective, the context length n plays the role of a statistical resource: it is the budget the model is given to estimate the population on which the ideal transformer would act. We are thus led to the central question of the paper:

> *How accurately does a finite-context contextual flow map approximate its infinite-context limit, as a function of the context length n, both at inference time (forward pass) and during training (backward pass)?*

This question is of practical as well as theoretical interest: although recent advances have substantially extended the usable context window of large language models [TGL$^+$24], long contexts remain constrained by computational costs, which in turn limit the emergence of new capabilities. Quantifying the dependence on n therefore informs how aggressively context can be reduced without compromising fidelity to the infinite-context reference. The McKean–Vlasov structure of (1.1) makes it tractable through the classical machinery of propagation of chaos [Kac56, Szn91, CD22a, CD22b].

Related work. The input-output behavior of transformers can be formalized in at least three ways: as a sequence-to-sequence map, a measure-to-measure map, and a contextual map; see Figure 1. The sequence-to-sequence viewpoint models a transformer $T : (\mathbb{R}^d)^n \to (\mathbb{R}^d)^n$ as a map from n input tokens $(x_1, \ldots, x_n)$ to n output tokens $(y_1, \ldots, y_n)$. This early and direct formalization was used to establish approximation results in [YBR$^+$20, ADTK23], albeit with architectures whose size grows exponentially in n. Although this dependence can be alleviated by allowing the embedding dimension to scale with n, the resulting growth in complexity remains restrictive in modern long-context applications. Early measure-theoretic foundations for attention were developed in [VBTdC20], which interpreted self-attention as a system of self-interacting particles and established regularity properties such as Lipschitz continuity under suitable assumptions. This line of work was substantially extended in the seminal paper [SABP22], where transformers

were analyzed as a mean-field interacting particle system implementing a measure-to-measure map. This perspective has enabled the use of powerful tools from partial differential equations and Wasserstein gradient flows; see [GLPR23, GKPR24, BPA25a, BPA25b, CLPR25, CACP25, Rig25, BCL$^+$26, CLPR26, GRR26, ÁLGRB26, AGRB25]. From a statistical viewpoint, a closely related recent contribution is [KS26], which studies transformers as measure-theoretic associative memory and establishes generalization guarantees together with a matching minimax lower bound under spectral assumptions. Despite its usefulness, however, the measure-to-measure viewpoint does not explicitly capture the central role of the distinguished token. This limitation is addressed by the contextual-map formulation introduced in the [FdHP24], which established universal approximation within the class of contextual flow maps using architectures whose size is independent of the context length. The relevance of this contextual framework extends well beyond approximation theory. It provides a natural language for the rapidly growing literature on *in-context learning* with transformers [BO20, XRLM22, GTLV22, BCW$^+$23, WS26, CPS$^+$26, BSS26]. Moreover, it captures the structure of settings in which transformers are explicitly deployed as *contextual velocity fields*, notably in diffusion models [PX23] and measure-to-measure regression [VWP$^+$26]. The present work adopts the same contextual viewpoint, but focuses on statistical questions rather than expressivity alone.

The statistical behavior of transformers as the context length grows has received relatively little attention. Recent works [BB25, BLPV25] study the convergence of a single attention head applied to n i.i.d. tokens to its infinite-token limit. While both establish a parametric $n^{-1/2}$ rate of convergence, the latter draws a more subtle picture near the hardmax limit. Our contribution differs in two major ways. First, we study a deep stack of attention layers using contextual flow maps rather than a single attention layer. In a deep stack the context tokens interact through the shared velocity field $\mathcal{V}$, so they lose their initial independence after the first layer; bounding the deviation from the infinite-context limit therefore requires tracking the growth of correlations across layers. Second, we also study the generalization of training dynamics from finite to infinite context length.

Our analysis is accordingly closer in spirit to the theory of *propagation of chaos* for McKean–Vlasov dynamics. We emphasize that in our setting the term refers to quantitative mean-field stability bounds rather than the convergence of finite marginals to product measures as in [Lac23]. Nevertheless, we adopt it as it has become standard in the machine learning literature [CB18, MMN18, GWB25a, GWB25b, GB26a, GB26b, Chi25, CCM26]. While our work does not require the full apparatus of the theory, we borrow its key structural insight: controlling deviations through regularity of the velocity fields, in both the forward and the backward pass. Uniform-in-time propagation of chaos for the backward pass of two-layer feedforward networks—simple maps $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$—was recently established in [GNS26]. We extend these results to contextual flow maps, which (i) account for depth and (ii) depend on an evolving measure.

Our contributions. We initiate a quantitative statistical theory of contextual flow maps in the long-context regime, treating the context length n as the central statistical resource. Under mild regularity assumptions on the velocity field $\mathcal{V}$, our main results are the following.

- **(C1)** *Forward propagation of chaos.* We establish a quantitative deviation bound between the finite-context contextual flow map (1.1) and its infinite-context limit, holding uniformly along the depth of the network (Theorem 2.1). The analysis explicitly tracks the propagation of correlations introduced at each layer by the shared velocity field—a mechanism that is absent from the single-layer analyses of [BB25, BLPV25].

- **(C2)** *Backward propagation of chaos.* We obtain a deviation bound between the Online Gradient Descent (OGD) training trajectory of a finite-context model and that of its infinite-context counterpart, holding over all iterations of the algorithm (Theorem 2.2). This extends the recent two-layer feedforward result of [GNS26] along two structural axes: *(i)* depth, through a stack of attention layers with parameters shared across tokens, and *(ii)* interaction with an evolving contextual measure that is itself transported across layers.

- **(C3)** *Adjoint equation for contextual flows.* We derive an Eulerian adjoint formulation of the loss gradient (Theorem 3.1 in Section 3) that couples a covector adjoint for the distinguished token to a scalar adjoint test function for the contextual measure. This formulation underlies all of our stability analysis and extends naturally to the Riemannian setting relevant to post-LayerNorm transformers.

(C4) *Stability estimates.* Our results rest on a new set of stability estimates for the contextual flow and the loss gradient (Theorems 4.1 and 4.2 in Section 4), established uniformly in depth and uniformly along the parameter path. We further show in Section 4.3 that the standard transformer architecture—both self-attention and MLP layers—satisfies these estimates with Lipschitz constants that are *independent of the embedding dimension and parameter count*, suggesting a structural reason why transformers train stably across model scales.

Assumptions. Our results rest on two sets of assumptions, stated as Assumptions A and B. The first is purely technical: routine regularity conditions (compactness, smoothness of the various objects entering the definition of CFMs) of the kind required by any analysis of this type. Some can likely be relaxed, but regularity of some form is unavoidable.

The second assumption plays a more substantive role. It is not needed to establish our slow rates $n^{-1/d}$; rather, it is the precise structural ingredient that upgrades these to the parametric rate $n^{-1/2}$. It posits that the velocity field $\mathcal{V}$ driving the CFM (1.1) takes the form

$$\mathcal{V}(x, \mu; \theta) = \mathcal{F}\left(\int_{\mathbb{R}^d} F(x, y; \theta)\, \mathrm{d}\mu(y)\right) + G(x; \theta).$$

for sufficiently regular functions $\mathcal{F}$, F and G, so that μ is applied *linearly*. Crucially, this is precisely the structure of self-attention (1.2) with a multi-layer perceptron (MLP): the transformer-based CFMs that motivate this work satisfy Assumption B *by construction*, and therefore enjoy the fast parametric rate. The full technical statements are deferred to the appendix.

Notation. We denote the Euclidean inner product and norm by $\langle \cdot, \cdot \rangle$ and $|\cdot|$ respectively. The operator norm of a matrix A is denoted by $\|A\| := \sup_{|h|=1} |Ah|$. For a map $\theta : [0,1] \to \mathbb{R}^p$, we use the following norms $\|\theta\|_{L^1} := \int_0^1 |\theta(s)| ds$ and $\|\theta\|_{L^\infty} := \sup_{s\in[0,1]} |\theta(s)|$. The 1-Wasserstein distance [see, e.g., CNWR25, Def. 1.2] between two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ is denoted by $\mathrm{W}_1(\mu, \nu)$. For a velocity field $\mathcal{V}(x, \mu; \theta) : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^p \to \mathbb{R}^d$, $D_x\mathcal{V}(x, \mu; \theta)$ and $D_\theta\mathcal{V}(x, \mu; \theta)$ denote the $d \times d$ (resp. $d \times p$) Jacobian matrix with respect to x (resp. θ) and $\frac{\delta\mathcal{V}}{\delta\mu}[x, \mu; \theta](z)$ denotes the Fréchet derivative with respect to μ and is a vector field in $\mathbb{R}^d$ indexed by z. The Wasserstein Jacobian (a.k.a the Lions derivative) is defined as the (spatial) Jacobian of the Fréchet derivative: $\mathbb{W}_\mu\mathcal{V}[x, \mu; \theta](z) = D_z\left(\frac{\delta\mathcal{V}}{\delta\mu}[x, \mu; \theta](z)\right)$.

2 Main results

We work throughout with the contextual flow map (1.1), which sends an input pair $(x_0, \mu_0) \in \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$ together with a parameter path $\theta : [0,1] \to \mathbb{R}^p$ to the output token $x_1 = x_1(x_0, \mu_0; \theta) \in \mathbb{R}^d$. All results below are stated for the inputs and the velocity field $\mathcal{V}$: (i) under the standard regularity hypotheses collected in Assumption A; or (ii) under the special form of $\mathcal{V}$ specified in Assumption B. We verify in Appendix E that both assumptions are met by the self-attention vector field (1.2) and standard MLP layers with dimension-free constants. We let $\mathbf{A}$ denote the bundle of constants appearing in the assumptions, and write $C = C(\mathbf{A})$ for any positive constant depending only on these parameters (in particular, independent of p, d, and n).

Loss functional. Given a target $y_0 \in \mathbb{R}^d$ and a point loss $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (e.g. squared error or cross-entropy), the prediction loss of a parameter path θ on a sample (x_0, μ_0, y_0) is

$$\mathcal{L}(\theta; \mu_0, x_0, y_0) \;\coloneqq\; \ell(x_1(x_0, \mu_0; \theta),\, y_0)\,. \tag{2.1}$$

Because the dependence of x_1 on the entire path $\theta(\cdot)$ is mediated implicitly through the coupled token–measure system (1.1), computing the Fréchet derivative $\delta\mathcal{L}/\delta\theta$ is non-trivial; an explicit Eulerian-adjoint formula is derived in Appendix B and underlies the regularity analysis below.

To obtain better stability properties, we consider the loss functional with a ridge penalty:

$$\mathcal{L}(\theta; \mu_0, x_0, y_0) + \frac{\lambda}{2}\int_0^1 |\theta(s)|^2\, \mathrm{d}s. \tag{2.2}$$

In Theorem 2.2 below, we establish two results: one for λ small (or zero) and one for λ large. The latter yields stability across all iterations of online gradient descent.

Online gradient descent. We train the model by online gradient descent (OGD)[1] on a stream of samples. At each iteration $k \in \mathbb{N}$, the parameter path is updated using a triple of observations (x_0^k, μ_0^k, y_0^k) according to

$$\theta_{k+1}(s) \;=\; (1-\eta\lambda)\,\theta_k(s) \;-\; \eta\,\frac{\delta\mathcal{L}(\theta_k;\mu_0^k,x_0^k,y_0^k)}{\delta\theta}(s), \qquad s\in[0,1], \tag{2.3}$$

where $\eta > 0$ is a fixed step size and $\lambda \in [0, \eta^{-1})$.

OGD is a faithful idealization of how large transformers are pretrained: a (close to) single pass over an enormous corpus, in which each step sees a new batch that need not be independent of other observations.

2.1 Propagation of chaos for the forward pass

We first analyze the inference-time behavior. Fix a parameter path θ, a distinguished token $x_0 \in \mathbb{R}^d$, and a contextual measure $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$, and let $z^{(1)}, \dots, z^{(n)} \overset{\text{iid}}{\sim} \mu_0$. We compare the contextual flow driven by the population context μ_0 to the one driven by the empirical context

$$\widehat{\mu}_0 \;\coloneqq\; \frac{1}{n}\sum_{i=1}^{n}\delta_{z^{(i)}}, \tag{2.4}$$

denoting the corresponding solutions of (1.1) by (x_s, μ_s) and $(\widehat{x}_s, \widehat{\mu}_s)$.

The central point is that the bound below is *uniform in the depth variable* $s \in [0,1]$, including at output time $s = 1$. Such a result is non-trivial since after the first layer the n context tokens are no longer independent: they are coupled through the shared velocity field $\mathcal{V}$. Controlling the build-up of correlations across layers is precisely what propagation-of-chaos arguments are designed for in the McKean–Vlasov literature [Szn91, CD22a], and it is what we adapt here to the contextual setting. The proof of the next theorem uses a standard Grönwall bound. It does not suffer from a classical exponential deterioration since we assume $s \le 1$.

Theorem 2.1 (Forward propagation of chaos)**.** *Adopt Assumption A and suppose that $d \ge 3$. There exists a constant $C = C(\mathbf{A})$ such that, with probability at least $1 - 4\exp\bigl(-n^{\,1-2/d}\bigr)$,*

$$\sup_{s\in[0,1]} \mathrm{W}_1(\mu_s,\widehat{\mu}_s) \;\le\; C\,n^{-1/d} \qquad \textit{and} \qquad \sup_{s\in[0,1]}\bigl|x_s-\widehat{x}_s\bigr| \;\le\; C\,n^{-1/d}. \tag{2.5}$$

If $\mathcal{V}$ additionally satisfies Assumption B, then for every $\delta \in (0, 1/2)$,

$$\sup_{s\in[0,1]}\bigl|x_s-\widehat{x}_s\bigr| \;\le\; C\sqrt{\frac{\log(\delta^{-1})}{n}}, \tag{2.6}$$

with probability at least $1 - \delta$.

A few comments are in order. First regarding the slow rate (2.5): without further structural assumptions on the CFM such as Assumption B, the rate $n^{-1/d}$ is the best one can hope for, as it is already a bottleneck at the level of the input contextual measure by the Wasserstein law of large numbers [CNWR25, Ch. 2]. The contextual flow therefore does *not* amplify the empirical noise; the finite-context output is statistically as good as the input it consumes. When μ_0 is supported on a submanifold of intrinsic dimension $d' < d$, the same bound holds with $n^{-1/d}$ replaced by $n^{-1/d'}$ [WB19]; this is the practically relevant regime in transformer pretraining, where token embeddings are believed to concentrate on a low-dimensional manifold despite the ambient dimension d being in the hundreds or thousands. The condition $d \ge 3$ is purely cosmetic: the more general Theorem A.2 in Appendix A holds for every $d \ge 1$, with the rate $n^{-1/d}$ adjusted according to the Wasserstein law of large numbers. More generally, these bounds instantiate a pattern recently articulated by Raginsky and Recht [RR26]: the deterministic core of the result is a Wasserstein-stability statement for the flow, into which the data assumption is plugged ex post via a Wasserstein law of large numbers, separating the geometry of the dynamics from the probabilistic assumptions on the data.

The fast/parametric rate (2.6) is quite striking: despite the presence of multiple layers that compromise independence, the parametric rate persists thanks to propagation of chaos across layers for structured CFMs such as transformers. Details are provided in Appendix D.

[1] OGD reduces to Stochastic Gradient Descent if we further assume that the samples are i.i.d.

2.2 Propagation of chaos along online gradient descent

We now turn to the corresponding question on the backward pass, which is significantly more delicate. Let $\{(x_0^k, \mu_0^k, y_0^k)\}_{k\in\mathbb{N}}$ be the population (infinite context) data stream and, at each step k, form the empirical contextual measure

$$\widehat{\mu}_0^k \;\coloneqq\; \frac{1}{n}\sum_{i=1}^{n} \delta_{z^{(k,i)}}. \tag{2.7}$$

where $z^{(k,i)} \sim \mu_0^k$ are independent for all $k \in \mathbb{N}$ and $i \in \{1, 2, \ldots, n\}$. Starting both trajectories from the same path θ_0, we compare the two OGD trajectories produced by (2.3): the *population OGD* $\{\theta_k\}_{k\in\mathbb{N}}$ driven by (x_0^k, μ_0^k, y_0^k), and the *empirical OGD* $\{\widehat{\theta}_k\}_{k\in\mathbb{N}}$ driven by $(x_0^k, \widehat{\mu}_0^k, y_0^k)$ with $\widehat{\theta}_0 = \theta_0$. The statistical quantity of interest is the deviation $\|\theta_k - \widehat{\theta}_k\|_{L^\infty}$.

The behavior of this deviation depends on the regularization parameter λ. Without enough regularization, parameter trajectories may drift apart after a large number of iterations and the deviation can in principle accumulate; strong enough regularization, by contrast, contracts the dynamics at every step and yields a bound at any prescribed iteration.

Theorem 2.2 (Backward propagation of chaos)**.** *Adopt Assumption A and suppose that $d \geq 3$. There exist constants $C = C(\mathbf{A})$, $\eta_c = \eta_c(\mathbf{A})$, and $\lambda_c = \lambda_c(\mathbf{A}) < \eta_c^{-1}$ such that the following hold.*

(i) Small or no ridge penalty. *For $\eta > 0$ and $\lambda \in [0, \eta^{-1})$, we let $k_c = (\log(1 + C\eta))^{-1}$. Then, for any $k \leq k_c$*

$$\left\|\theta_k - \widehat{\theta}_k\right\|_{L^\infty} \;\leq\; C\, n^{-1/d}, \tag{2.8}$$

with probability at least $1 - 2\exp\left(-\eta^{-1} n^{\,1-2/d}\right)$.

If $\mathcal{V}$ additionally satisfies Assumption B, then for every $\delta \in (0, 1/2)$, $k \leq k_c$,

$$\left\|\theta_k - \widehat{\theta}_k\right\|_{L^\infty} \;\leq\; C\,\left(1 + \sqrt{\eta \log(\delta^{-1})}\right)\sqrt{\frac{d}{n}}, \tag{2.9}$$

with probability at least $1 - \delta$.

(ii) Large ridge penalty. *If $\eta < \eta_c$ and $\lambda \in (\lambda_c, \eta^{-1})$, then for every $k \in \mathbb{N}$*

$$\left\|\theta_k - \widehat{\theta}_k\right\|_{L^\infty} \;\leq\; C\, n^{-1/d}, \tag{2.10}$$

with probability at least $1 - 2\exp\left(-\eta^{-1}\, n^{\,1-2/d}\right)$.

If $\mathcal{V}$ additionally satisfies Assumption B, then for every $\delta \in (0, 1/2)$, $k \in \mathbb{N}$,

$$\left\|\theta_k - \widehat{\theta}_k\right\|_{L^\infty} \;\leq\; C\,(1 + \sqrt{\eta \log(\delta^{-1})})\sqrt{\frac{d}{n}}, \tag{2.11}$$

with probability at least $1 - \delta$.

The proofs are provided in Appendix C and Appendix D. Also, three features of this result are worth emphasizing.

Rate. Both bounds (2.8),(2.10) match the inference-time rate $n^{-1/d}$ of Theorem 2.1 (sharpening to $n^{-1/d'}$ on a d'-dimensional submanifold) and $n^{1/2}$ under Assumption B: training incurs no statistical loss beyond, even though the OGD iterates depend on the full history of empirical samples and gradients.

Role of regularization. The dichotomy between (i) and (ii) is genuine. Statement (i) is unconditional but limited to a horizon $k \lesssim 1/\eta$, beyond which Grönwall bounds deteriorate exponentially; statement (ii) trades this for sufficient ridge regularization $\lambda > \lambda_c$, ensuring that the per-step contraction beats the noise injected by replacing μ_0^k with $\widehat{\mu}_0^k$.

Comparison with prior work. The result extends the uniform-in-time backward propagation of chaos of [GNS26] along two structural axes: *depth*, through a stack of attention layers with parameters shared across tokens, and *measure interaction*, through a contextual measure transported by the dynamics. The technical crux is the Lipschitz dependence of $\delta\mathcal{L}/\delta\theta$ on the contextual measure (Section 4), which transfers forward concentration into training stability.

3 The adjoint equation

The proof of Theorem 2.2 hinges on a precise analytical handle on the gradient of the loss with respect to the parameter path θ. Because the dependence of x_1 on $\theta(\cdot)$ is mediated implicitly through the coupled token–measure system (1.1), this gradient does not admit a closed-form chain-rule expression. It is instead given by an *Eulerian adjoint* construction in which two Lagrange multipliers—one for the token dynamics (1.1a) and one for the contextual measure dynamics (1.1b)—propagate the error signal backward through depth. We state the result here and defer the derivation to Appendix B.

Recall from (2.1) that, given a target $y_0 \in \mathbb{R}^d$ and a smooth point loss $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the prediction loss of a parameter path θ on the sample (x_0, μ_0, y_0) is $\mathcal{L}(\theta; \mu_0, x_0, y_0) \;\coloneqq\; \ell\big(x_1(x_0, \mu_0; \theta), y_0\big)$ where x_1 is obtained by integrating (1.1) from depth $s = 0$ to $s = 1$.

Theorem 3.1 (Gradient via adjoint variables)**.** *Adopt Assumption A. Let (x_s, μ_s) solve the contextual flow* (1.1) *with initial condition $(x_0, \mu_0) \in \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$. Then there exist a covector path $p_\cdot : [0,1] \to \mathbb{R}^d$ and a scalar field $\phi_\cdot : [0,1] \to C^{1,1}(\mathbb{R}^d)$ such that the Fréchet derivative of the loss with respect to the parameter path satisfies*

$$\frac{\delta \mathcal{L}(\theta; \mu_0, x_0, y_0)}{\delta \theta}(s) \;=\; D_\theta \mathcal{V}(x_s, \mu_s; \theta(s))^\top \, p_s \;+\; \int_{\mathbb{R}^d} D_\theta \mathcal{V}(z, \mu_s; \theta(s))^\top \, \nabla \phi_s(z) \, \mathrm{d}\mu_s(z), \tag{3.1}$$

where the token adjoint p_s *and the* measure adjoint ϕ_s *solve the backward system*

$$\begin{cases} \dot{p}_s \;=\; -D_x \mathcal{V}(x_s, \mu_s; \theta(s))^\top \, p_s, \\ p_1 \;=\; \nabla_x \ell(x_1, y_0), \end{cases} \tag{3.2}$$

$$\begin{cases} \partial_s \phi_s(z) + \nabla \phi_s(z) \cdot \mathcal{V}(z, \mu_s; \theta(s)) \\ \qquad = \; - p_s \cdot \dfrac{\delta \mathcal{V}}{\delta \mu}[x_s, \mu_s; \theta(s)](z) \;-\; \displaystyle\int_{\mathbb{R}^d} \nabla \phi_s(\xi) \cdot \frac{\delta \mathcal{V}}{\delta \mu}[\xi, \mu_s; \theta(s)](z) \, \mathrm{d}\mu_s(\xi), \\ \phi_1(z) \equiv 0. \end{cases} \tag{3.3}$$

Moreover, the system (3.2)–(3.3) *is well-posed, and there exists a constant $C = C(\mathbf{A})$ such that, for every $s \in [0,1]$ and $z, \widetilde{z} \in \mathbb{R}^d$,*

$$|p_s| \le C, \qquad |\phi_s(z)| \le C(|z| + 1), \qquad |\nabla \phi_s(z)| \le C, \qquad |\nabla \phi_s(z) - \nabla \phi_s(\widetilde{z})| \le C|z - \widetilde{z}|. \tag{3.4}$$

The structure of (3.2)–(3.3) deserves comment. The covector p_s runs backward in depth and tracks the sensitivity of the terminal loss to the position of the distinguished token at depth s; it is a direct generalization of the classical adjoint variable in the neural-ODE literature [CRBD18]. The scalar field ϕ_s plays the analogous role for the contextual measure: it tracks the sensitivity of the loss to perturbations of μ_s at depth s, and its evolution couples back to the token adjoint through the source term involving $\delta \mathcal{V}/\delta \mu$. The $C^{1,1}$-regularity stated in (3.4)—and in particular the Lipschitz continuity of $\nabla \phi_s$—is what ultimately enables Wasserstein-stability bounds on the gradient (3.1): a less regular adjoint test function would not interact stably with empirical contextual measures.

Derivation in one sentence. The adjoint system (3.2)–(3.3) arises by introducing a Lagrangian for the constrained optimization problem of minimizing $\mathcal{L}$ subject to (1.1), with p_s enforcing the token equation and ϕ_s enforcing the continuity equation, and then choosing the multipliers so that the variations of the Lagrangian with respect to x_s and μ_s vanish. The full computation, including the integration-by-parts manipulations that produce the cross terms in (3.3) and the well-posedness analysis of (3.3) via a contraction argument in a weighted C^1 space, is given in Appendix B.

Relation to control theory. Adjoint systems featuring a Wasserstein-derivative term are well-known in the Pontryagin theory of McKean–Vlasov control [CD15, CD18, BFY13] and in the optimal control of continuity equations [BF21]. The structure of (3.2)–(3.3) is closest in spirit to the major–minor mean-field formalism [Hua10, CZ16], in which a distinguished agent interacts with a continuum population. Our contribution is not only the specific adjoint equation for contextual flow maps but also the stability estimates of Section 4 which are absent from the control literature.

Riemannian extension. The Eulerian Lagrange-multiplier argument is geometrically transparent and extends without modification to settings where the contextual flow is constrained to a Riemannian submanifold $\mathbf{M} \subseteq \mathbb{R}^d$—for example, the unit sphere $\mathbb{S}^{d-1}$ relevant to post-LayerNorm transformers as in [GLPR25]. In that case p_s becomes a tangent vector in $T_{x_s}\mathbf{M}$, the spatial Jacobian $D_x\mathcal{V}$ is replaced by the covariant derivative $\nabla^{\mathbf{M}}\mathcal{V}$, and (3.2) is modified accordingly.

4 Stability estimates

The forward and backward propagation-of-chaos bounds of Section 2 ultimately reduce to a number of *stability estimates*: quantitative statements asserting that the contextual flow (1.1), the adjoint system (3.2)–(3.3), and the loss gradient (3.1) all depend Lipschitz-continuously on their inputs, with constants that are uniform in depth. We collect the two estimates that drive the analysis in this section, and outline how they imply the main theorems. Full proofs are deferred to Appendices A and B.

4.1 Stability of the contextual flow

Let (x_s, μ_s) and $(\widetilde{x}_s, \widetilde{\mu}_s)$ denote the solutions of (1.1) associated with two parameter paths $\theta, \widetilde{\theta}$ and two initial conditions (x_0, μ_0) and $(\widetilde{x}_0, \widetilde{\mu}_0)$ in $\mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$, respectively.

Theorem 4.1 (Stability of the contextual flow)**.** *Adopt Assumption A. There exists $C = C(\mathbf{A})$ such that, for every $s \in [0,1]$,*

$$\mathrm{W}_1(\mu_s, \widetilde{\mu}_s) \;\le\; C\big(\mathrm{W}_1(\mu_0, \widetilde{\mu}_0) \;+\; \|\theta - \widetilde{\theta}\|_{L^1}\big), \tag{4.1}$$

$$|x_s - \widetilde{x}_s| \;\le\; C\big(\mathrm{W}_1(\mu_0, \widetilde{\mu}_0) \;+\; |x_0 - \widetilde{x}_0| \;+\; \|\theta - \widetilde{\theta}\|_{L^1}\big). \tag{4.2}$$

The proof, given in Appendix A, follows the classical Grönwall pattern, lifting the difference between flows to a coupling of trajectories along their common characteristics. The key point is that the time horizon is fixed at $s = 1$, which prevents the exponential blow-up typical of long-time Grönwall arguments and yields dimension-free constants.

Theorem 4.1 immediately yields the forward propagation of chaos bound: instantiating it with the empirical contextual measure $\widehat{\mu}_0 = \frac{1}{n}\sum_{i=1}^n \delta_{z^{(i)}}$ in place of $\widetilde{\mu}_0$, with the same parameter path and the same initial token, gives

$$\sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}_s) \;+\; \sup_{s\in[0,1]} |x_s - \widehat{x}_s| \;\le\; C\,\mathrm{W}_1(\mu_0, \widehat{\mu}_0),$$

and Theorem 2.1 then follows by combining this deterministic bound with the Wasserstein law of large numbers and a McDiarmid-type concentration inequality (Appendix A).

4.2 Stability of the loss gradient

The backward analysis requires substantially more: it requires that the loss gradient (3.1) inherit the stability of the forward flow. This is the most technically demanding ingredient of the paper, because the gradient depends on the entire trajectory $(x_s, \mu_s, p_s, \phi_s)$ of the coupled forward–adjoint system.

Theorem 4.2 (Stability of the loss gradient)**.** *Adopt Assumption A. There exists $C = C(\mathbf{A})$ such that, for every $s \in [0,1]$,*

$$\begin{aligned}\left|\frac{\delta\mathcal{L}(\theta;\mu_0,x_0,y_0)}{\delta\theta}(s) \;-\; \frac{\delta\mathcal{L}(\widetilde{\theta};\widetilde{\mu}_0,\widetilde{x}_0,\widetilde{y}_0)}{\delta\theta}(s)\right| \\ \le\; C\Big(\mathrm{W}_1(\mu_0,\widetilde{\mu}_0) + |x_0-\widetilde{x}_0| + |y_0-\widetilde{y}_0| + \|\theta-\widetilde{\theta}\|_{L^1} + |\theta(s)-\widetilde{\theta}(s)|\Big).\end{aligned} \tag{4.3}$$

The bound is uniform in $s \in [0,1]$ and *linear* in each of the input perturbations. The proof proceeds in three steps, all carried out in Appendix B. *(i)* Theorem 4.1 provides the stability of the forward trajectory (x_s, μ_s). *(ii)* The token adjoint p_s obeys a linear ODE whose coefficients depend on $(x_s, \mu_s, \theta(s))$, so a Grönwall argument transfers the stability of the forward flow into a stability bound for p_s. *(iii)* The measure adjoint ϕ_s is the most delicate: it solves a transport equation with a non-local source term, and its regularity is established by a fixed-point argument in a weighted C^1 space (the

same Banach space used to prove well-posedness in Theorem 3.1); the contraction constant degrades by a controlled amount under perturbation of the data, yielding a Lipschitz bound on $(\phi_s, \nabla\phi_s)$.

The Lipschitz dependence in (4.3) is exactly what is needed to transfer the forward concentration of Theorem 2.1 into the stability of the OGD trajectory. Indeed, applying Theorem 4.2 to the population update $\theta_{k+1} = (1-\eta\lambda)\theta_k - \eta\,\delta\mathcal{L}(\theta_k; \mu_0^k, x_0^k, y_0^k)/\delta\theta$ and its empirical counterpart driven by $\widehat{\mu}_0^k$ produces the recursion

$$\|\theta_{k+1} - \widehat{\theta}_{k+1}\|_{L^\infty} \;\le\; \big(1-\eta(\lambda - C)\big)\,\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \;+\; \eta\, C\, \mathrm{W}_1(\mu_0^k, \widehat{\mu}_0^k).$$

In the regime $\lambda > \lambda_c$, the per-step contraction $1-\eta(\lambda - C)$ is strictly less than 1, and iterating the recursion yields the bound (2.10) in Theorem 2.2; in the regime with small or no regularization, iterating only over $k \lesssim 1/\eta$ steps produces the finite-horizon bound (2.8).

4.3 Stability estimates for transformers

All the estimates above are stated under the abstract regularity assumptions collected in Assumption A on the velocity field $\mathcal{V}$. In Appendix E, we verify that the concrete self-attention vector field (1.2), together with standard MLP layers, satisfies these assumptions. Crucially, all resulting Lipschitz constants (and higher-order regularity constants) are *independent of the embedding dimension d and the parameter dimension p*—i.e., they are truly dimension-free. Our analysis builds upon prior work on the smoothness of attention. For example, [CAP24, FMS26] established Lipschitz continuity of the attention mechanism with respect to the input tokens (corresponding to $D_x\mathcal{V}$ and $\mathbb{W}_\mu\mathcal{V}$ in our setting). In addition, we establish full Lipschitz continuity (together with the necessary higher-order regularity) with respect to *both* the tokens *and* the trainable parameter matrices Q, K, V themselves. Combined with the dimension-free nature of our constants, all the rates stated in Theorem 2.1 and Theorem 2.2 hold for realistic Transformer architectures with absolute constants; the only unavoidable dependence is on the input-support radius R and the parameter-path bound M.

5 Assumptions

We collect here the standard regularity hypotheses on the velocity field $\mathcal{V}$, the loss function ℓ, and the input data that are referenced throughout the paper. Different results require different subsets of these hypotheses; we make this explicit in each statement by citing the bundle **A** of constants involved.

Assumption A (Regularity)**.** We collect the following regularity conditions on the input data, the loss ℓ, and the velocity field $\mathcal{V}$:

(A1) *(Data and initialization).* The contextual measure μ_0 has compact support, and x_0, y_0 are uniformly bounded: there exist $R, M > 0$ such that $\operatorname{supp}\mu_0 \subseteq B_R(0)$, $|x_0| \le R$, and $|y_0| \le R$. The initial parameter is uniformly bounded and Lipschitz: $|\theta(s)| \le M$ and $|\theta(s) - \theta(t)| \le M|s-t|$ for all $s, t \in [0,1]$.

(A2) *(Smoothness of the loss).* There exist $M_\ell, L_\ell > 0$ such that, for all $x, \widetilde{x}, y, \widetilde{y} \in \mathbb{R}^d$,

$$|\nabla_x\ell(x,y)| \le M_\ell(|x| + |y| + 1), \qquad |\nabla_x\ell(x,y) - \nabla_x\ell(\widetilde{x},\widetilde{y})| \le L_\ell\big(|x-\widetilde{x}| + |y-\widetilde{y}|\big).$$

(A3) *(Uniform Lipschitz continuity of $\mathcal{V}$).* There exists $L_0 > 1$ such that, for all $x, \widetilde{x}, z, \widetilde{z} \in \mathbb{R}^d$, $\mu, \widetilde{\mu} \in \mathcal{P}(\mathbb{R}^d)$, and $\theta, \widetilde{\theta} \in \mathbb{R}^p$,

$$|\mathcal{V}(x,\mu;\theta) - \mathcal{V}(\widetilde{x},\widetilde{\mu};\widetilde{\theta})| \le L_0\big(|x-\widetilde{x}| + \mathrm{W}_1(\mu,\widetilde{\mu}) + |\theta-\widetilde{\theta}|\big),$$
$$\left|\tfrac{\delta\mathcal{V}}{\delta\mu}[x,\mu;\theta](z) - \tfrac{\delta\mathcal{V}}{\delta\mu}[\widetilde{x},\widetilde{\mu};\widetilde{\theta}](\widetilde{z})\right| \le L_0\big(|x-\widetilde{x}| + \mathrm{W}_1(\mu,\widetilde{\mu}) + |\theta-\widetilde{\theta}| + |z-\widetilde{z}|\big).$$

Furthermore, there exists $M_0 > 1$ such that

$$|\mathcal{V}(x,\mu;\theta)| \le M_0(|x| + |\theta| + 1), \qquad \left|\tfrac{\delta\mathcal{V}}{\delta\mu}[x,\mu;\theta](z)\right| \le M_0(|x| + |\theta| + |z| + 1).$$

(A4) *(Smoothness of the derivatives).* The Jacobians $D_x\mathcal{V}$ and $D_\theta\mathcal{V}$, and the Wasserstein Jacobian $\mathbb{W}_\mu\mathcal{V}$, are uniformly bounded and Lipschitz continuous with respect to (x, μ, θ): there exist

$L_1, L_2, L_3 > 0$ such that, for all $x, \widetilde{x}, z, \widetilde{z} \in \mathbb{R}^d$, $\mu, \widetilde{\mu} \in \mathcal{P}(\mathbb{R}^d)$, and $\theta, \widetilde{\theta} \in \mathbb{R}^p$,

$$\|D_x\mathcal{V}(x,\mu;\theta) - D_x\mathcal{V}(\widetilde{x},\widetilde{\mu};\widetilde{\theta})\| \le L_1\big(|x-\widetilde{x}| + \mathrm{W}_1(\mu,\widetilde{\mu}) + |\theta-\widetilde{\theta}|\big),$$
$$\|\mathbb{W}_\mu\mathcal{V}(x,\mu;\theta)(z) - \mathbb{W}_\mu\mathcal{V}(\widetilde{x},\widetilde{\mu};\widetilde{\theta})(\widetilde{z})\| \le L_2\big(|x-\widetilde{x}| + \mathrm{W}_1(\mu,\widetilde{\mu}) + |\theta-\widetilde{\theta}| + |z-\widetilde{z}|\big),$$
$$\|D_\theta\mathcal{V}(x,\mu;\theta) - D_\theta\mathcal{V}(\widetilde{x},\widetilde{\mu};\widetilde{\theta})\| \le L_3\big(|x-\widetilde{x}| + \mathrm{W}_1(\mu,\widetilde{\mu}) + |\theta-\widetilde{\theta}|\big).$$

We write $\mathbf{A} = \{R, M, M_\ell, L_\ell, L_0, M_0, L_1, L_2, L_3\}$ for the bundle of constants in Assumption A and $C = C(\mathbf{A})$ for any positive constant depending only on these parameters; in particular, they are independent of the embedding dimension d, the parameter dimension p, and the context length n.

For $\mathcal{V}$ with the following form, we can obtain the parametric rate $n^{-1/2}$ in Theorem 2.1 and Theorem 2.2.

Assumption B. The vector field $\mathcal{V}(x,\mu;\theta)$ has the form

$$\mathcal{V}(x,\mu;\theta) = \mathcal{F}\left(\int_{\mathbb{R}^d} F(x,y;\theta)\,\mathrm{d}\mu(y)\right) + G(x;\theta), \tag{5.1}$$

where $F(x,y;\theta) : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^m$ for some $m \in \mathbb{Z}_+$, $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^d$, and $G(x;\theta) : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$. Also, there exists an $L_4 > 0$, such that for all $x, y \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^p$,

$$|F(x,y;\theta)| \le L_4(|x|+|y|+|\theta|+1), \quad \|DF(x,y;\theta)\| \le L_4, \quad \|D^2F(x,y;\theta)\| \le L_4, \tag{5.2}$$

and

$$|G(x;\theta)| \le L_4(|x|+|\theta|+1), \quad \|DG(x;\theta)\| \le L_4, \quad \|D^2G(x;\theta)\| \le L_4, \tag{5.3}$$

Here, D, D^2 denotes the Jacobian and the Hessian of F, G with respect to x, y, θ. Also, for all $w \in \mathbb{R}^m$,

$$|\mathcal{F}(w)| \le L_4(|w|+1), \quad \|D\mathcal{F}(w)\| \le L_4, \quad \|D^2\mathcal{F}(w)\| \le L_4. \tag{5.4}$$

If the activation function in the MLP layer is ReLU, it is not C^2-regular as required by Assumption B. Nevertheless, the theory developed in this paper also applies to this case. See Remark E.8 for further discussion.

Remark on localization. It suffices for Assumptions (A3)–(A4) and Assumption B to hold on the compact sets furnished by (A1): our proofs invoke $\mathcal{V}$ and its derivatives only along the trajectories of the contextual flow (1.1) and along the OGD iterates (2.3). We establish the requisite uniform-in-depth boundedness of both in Lemma C.1 and in Theorem A.1 below. In particular, when $\mathcal{V}$ is a transformer block (a self-attention layer or an MLP), Assumptions (A3)–(A4) and Assumption B are derived from the single global hypothesis (A1) with constants that are dimension-free (Appendix E).

6 Conclusion

We introduced *contextual flow maps* as a unifying abstraction for analyzing transformers and developed a quantitative statistical theory in which the context length n is the central statistical resource. Within this framework, we established forward and backward propagation-of-chaos bounds at the Wasserstein rate $n^{-1/d}$, uniformly along the depth of the network and, in the regularized regime, uniformly along the iterates of online gradient descent. To our knowledge, these are the first such guarantees for a deep stack of attention layers, where tokens lose their initial independence after the first layer through the shared velocity field. Under the additional structural condition that contextual measure μ is applied linearly when computing the vector field $\mathcal{V}$—a condition satisfied by self-attention—we sharpen these bounds to the parametric rate $n^{-1/2}$, both for the distinguished token and along the OGD trajectory. Identifying weaker structural conditions under which the parametric rate persists is a natural direction for future work.

A central technical ingredient, of independent interest, is the Eulerian adjoint equation of Section 3. In contrast to the classical neural-ODE adjoint, which features a single covector for the distinguished token, the contextual setting requires *two* coupled adjoint variables: a covector p_s tracking sensitivity to the token, and a scalar test field ϕ_s tracking sensitivity to the contextual measure. The $C^{1,1}$-regularity of ϕ_s is exactly what allows Wasserstein stability of the loss gradient, and we expect this two-variable construction to be useful for stability analysis of trained transformers and for the Riemannian extension to post-LayerNorm architectures.

Acknowledgment

PR is partially supported by NSF grant DMS-2509011.

References

[ADTK23] Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient transformers. In *TAG-ML*, 2023.

[AGRB25] Albert Alcalde, Borjan Geshkovski, and Domènec Ruiz-Balet. Attention's forward pass and Frank–Wolfe, 2025. arXiv:2508.09628.

[ÁLGRB26] Antonio Álvarez-López, Borjan Geshkovski, and Domènec Ruiz-Balet. Perceptrons and localization of attention's mean-field landscape, 2026. arXiv:2601.21366.

[BB25] Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective, 2025. arXiv:2512.11784.

[BCL$^+$26] Giuseppe Bruno, Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Scaling limits of long-context transformers, 2026. arXiv:2605.08505.

[BCW$^+$23] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *NeurIPS*, 2023.

[BF21] Benoît Bonnet and Hélène Frankowska. Necessary optimality conditions for optimal control problems in Wasserstein spaces. *Applied Mathematics & Optimization*, 84(Suppl. 2):S1281–S1330, 2021.

[BFY13] Alain Bensoussan, Jens Frehse, and Phillip Yam. *Mean Field Games and Mean Field Type Control Theory*. SpringerBriefs in Mathematics. Springer, New York, 2013.

[BLPV25] Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and François-Xavier Vialard. Token sample complexity of attention, 2025. arXiv:2512.10656.

[BO20] Tom B. Brown and Others. Language models are few-shot learners. In *NeurIPS*, 2020.

[BPA25a] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *ICLR*, 2025.

[BPA25b] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In *NeurIPS*, 2025.

[BSS26] Nicholas Barnfield, Subhabrata Sen, and Pragya Sur. Multi-layer cross-attention is provably optimal for multi-modal in-context learning, 2026. arXiv:2602.04872.

[CACP25] Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers, 2025. arXiv:2501.18322.

[CAP24] Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *ICML*, 2024.

[CB18] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NeurIPS*, 2018.

[CCM26] Louis-Pierre Chaintron, Lénaïc Chizat, and Javier Maas. Resnets of all shapes and sizes: Convergence of training dynamics in the large-scale limit, 2026. arXiv:2603.18168.

[CD15] René Carmona and François Delarue. Forward-backward stochastic differential equations and controlled McKean-Vlasov dynamics. *The Annals of Probability*, 43(5):2647–2700, 2015.

[CD18] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I–II*, volume 83–84 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2018.

[CD22a] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. I. Models and methods. *Kinetic and Related Models*, 15(6):895–1015, 2022.

[CD22b] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. II. Applications. *Kinetic and Related Models*, 15(6):1017–1173, 2022.

[Chi25] Lénaïc Chizat. The hidden width of deep resnets: Tight error bounds and phase diagrams, 2025. arXiv:2509.10167.

[CLPR25] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models, 2025. arXiv:2504.14697.

[CLPR26] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *ICLR*, 2026.

[CNWR25] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*, volume 2364 of *Lecture Notes in Mathematics*. Springer, Cham, 2025.

[CPS+26] Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax-optimal in-context nonparametric regression with transformers. In *ICML*, 2026.

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[CZ16] René Carmona and Xiuneng Zhu. A probabilistic approach to mean field games with major and minor players. *The Annals of Applied Probability*, 26(3):1535–1580, 2016.

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[E17] Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5:1 – 11, 2017.

[FdHP24] Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *ICLR*, 2024.

[FMS26] Takashi Furuya, Davide Murari, and Carola-Bibiane Schönlieb. Approximation theory for Lipschitz-continuous transformers, 2026. arXiv:2602.15503.

[GB26a] Margalit Glasgow and Joan Bruna. Quantitative propagation of chaos in mean-field NNs: Sufficient conditions, lower bounds, and uniform-in-time guarantees, 2026.

[GB26b] Margalit Glasgow and Joan Bruna. Uniform-in-time weak propagation-of-chaos in shallow neural networks, 2026. arXiv:2605.22010.

[GKPR24] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model, 2024. arXiv:2410.06833.

[GLPR23] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *NeurIPS*, 2023.

[GLPR25] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

[GNS26] Arnaud Guillin, Boris Nectoux, and Paul Stos. Uniform-in-time concentration in two-layer neural networks via transportation inequalities, 2026. arXiv:2603.01842.

[GRR26] Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using Transformers, 2026. arXiv:2411.04551.

[GTLV22] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *NeurIPS*, 2022.

[GWB25a] Margalit Glasgow, Denny Wu, and Joan Bruna. Mean-field analysis of polynomial-width two-layer neural network beyond finite time horizon. In *COLT*, 2025.

[GWB25b] Margalit Glasgow, Denny Wu, and Joan Bruna. Propagation of chaos in one-hidden-layer neural networks beyond logarithmic time, 2025. arXiv:2504.13110.

[HR18] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2018.

[Hua10] Minyi Huang. Large-population LQG games involving a major player: The Nash certainty equivalence principle. *SIAM Journal on Control and Optimization*, 48(5):3318–3353, 2010.

[Kac56] Mark Kac. Foundations of kinetic theory. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 171–197, Berkeley, 1956. University of California Press.

[Kos08] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, 2008.

[KS26] Ryotaro Kawata and Taiji Suzuki. Transformers as measure-theoretic associative memory: A statistical perspective and minimax optimality, 2026. arXiv:2602.01863.

[Lac23] Daniel Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *Probability and Mathematical Physics*, 4(2):377–432, 2023.

[MMN18] 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.

[PX23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[Rig25] Philippe Rigollet. The mean-field dynamics of transformers, 2025. arXiv:2512.01868.

[RR26] Maxim Raginsky and Benjamin Recht. Separating geometry from probability in the analysis of generalization, 2026. arXiv:2604.19560.

[SABP22] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *AISTATS*, 2022.

[Szn91] Alain-Sol Sznitman. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX — 1989*, volume 1464 of *Lecture Notes in Mathematics*, pages 165–251. Springer, Berlin, Heidelberg, 1991.

[TGL+24] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. arXiv:2403.05530.

[VBTdC20] James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention, 2020. arXiv:2007.02876.

[VDVW96] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes: with applications to statistics*, pages 16–28. Springer, 1996.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[VWP+26] Matthew Vandergrift, Martha White, Yury Polyanskiy, Philippe Rigollet, and Lazar Atanackovic. Measure-to-measure regression with transformers, 2026. arXiv:2605.28075.

[WB19] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[WS26] Tomoya Wakayama and Taiji Suzuki. In-context learning is provably Bayesian inference: A generalization theory for meta-learning. In *ICML*, 2026.

[XRLM22] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *ICLR*, 2022.

[YBR+20] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *ICLR*, 2020.

A Stability of the contextual flow: Proofs of Theorem 4.1 and 2.1

This appendix proves the stability bound for the forward flow stated in Theorem 4.1, establishes the corresponding concentration result for empirical contextual measures, and combines the two to deduce Theorem 2.1 under Assumption A.

A.1 Stability of the contextual flow

We restate Theorem 4.1 in a slightly stronger form that also records the uniform-in-depth boundedness of (x_s, μ_s). Throughout the section, $\theta, \widetilde{\theta} : [0,1] \to \mathbb{R}^p$ are two parameter paths, $\mu_0, \widetilde{\mu}_0 \in \mathcal{P}(\mathbb{R}^d)$ are two initial measures, and $x_0, \widetilde{x}_0 \in \mathbb{R}^d$ are two initial tokens, all satisfying (A1). The corresponding solutions to (1.1) are denoted (x_s, μ_s) and $(\widetilde{x}_s, \widetilde{\mu}_s)$.

Theorem A.1 (Stability of the contextual flow—precise version)**.** *Adopt Assumptions* (A1) *and* (A3)*. There exists $C = C(\mathbf{A})$ such that, for all $s \in [0,1]$,*

$$\operatorname{supp} \mu_s \subseteq B_C(0), \qquad |x_s| \le C, \tag{A.1}$$

$$\mathrm{W}_1(\mu_s, \widetilde{\mu}_s) \le C\big(\mathrm{W}_1(\mu_0, \widetilde{\mu}_0) + \|\theta - \widetilde{\theta}\|_{L^1}\big), \tag{A.2}$$

$$|x_s - \widetilde{x}_s| \le C\big(\mathrm{W}_1(\mu_0, \widetilde{\mu}_0) + |x_0 - \widetilde{x}_0| + \|\theta - \widetilde{\theta}\|_{L^1}\big). \tag{A.3}$$

The proof rests on the characteristic ODEs associated with (1.1): let $\Phi_s, \widetilde{\Phi}_s : \mathbb{R}^d \to \mathbb{R}^d$ solve

$$\begin{cases} \partial_s \Phi_s(x) = \mathcal{V}(\Phi_s(x), \mu_s; \theta(s)), \\ \Phi_0(x) = x, \end{cases} \qquad \begin{cases} \partial_s \widetilde{\Phi}_s(x) = \mathcal{V}(\widetilde{\Phi}_s(x), \widetilde{\mu}_s; \widetilde{\theta}(s)), \\ \widetilde{\Phi}_0(x) = x. \end{cases} \tag{A.4}$$

Then $x_s = \Phi_s(x_0)$ and $\mu_s = (\Phi_s)_{\#}\mu_0$, and analogously for the tilded quantities.

Proof of (A.1). By the linear-growth bound in (A3) and (A1),

$$\tfrac{\mathrm{d}}{\mathrm{d}s}|x_s| \le |\partial_s \Phi_s(x_0)| = |\mathcal{V}(x_s, \mu_s; \theta(s))| \le M_0(|x_s| + M + 1),$$

and Grönwall's inequality yields $|x_s| \le (R + M + 1)e^{M_0} + (M+1)$. The same argument applied uniformly over $x \in \operatorname{supp} \mu_0$ and the identity $\mu_s = (\Phi_s)_{\#}\mu_0$ give the support bound. □

Proof of (A.2). Let γ_0 be any coupling of μ_0 and $\widetilde{\mu}_0$. Then $(\Phi_s, \widetilde{\Phi}_s)_{\#}\gamma_0$ is a coupling of μ_s and $\widetilde{\mu}_s$, so

$$\mathrm{W}_1(\mu_s, \widetilde{\mu}_s) \le \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y| \, \mathrm{d}(\Phi_s, \widetilde{\Phi}_s)_{\#}\gamma_0(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_s(x) - \widetilde{\Phi}_s(y)| \mathrm{d}\gamma_0(x,y). \tag{A.5}$$

Differentiating in s and using (A.4) and (A3),

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}s} \int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_s(x) - \widetilde{\Phi}_s(y)| \mathrm{d}\gamma_0(x,y) \le \int_{\mathbb{R}^d \times \mathbb{R}^d} |\partial_s \Phi_s(x) - \partial_s \widetilde{\Phi}_s(y)| \mathrm{d}\gamma_0(x,y) \\ &\le L_0 \int_{\mathbb{R}^d \times \mathbb{R}^d} \Big(|\Phi_s(x) - \widetilde{\Phi}_s(y)| + \mathrm{W}_1(\mu_s, \widetilde{\mu}_s) + |\theta(s) - \widetilde{\theta}(s)| \Big) \mathrm{d}\gamma_0(x,y) \\ &\le 2L_0 \int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_s(x) - \widetilde{\Phi}_s(y)| \mathrm{d}\gamma_0(x,y) + L_0 |\theta(s) - \widetilde{\theta}(s)|, \end{aligned}$$

where the last step uses (A.5). Grönwall's inequality yields, for $s \in [0,1]$,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_s(x) - \widetilde{\Phi}_s(y)| \mathrm{d}\gamma_0(x,y) \le e^{2L_0 s} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y| \mathrm{d}\gamma_0(x,y) + L_0 \|\theta - \widetilde{\theta}\|_{L^1} \right).$$

Taking the infimum over couplings γ_0 gives (A.2). □

Proof of (A.3). Since $|x_s - \widetilde{x}_s| = |\Phi_s(x_0) - \widetilde{\Phi}_s(\widetilde{x}_0)|$, applying (A3) and (A.2) gives

$$\tfrac{\mathrm{d}}{\mathrm{d}s}|x_s - \widetilde{x}_s| \le L_0\big(|x_s - \widetilde{x}_s| + \mathrm{W}_1(\mu_s, \widetilde{\mu}_s) + |\theta(s) - \widetilde{\theta}(s)|\big).$$

Substituting (A.2) into the right-hand side and applying Grönwall yields (A.3). □

Theorem 4.1 follows directly from the estimates (A.2)–(A.3) in Theorem A.1.

A.2 Concentration of the empirical contextual flow

We now turn to the central random object. Let $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ satisfy (A1), draw $z^{(1)}, \dots, z^{(n)} \overset{\text{iid}}{\sim} \mu_0$, and form the empirical measure

$$\widehat{\mu}_0 \;\coloneqq\; \frac{1}{n}\sum_{i=1}^{n} \delta_{z^{(i)}}. \tag{A.6}$$

We compare the contextual flow (1.1) starting from μ_0 with the one starting from $\widehat{\mu}_0$, both evolved with the same parameter path θ and the same initial token x_0. The corresponding solutions are denoted (x_s, μ_s) and $(\widehat{x}_s, \widehat{\mu}_s)$.

Theorem A.2 (Concentration of the empirical contextual flow)**.** *Adopt Assumptions* (A1) *and* (A3)*. Suppose $d \geq 3$. There exist constants $C, C_1, \alpha > 0$ depending on* **A** *such that*

$$\mathbb{E}\Big[\sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}_s)\Big] \;\leq\; Cn^{-1/d}, \qquad \mathbb{E}\Big[\sup_{s\in[0,1]} |x_s - \widehat{x}_s|\Big] \;\leq\; Cn^{-1/d}. \tag{A.7}$$

For every $\varepsilon > 0$,

$$\mathbb{P}\Big(\big|\sup_s \mathrm{W}_1(\mu_s, \widehat{\mu}_s) - \mathbb{E}[\sup_s \mathrm{W}_1(\mu_s, \widehat{\mu}_s)]\big| > \varepsilon\Big) \leq 2\exp(-\alpha n\varepsilon^2), \tag{A.8}$$

$$\mathbb{P}\Big(\big|\sup_s |x_s - \widehat{x}_s| - \mathbb{E}[\sup_s |x_s - \widehat{x}_s|]\big| > \varepsilon\Big) \leq 2\exp(-\alpha n\varepsilon^2). \tag{A.9}$$

Consequently, for every $\beta > 0$, with probability at least $1 - 4e^{-\beta}$,

$$\sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}_s) \;+\; \sup_{s\in[0,1]} |x_s - \widehat{x}_s| \;\leq\; Cn^{-1/d} + C_1\beta^{1/2}n^{-1/2}. \tag{A.10}$$

The proof rests on two classical facts that we record here for the reader's convenience.

Lemma A.3 (Wasserstein law of large numbers)**.** *Let $d \geq 3$ and let $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ with $\operatorname{supp}\mu_0 \subseteq B_R(0)$. There exists $C = C(R) > 0$ such that*

$$\mathbb{E}[\mathrm{W}_1(\mu_0, \widehat{\mu}_0)] \leq Cn^{-1/d}.$$

For $d = 1$ the rate is $n^{-1/2}$; for $d = 2$ it is $n^{-1/2}\log n$. If μ_0 is supported on a d'-dimensional submanifold of $\mathbb{R}^d$, the rate becomes $n^{-1/d'}$ [WB19].

Lemma A.4 (McDiarmid's inequality)**.** *Let $\xi_1, \dots, \xi_N$ be independent random variables in $\mathbb{R}^q$, and let $f : (\mathbb{R}^q)^N \to \mathbb{R}$ satisfy: for each i, there exists $c_i > 0$ such that*

$$|f(\dots, x_i, \dots) - f(\dots, \widetilde{x}_i, \dots)| \leq c_i$$

whenever the two arguments differ only in the i-th coordinate. Then, for every $\varepsilon > 0$,

$$\mathbb{P}\big(|f(\xi_1, \dots, \xi_N) - \mathbb{E}f(\xi_1, \dots, \xi_N)| \geq \varepsilon\big) \leq 2\exp\Big(-\tfrac{2\varepsilon^2}{\sum_{i=1}^N c_i^2}\Big).$$

Proof of Theorem A.2. Theorem A.1, applied with $\widetilde{\mu}_0 = \widehat{\mu}_0$, $\widetilde{x}_0 = x_0$, and $\widetilde{\theta} = \theta$, yields the deterministic bounds

$$\sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}_s) \;+\; \sup_{s\in[0,1]} |x_s - \widehat{x}_s| \;\leq\; C\,\mathrm{W}_1(\mu_0, \widehat{\mu}_0). \tag{A.11}$$

Taking expectations and applying Lemma A.3 gives (A.7).

For the concentration claims, view $f(z^{(1)}, \dots, z^{(n)}) \coloneqq \sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}_s)$ as a function of the i.i.d. sample. Replacing one sample $z^{(i)}$ by an independent copy $\widetilde{z}^{(i)}$ changes the empirical measure to $\widehat{\mu}'_s$. Because

$$\left|\sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}_s) - \sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}'_s)\right| \leq \sup_{s\in[0,1]} |\mathrm{W}_1(\mu_s, \widehat{\mu}_s) - \mathrm{W}_1(\mu_s, \widehat{\mu}'_s)| \leq \sup_{s\in[0,1]} \mathrm{W}_1(\widehat{\mu}_s, \widehat{\mu}'_s),$$

which is at most $C\mathrm{W}_1(\widehat{\mu}_0, \widehat{\mu}'_0) \leq C2R/n$ by (A.2). So, the value of f changes by at most $C \cdot 2R/n$. McDiarmid's inequality (Lemma A.4) with $c_i = 2RC/n$ gives $\sum c_i^2 \leq 4R^2C^2/n$, and (A.8) follows with $\alpha = 1/(2R^2C^2)$. The same argument applied to $|x_s - \widehat{x}_s|$ yields (A.9), and combining mean and concentration gives (A.10). □

A.3 Proof of Theorem 2.1

(2.5) in Theorem 2.1 follows immediately from Theorem A.2 by choosing $\beta = n^{1-2/d}$. Indeed, $C_1\beta^{1/2}n^{-1/2} = C_1 n^{-1/d}$, so (A.10) delivers

$$\sup_{s\in[0,1]} \mathrm{W}_1(\mu_s, \widehat{\mu}_s) + \sup_{s\in[0,1]} |x_s - \widehat{x}_s| \;\le\; (C + C_1)\, n^{-1/d}$$

with probability at least $1 - 4\exp(-n^{1-2/d})$. The case $d = 1, 2$ is recovered from the corresponding statement of Lemma A.3.

(2.6) in Theorem 2.1 follows from the later Theorem D.2 by choosing β such that $2e^{-\beta} = \delta$. □

B The adjoint system: Proofs of Theorem 3.1 and 4.2

This appendix derives the gradient formula (3.1) together with the adjoint equations (3.2)–(3.3) (Theorem 3.1 in the body), establishes the well-posedness, regularity, and stability of the adjoint variables, and proves the stability of the loss gradient (Theorem 4.2 in the body).

B.1 Eulerian derivation of the gradient formula

We use the Eulerian Lagrangian approach: introduce two multipliers, $p_s : [0,1] \to \mathbb{R}^d$ for the token equation and $\phi_s : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ for the continuity equation, and choose them so that the variation of the Lagrangian with respect to (x_s, μ_s) vanishes identically. The remaining variation, with respect to θ, yields (3.1).

Fix $x_0 \in \mathbb{R}^d$, $y_0 \in \mathbb{R}^d$, an arbitrary continuous direction $\eta : [0,1] \to \mathbb{R}^p$, and set $\theta^\varepsilon(s) := \theta(s) + \varepsilon\eta(s)$. Let $(x_s^\varepsilon, \mu_s^\varepsilon)$ denote the corresponding solution of (1.1). We start with the differentiability of $(x_s^\varepsilon, \mu_s^\varepsilon)$ in ε at $\varepsilon = 0$.

Differentiability with respect to ε. Let Φ_s^ε be the characteristic map associated with μ_s^ε:

$$\begin{cases} \partial_s \Phi_s^\varepsilon(x) = \mathcal{V}(\Phi_s^\varepsilon(x), \mu_s^\varepsilon; \theta^\varepsilon(s)), \\ \Phi_0^\varepsilon(x) = x. \end{cases} \tag{B.1}$$

Then $\mu_s^\varepsilon = (\Phi_s^\varepsilon)_{\#}\mu_0$ and $x_s^\varepsilon = \Phi_s^\varepsilon(x_0)$. Write $\Phi_s := \Phi_s^0$ and note that $\mu_s = \mu_s^0$. For $|\varepsilon|$ sufficiently small, the paths $\{\theta_s^\varepsilon\}$ are uniformly bounded, and the support estimate places all relevant perturbed characteristics in a common compact ball on which Assumptions (A3) and (A4) hold. In what follows, every $\sup_a$ is taken over for $a \in \operatorname{supp}\mu_0 \cup \{x_0\}$. We show Φ_s^ε is differentiable in ε in the following.

Similar to the argument of Theorem A.1, applying Assumption (A3) and Grönwall's inequality also gives the preliminary stability estimate

$$\sup_{s\in[0,1]} \sup_a |\Phi_s^\varepsilon(a) - \Phi_s(a)| \le C|\varepsilon|\|\eta\|_{L^\infty}, \tag{B.2}$$

for $C = C(\mathbf{A})$. Now, let $U_s(a)$ be the unique $\mathbb{R}^d$-valued solution, continuous in a and absolutely continuous in s, of the following equation, understood for almost every s:

$$\begin{cases} \partial_s U_s(a) = D_x\mathcal{V}(\Phi_s(a), \mu_s; \theta(s))U_s(a) + \displaystyle\int_{\mathbb{R}^d} \mathbb{W}_\mu\mathcal{V}[\Phi_s(a), \mu_s; \theta(s)](\Phi_s(b))U_s(b)\,\mathrm{d}\mu_0(b) \\ \qquad\qquad + D_\theta\mathcal{V}(\Phi_s(a), \mu_s; \theta(s))\eta(s), \\ U_0(a) = 0. \end{cases} \tag{B.3}$$

The form of (B.3) comes from taking derivative of $\Phi_s^\varepsilon(a)$ in ε formally, whose existence, uniqueness, and regularity, follow directly from similar arguments in Proposition B.2 and Proposition B.3. Assumption (A4) implies that the first-order Taylor remainder of $\mathcal{V}$ is quadratic. Thus, subtracting the integral forms of the characteristic equations for both Φ^ε and Φ, dividing by ε, applying the preceding Taylor expansion, then subtracting (B.3), and applying (B.2) for the remainder term, gives, for every $s \in [0,1]$,

$$\sup_a \left| \frac{\Phi_s^\varepsilon(a) - \Phi_s(a)}{\varepsilon} - U_s(a) \right| \le C\int_0^s \sup_a \left| \frac{\Phi_r^\varepsilon(a) - \Phi_r(a)}{\varepsilon} - U_r(a) \right| \mathrm{d}r + C|\varepsilon|\|\eta\|_{L^\infty}^2.$$

Therefore, another application of Grönwall's inequality yields $\varepsilon \mapsto \Phi^\varepsilon$ is differentiable in the uniform path norm, with $\partial_\varepsilon \Phi_s^\varepsilon|_{\varepsilon=0} = U_s$.

Consequently, for every $f \in C^1(\mathbb{R}^d)$, the pushforward identity and the mean-value formula give, uniformly in s,

$$\langle \sigma_s, f\rangle \coloneqq \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \int_{\mathbb{R}^d} f(z)\,\mathrm{d}\mu_s^\varepsilon(z) = \int_{\mathbb{R}^d} \nabla f(\Phi_s(a))\cdot U_s(a)\,\mathrm{d}\mu_0(a).$$

This formula defines σ_s as a compactly supported distribution of order at most one, with $\operatorname{supp}\sigma_s \subseteq \Phi_s(\operatorname{supp}\mu_0) \subseteq B_C(0)$ for $C = C(\mathbf{A})$ using Theorem A.1, and $\sigma_0 = 0$ because $\mu_0^\varepsilon = \mu_0$. Moreover, for this $C = C(\mathbf{A})$, let $f_s^\varepsilon \in C_c^1(\mathbb{R}^d)$ be a family of test function with $\operatorname{supp} f_s^\varepsilon \subseteq B_{2C}(0)$ and suppose further that

$$\frac{f_s^\varepsilon(z) - f_s^0(z)}{\varepsilon} \longrightarrow \partial_\varepsilon f_s^\varepsilon(z)|_{\varepsilon=0}$$

uniformly in (s,z) on this ball, with a jointly continuous limit. Then,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \int_{\mathbb{R}^d} f_s^\varepsilon(z)\,\mathrm{d}\mu_s^\varepsilon(z) = \big\langle \sigma_s, f_s^0\big\rangle + \int_{\mathbb{R}^d} \partial_\varepsilon f_s^\varepsilon(z)|_{\varepsilon=0}\,\mathrm{d}\mu_s(z),$$

which will be frequently used below to differentiate the Lagrangian.

Weak measure-valued Lagrangian. For a smooth multiplier ϕ_s and an absolutely continuous multiplier p_s, define the token and measure constraint terms by

$$\begin{aligned}
\mathsf{T}(\theta^\varepsilon) &\coloneqq \int_0^1 p_s \cdot \big(\dot{x}_s^\varepsilon - \mathcal{V}(x_s^\varepsilon, \mu_s^\varepsilon; \theta^\varepsilon(s))\big)\,\mathrm{d}s,\\
\mathsf{M}(\theta^\varepsilon) &\coloneqq \int_{\mathbb{R}^d} \phi_1(z)\,\mathrm{d}\mu_1^\varepsilon(z) - \int_{\mathbb{R}^d} \phi_0(z)\,\mathrm{d}\mu_0(z)\\
&\quad - \int_0^1 \int_{\mathbb{R}^d} \Big[\partial_s\phi_s(z) + \nabla\phi_s(z)\cdot \mathcal{V}(z, \mu_s^\varepsilon; \theta^\varepsilon(s))\Big]\,\mathrm{d}\mu_s^\varepsilon(z)\,\mathrm{d}s.
\end{aligned}$$

The identity $\mathsf{M}(\theta^\varepsilon) = 0$ is exactly the weak formulation of

$$\partial_s \mu_s^\varepsilon + \nabla\cdot\big(\mu_s^\varepsilon \mathcal{V}(\,\cdot\,, \mu_s^\varepsilon; \theta^\varepsilon(s))\big) = 0.$$

It is meaningful for an arbitrary probability measure. Compact support removes boundary terms at infinity; equivalently, one may multiply ϕ by a smooth cutoff that equals one on the common support.

Define the augmented Lagrangian

$$L(\theta^\varepsilon) \coloneqq \ell(x_1^\varepsilon, y_0) - \mathsf{T}(\theta^\varepsilon) - \mathsf{M}(\theta^\varepsilon).$$

Since $(x_s^\varepsilon, \mu_s^\varepsilon)$ satisfies (1.1), both constraints $\mathsf{T}(\theta^\varepsilon)$ and $\mathsf{M}(\theta^\varepsilon)$ vanish for every choice of (p_s, ϕ_s), and so $L(\theta^\varepsilon) = \mathcal{L}(\theta^\varepsilon; \mu_0, x_0, y_0)$ for all ε.

Token term. Integration by parts in s, Using $x_0^\varepsilon = x_0$ and the measure chain rule give

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \mathsf{T}(\theta^\varepsilon) &= p_1\cdot\partial_\varepsilon x_1^0 - \int_0^1 \big(\dot{p}_s + D_x\mathcal{V}(x_s, \mu_s; \theta(s))^\top p_s\big)\cdot \partial_\varepsilon x_s^0\,\mathrm{d}s\\
&\quad - \int_0^1 \Big\langle \sigma_s,\, p_s \cdot \frac{\delta \mathcal{V}}{\delta \mu}[x_s, \mu_s; \theta(s)](\,\cdot\,)\Big\rangle\,\mathrm{d}s\\
&\quad - \int_0^1 p_s\cdot\big(D_\theta \mathcal{V}(x_s, \mu_s; \theta(s))\eta(s)\big)\,\mathrm{d}s.
\end{aligned} \tag{B.4}$$

Measure term. Differentiating the weak measure constraint and using $\sigma_0 = 0$ gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \mathsf{M}(\theta^\varepsilon) = \langle \sigma_1, \phi_1 \rangle \\ - \int_0^1 \Big\langle \sigma_s,\, \partial_s \phi_s(\,\cdot\,) + \nabla \phi_s(\,\cdot\,) \cdot \mathcal{V}(\,\cdot\,, \mu_s; \theta(s)) \\ + \int_{\mathbb{R}^d} \nabla \phi_s(\xi) \cdot \frac{\delta \mathcal{V}}{\delta \mu}[\xi, \mu_s; \theta(s)](\,\cdot\,)\, \mathrm{d}\mu_s(\xi) \Big\rangle \, \mathrm{d}s \\ - \int_0^1 \int_{\mathbb{R}^d} \nabla \phi_s(z) \cdot \big(D_\theta \mathcal{V}(z, \mu_s; \theta(s)) \eta(s) \big)\, \mathrm{d}\mu_s(z)\, \mathrm{d}s.\end{aligned} \tag{B.5}$$

For smooth ϕ, every pairing in (B.5) is well-defined through the explicit action of σ_s on C^1 test functions.

Combination. Adding $\nabla_x \ell(x_1^0, y_0) \cdot \partial_\varepsilon x_1^0$ to (B.4) and (B.5), and choosing (p_s, ϕ_s) so that the coefficients of $\partial_\varepsilon x_1^0$, $\partial_\varepsilon x_s^0$, σ_1, σ_s all vanish, leads precisely to the adjoint system (3.2)–(3.3) of the body. The remaining terms involve only η and yield, after collecting:

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \mathcal{L}(\theta^\varepsilon; \mu_0, x_0, y_0) \;=\; \int_0^1 \Big[(D_\theta \mathcal{V})^\top p_s + \int_{\mathbb{R}^d} (D_\theta \mathcal{V}(z, \mu_s; \theta(s)))^\top \nabla \phi_s(z)\, \mathrm{d}\mu_s(z) \Big] \cdot \eta(s)\, \mathrm{d}s.$$

Since η is arbitrary, this is the formula (3.1). □

Riemannian extension. The argument extends without modification when (x_s, μ_s) is constrained to a Riemannian submanifold $\mathbf{M} \subseteq \mathbb{R}^d$. The Lagrangian is unchanged (with the inner product on $\mathbf{M}$ in place of the Euclidean one), the multiplier p_s becomes a tangent vector in $T_{x_s}\mathbf{M}$, and the spatial Jacobian $D_x \mathcal{V}$ is replaced by the covariant derivative $\nabla^{\mathbf{M}} \mathcal{V}$. The resulting backward equation for p_s reads

$$\dot{p}_s \;=\; -\big[\nabla^{\mathbf{M}} \mathcal{V}(x_s, \mu_s; \theta(s))\big]^* p_s,$$

with $[\nabla^{\mathbf{M}} \mathcal{V}]^*$ the metric dual of the $(1,1)$-tensor $\nabla^{\mathbf{M}} \mathcal{V}$ at x_s; the equation for ϕ_s is unchanged in form once the inner products are reinterpreted by the Riemannian metric on $\mathbf{M}$.

B.2 Token adjoint: well-posedness and stability

We now establish the bounds on p_s and its dependence on the data. Existence and uniqueness of the solution to (3.2) follow from classical linear ODE theory.

Proposition B.1 (Bounds and stability for p_s)**.** *Adopt the assumptions of Theorem 3.1. There exist constants C_1, C_2 depending on $\mathbf{A}$ such that, with p_s and $\widetilde{p}_s$ the solutions of* (3.2) *associated with (x_s, μ_s, θ) and $(\widetilde{x}_s, \widetilde{\mu}_s, \widetilde{\theta})$ respectively,*

$$\sup_{s \in [0,1]} |p_s| \;\le\; C_1, \qquad \sup_{s \in [0,1]} |\widetilde{p}_s| \;\le\; C_1, \tag{B.6}$$

$$\sup_{s \in [0,1]} |p_s - \widetilde{p}_s| \;\le\; C_2\, \mathcal{K}, \qquad \mathcal{K} := \mathrm{W}_1(\mu_0, \widetilde{\mu}_0) + |x_0 - \widetilde{x}_0| + |y_0 - \widetilde{y}_0| + \|\theta - \widetilde{\theta}\|_{L^1}. \tag{B.7}$$

Proof. By (3.2) and (A4), $\frac{\mathrm{d}}{\mathrm{d}s}|p_s| \le L_0 |p_s|$. The terminal condition $|p_1| \le M_\ell (2C + 1)$ (from (A2) and the bounds in Theorem A.1) and Grönwall's inequality yield (B.6). For (B.7), note first that

$$|p_1 - \widetilde{p}_1| \;\le\; L_\ell \big(|x_1 - \widetilde{x}_1| + |y_0 - \widetilde{y}_0| \big).$$

Subtracting the two adjoint equations and using (A3), (A4), and (B.6),

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}s}|p_s - \widetilde{p}_s| &\le \big| D_x \mathcal{V}(x_s, \mu_s; \theta(s))^\top p_s - D_x \mathcal{V}(\widetilde{x}_s, \widetilde{\mu}_s; \widetilde{\theta}(s))^\top \widetilde{p}_s \big| \\ &\le L_1 \big(|x_s - \widetilde{x}_s| + \mathrm{W}_1(\mu_s, \widetilde{\mu}_s) + |\theta(s) - \widetilde{\theta}(s)| \big)\, C_1 + L_0 |p_s - \widetilde{p}_s|.\end{aligned}$$

Substituting the bounds of Theorem A.1 and applying Grönwall yields (B.7). □

B.3 Measure adjoint: well-posedness, regularity, and stability

The well-posedness of (3.3) is more delicate than (3.2) because the right-hand side depends on $\nabla\phi_s$ in a nonlocal way. We resolve this by a contraction argument in a weighted C^1 space.

Proposition B.2 (Existence and regularity of ϕ_s)**.** *Adopt the assumptions of Theorem 3.1. There exists a unique solution ϕ_s of* (3.3) *in the class $\{\phi \in C([0,1]; C^1(\mathbb{R}^d)) : \phi(s,\cdot) \in C^{1,1}(\mathbb{R}^d)\}$. There exists $C = C(\mathbf{A})$ such that, for all $s \in [0,1]$ and $z, \widetilde{z} \in \mathbb{R}^d$,*

$$|\phi_s(z)| \le C(|z|+1), \qquad |\nabla\phi_s(z)| \le C, \qquad |\nabla\phi_s(z) - \nabla\phi_s(\widetilde{z})| \le C|z-\widetilde{z}|.$$

Proof. Define the Banach space

$$X \coloneqq \big\{\psi : [0,1]\times\mathbb{R}^d \to \mathbb{R} \,:\, \psi(s,\cdot) \in C^1(\mathbb{R}^d),\ \|\psi\|_X < \infty\big\}, \tag{B.8}$$

with norm

$$\|\psi\|_X \coloneqq \sup_{s\in[0,1],\ z\in\mathbb{R}^d} e^{\alpha(s-1)} \left(\frac{|\psi(s,z)|}{|z|+1} + |\nabla_z\psi(s,z)|\right), \tag{B.9}$$

where $\alpha = \alpha(\mathbf{A}) > 0$ will be chosen below. For $\psi \in X$, set

$$f(s,z;\psi) \coloneqq -p_s \cdot \tfrac{\delta\mathcal{V}}{\delta\mu}[x_s,\mu_s;\theta(s)](z) - \int_{\mathbb{R}^d} \nabla\psi_s(\xi)\cdot\tfrac{\delta\mathcal{V}}{\delta\mu}[\xi,\mu_s;\theta(s)](z)\,\mathrm{d}\mu_s(\xi). \tag{B.10}$$

By (A3) and Proposition B.1, there exists $C = C(\mathbf{A})$ such that

$$e^{\alpha(s-1)}|f(s,z;\psi)| \le C(|z|+1)\big(\|\psi\|_X + e^{\alpha(s-1)}\big),\ e^{\alpha(s-1)}|\nabla_z f(s,z;\psi)| \le C\big(\|\psi\|_X + e^{\alpha(s-1)}\big). \tag{B.11}$$

Let $\Psi_\tau(z)$ be the solution from depth s of $\partial_\tau\Psi_\tau = \mathcal{V}(\Psi_\tau,\mu_\tau;\theta(\tau))$ with $\Psi_s(z) = z$, $\forall z \in \mathbb{R}^d$. Then, the transport equation

$$\partial_s u(s,z) + \nabla_z u(s,z)\cdot\mathcal{V}(z,\mu_s;\theta(s)) = f(s,z;\psi), \qquad u(1,z) = 0 \tag{B.12}$$

admits the explicit solution

$$u(s,z) \;=\; -\int_s^1 f(\tau,\Psi_\tau(z);\psi)\,\mathrm{d}\tau. \tag{B.13}$$

We show that this $u(s,z) \in X$. First, by Assumption (A3),

$$\partial_\tau|\Psi_\tau(z)| \le |\partial_\tau\Psi_\tau(z)| \le M_0(|\Psi_\tau(z)| + M).$$

A Grönwall argument yields that $|\Psi_\tau(z)| \le Me^{M_0}(|z|+1)$ for any $\tau \in [s,1]$ and any $z \in \mathbb{R}^d$. Similarly, because

$$\partial_\tau D_z\Psi_\tau(z) = D_x\mathcal{V}(\Psi_\tau(z),\mu_\tau;\theta(\tau))D_z\Psi_\tau(z), \quad D_z\Psi_s(z) = \mathrm{Id}, \tag{B.14}$$

We see that by (A3),

$$\partial_\tau\|D_z\Psi_\tau(z)\| \le L_0\|D_z\Psi_\tau(z)\|.$$

By Grönwall, $|\Psi_\tau(z)| \le Me^{M_0}(|z|+1)$ and $\|D_z\Psi_\tau(z)\| \le e^{L_0}$ uniformly on $[s,1]\times\mathbb{R}^d$. Setting $s = \tau$ in (B.11) and combining with (B.13),

$$\begin{aligned}\|u\|_X &\le C\left(\|\psi\|_X \sup_{s\in[0,1]} e^{\alpha(s-1)}\int_s^1 e^{\alpha(1-\tau)}\,\mathrm{d}\tau + 1\right)\\ &= C\left(\frac{1}{\alpha}\|\psi\|_X \sup_{s\in[0,1]}(1-e^{\alpha(s-1)}) + 1\right) \le C\left(\frac{1}{\alpha}\|\psi\|_X + 1\right),\end{aligned}$$

Choosing $\alpha = 4C(C+1)$, the map $\psi \mapsto u$ leaves the ball $\{\psi \in X : \|\psi\|_X \le C+1\}$ invariant.

For any two $\psi_1, \psi_2 \in X$, we can similarly obtain u_1, u_2. By the linearity of ψ in the function $f(s,z;\psi)$, we see that $f(s,z;\psi_1-\psi_2)$ is homogeneous with respect to $\psi_1-\psi_2$. We can thus similarly show that

$$\|u_1 - u_2\|_X \le \frac{C}{\alpha}\|\psi_1-\psi_2\|_X = \frac{1}{4}\|\psi_1-\psi_2\|_X.$$

Thus, the map from ψ to u is a contraction. By the Banach fixed-point theorem, there is a unique fixed point $\phi(s,z) \in X$ which satisfies

$$\|\phi\|_X \le C + 1 \tag{B.15}$$

and solves (B.12) for $u = \psi = \phi$. This $\phi(s,z)$ is exactly the unique solution to (3.3).

For the Lipschitz bound on $\nabla\phi_s$, differentiate (B.13) in z:

$$\nabla_z \phi(s,z) \;=\; -\int_s^1 (D_z\Psi_\tau(z))^\top \big(\nabla_z f(\tau, \Psi_\tau(z); \phi)\big)\,\mathrm{d}\tau.$$

We first notice that $D_z\Psi_\tau(z)$ is uniformly Lipschitz continuous, which follows from a direct Grönwall argument for $\|D_z\Psi_\tau(z_1) - D_z\Psi_\tau(z_2)\|, \forall z_1, z_2 \in \mathbb{R}^d$. Indeed, by (B.14),

$$\begin{aligned}\partial_\tau \|D_z\Psi_\tau(z_1) - D_z\Psi_\tau(z_2)\| &\le \|D_x\mathcal{V}(\Psi_\tau(z_1), \mu_\tau; \theta(\tau))\| \cdot \|D_z\Psi_\tau(z_1) - D_z\Psi_\tau(z_2)\| \\ &\quad + \|D_x\mathcal{V}(\Psi_\tau(z_1), \mu_\tau; \theta(\tau)) - D_x\mathcal{V}(\Psi_\tau(z_2), \mu_\tau; \theta(\tau))\| \cdot \|D_z\Psi_\tau(z_2)\| \\ &\le L_0\|D_z\Psi_\tau(z_1) - D_z\Psi_\tau(z_2)\| + L_1 e^{2L_0}|z_1 - z_2|,\end{aligned}$$

where the last inequality follows from the assumptions (A3) and (A4) on $D_x\mathcal{V}$, together with the previous proved fact that $\|D_z\Psi_\tau(z)\| \le e^{L_0}$. A Grönwall type argument then shows that $D_z\Psi_\tau(z)$ is Lipschitz with constant depending only on $\mathbf{A}$.

Next, using the explicit form (B.10),

$$(\nabla_z f(s,z;\phi))^\top = -p_s \cdot (\mathbb{W}_\mu\mathcal{V})[x_s, \mu_s; \theta(s)](z) - \int_{\mathbb{R}^d} \nabla\phi_s(\xi) \cdot (\mathbb{W}_\mu\mathcal{V})[\xi, \mu_s; \theta(s)](z)\,\mathrm{d}\mu_s(\xi).$$

By the Lipschitz assumption (A4) on $\mathbb{W}_\mu\mathcal{V}$ with respect to the z-variable, and the fact that $|\nabla\phi_s(\xi)| \le \|\phi\|_X \le C + 1$ in (B.15), we conclude that $\nabla_z f(s,z;\phi)$ is also Lipschitz continuous in z with constant depending only on $\mathbf{A}$. Thus, together with the uniform Lipschitz continuity of $D_z\Psi_\tau(z)$ showed above, $\nabla_z\phi(s,\cdot)$ is uniformly Lipschitz in z, with constant depending only on $\mathbf{A}$.

□

We turn to stability. Let $\theta, \widetilde{\theta}$, $\mu_0, \widetilde{\mu}_0$, $x_0, \widetilde{x}_0$, $y_0, \widetilde{y}_0$ all satisfy (A1), with corresponding solutions (μ_s, p_s, ϕ_s) and $(\widetilde{\mu}_s, \widetilde{p}_s, \widetilde{\phi}_s)$.

Proposition B.3 (Stability of ϕ_s)**.** *Adopt the assumptions of Theorem 3.1. There exist C_1, C_2 depending on $\mathbf{A}$ such that, for all $s \in [0,1]$ and $z \in \mathbb{R}^d$,*

$$|\phi_s(z) - \widetilde{\phi}_s(z)| \le C_1\,\mathcal{K}\,(|z| + 1), \qquad |\nabla\phi_s(z) - \nabla\widetilde{\phi}_s(z)| \le C_2\,\mathcal{K},$$

where $\mathcal{K} = \mathrm{W}_1(\mu_0, \widetilde{\mu}_0) + |x_0 - \widetilde{x}_0| + |y_0 - \widetilde{y}_0| + \|\theta - \widetilde{\theta}\|_{L^1}$.

Proof. Set $\Gamma_s(z) \coloneqq \widetilde{\phi}_s(z) - \phi_s(z)$. Subtracting the two equations (3.3) and rearranging,

$$\partial_s\Gamma_s(z) + \nabla_z(\Gamma_s(z)) \cdot \mathcal{V}(z, \widetilde{\mu}_s; \widetilde{\theta}(s)) + \int_{\mathbb{R}^d} \nabla(\Gamma_s(\xi)) \cdot \tfrac{\delta\mathcal{V}}{\delta\mu}[\xi, \widetilde{\mu}_s; \widetilde{\theta}(s)](z)\mathrm{d}\widetilde{\mu}_s(\xi) = g(s,z),$$

with $\Gamma_1 \equiv 0$ and source term

$$\begin{aligned}g(s,z) = &-\nabla\phi_s(z) \cdot \big(\mathcal{V}(z, \widetilde{\mu}_s; \widetilde{\theta}(s)) - \mathcal{V}(z, \mu_s; \theta(s))\big) \\ &- \big(\widetilde{p}_s \cdot \tfrac{\delta\mathcal{V}}{\delta\mu}[\widetilde{x}_s, \widetilde{\mu}_s; \widetilde{\theta}(s)](z) - p_s \cdot \tfrac{\delta\mathcal{V}}{\delta\mu}[x_s, \mu_s; \theta(s)](z)\big) \\ &- \int_{\mathbb{R}^d} \nabla\phi_s(\xi) \cdot \big(\tfrac{\delta\mathcal{V}}{\delta\mu}[\xi, \widetilde{\mu}_s; \widetilde{\theta}(s)](z)\,\mathrm{d}\widetilde{\mu}_s(\xi) - \tfrac{\delta\mathcal{V}}{\delta\mu}[\xi, \mu_s; \theta(s)](z)\,\mathrm{d}\mu_s(\xi)\big).\end{aligned}$$

Combining the Lipschitz bounds in (A3), (A4), Theorem A.1, Proposition B.1, and Proposition B.2, we obtain

$$|g(s,z)| \le C(|z| + 1)\big(\mathcal{K} + |\theta(s) - \widetilde{\theta}(s)|\big), \qquad |\nabla_z g(s,z)| \le C\big(\mathcal{K} + |\theta(s) - \widetilde{\theta}(s)|\big). \tag{B.16}$$

The argument of Proposition B.2, applied to Γ_s in the same Banach space X but with the inhomogeneous source g, gives

$$\|\Gamma\|_X \le C\Big(\tfrac{1}{\alpha}\|\Gamma\|_X + \mathcal{K}\Big),$$

and choosing $\alpha = 4C$ yields $\|\Gamma\|_X \le 4C\,\mathcal{K}$, which is the stated bound. □

B.4 Stability of the loss gradient

We are now in position to prove Theorem 4.2, restated below in its precise form.

Theorem B.4 (Stability of the loss gradient)**.** *Adopt the assumptions of Theorem 3.1. There exists $C = C(\mathbf{A})$ such that, for all $s \in [0,1]$,*

$$\begin{aligned}&\left|\tfrac{\delta\mathcal{L}(\theta;\mu_0,x_0,y_0)}{\delta\theta}(s) - \tfrac{\delta\mathcal{L}(\widetilde{\theta};\widetilde{\mu}_0,\widetilde{x}_0,\widetilde{y}_0)}{\delta\theta}(s)\right| \\ &\quad\le C\big(\mathrm{W}_1(\mu_0,\widetilde{\mu}_0) + |x_0-\widetilde{x}_0| + |y_0-\widetilde{y}_0| + \|\theta-\widetilde{\theta}\|_{L^1} + |\theta(s)-\widetilde{\theta}(s)|\big).\end{aligned}$$

Proof. Apply the triangle inequality to the gradient formula (3.1) and bound each piece using:

- the Lipschitz bound on $D_\theta\mathcal{V}$ in (x,μ,θ) from (A4),
- the bounds on $|p_s|$ and $|p_s - \widetilde{p}_s|$ from Proposition B.1,
- the bounds on $|\nabla\phi_s|$ and $|\nabla\phi_s - \nabla\widetilde{\phi}_s|$ from Propositions B.2 and B.3,
- the bounds on $|x_s - \widetilde{x}_s|$ and $\mathrm{W}_1(\mu_s,\widetilde{\mu}_s)$ from Theorem A.1.

Each contribution is linear in $\mathcal{K} + |\theta(s) - \widetilde{\theta}(s)|$ with $\mathcal{K}$ defined in Proposition B.3, and the conclusion follows. □

C Non-asymptotic OGD analysis: proof of Theorem 2.2

This appendix proves Theorem 2.2 under Assumption A via three intermediate results: a uniform-in-iteration boundedness lemma for the OGD trajectory, a deterministic stability bound with respect to data and initialization, and a McDiarmid-type concentration bound for the empirical OGD trajectory.

C.1 Uniform boundedness of the OGD iterates

The first lemma ensures that all results of Appendices A–B apply along the OGD trajectory (2.3), with (A1) replaced by the slightly enlarged bound $\|\theta\|_{L^\infty} \le M + 3$.

Lemma C.1 (Boundedness of OGD)**.** *Adopt the assumptions of Theorem 3.1. There exists $C = C(\mathbf{A})$ such that, for any sequence of training inputs $\{(\mu_0^k, x_0^k, y_0^k)\}_{k\in\mathbb{N}}$ satisfying* (A1) *and any initialization θ_0 with $\|\theta_0\|_{L^\infty} \le M$, the OGD iterates* (2.3) *satisfy:*

(i) *For every $\eta > 0$, $\lambda \in [0, \eta^{-1})$, and $k \le (\log(1+C\eta))^{-1}$,*

$$\|\theta_k\|_{L^\infty} \le \|\theta_0\|_{L^\infty} + 3. \tag{C.1}$$

(ii) *If $\eta^{-1} > C$ and $\lambda \in (C, \eta^{-1})$, then for every $k \in \mathbb{N}$,*

$$\|\theta_k\|_{L^\infty} \le \|\theta_0\|_{L^\infty} + 1. \tag{C.2}$$

Proof. By Theorem 3.1 and the boundedness of $D_\theta\mathcal{V}$ in (A4), $\|\delta\mathcal{L}/\delta\theta\|_{L^\infty} \le C$ for some $C = C(\mathbf{A})$. Iterating (2.3) gives

$$\|\theta_k\|_{L^\infty} \le (1-\eta\lambda)^k\|\theta_0\|_{L^\infty} + \eta C\sum_{i=0}^{k-1}(1-\eta\lambda)^i.$$

Each term in the geometric sum can be bounded by $(1+C\eta)^i$, so the second term is bounded by $(1+C\eta)^k = e < 3$ when $k \le (\log(1+C\eta))^{-1}$, giving (i); when $\lambda > C$, it is bounded by $C/\lambda < 1$ for every k, giving (ii). □

C.2 Stability of the OGD iterates

Let $\{(\mu_0^k, x_0^k, y_0^k)\}_{k\in\mathbb{N}}$ and $\{(\widetilde{\mu}_0^k, \widetilde{x}_0^k, \widetilde{y}_0^k)\}_{k\in\mathbb{N}}$ be two streams of training data, and $\theta_0, \widetilde{\theta}_0$ two initializations, all satisfying (A1). Let $\{\theta_k\}$ and $\{\widetilde{\theta}_k\}$ be the corresponding OGD trajectories, and define the data discrepancy

$$\mathcal{E}_k \;\coloneqq\; \max_{\ell\le k-1}\big(\mathrm{W}_1(\mu_0^\ell, \widetilde{\mu}_0^\ell) + |x_0^\ell - \widetilde{x}_0^\ell| + |y_0^\ell - \widetilde{y}_0^\ell|\big). \tag{C.3}$$

Theorem C.2 (Finite- and long-horizon stability of OGD)**.** *Adopt the assumptions of Theorem 3.1. There exist $C = C(\mathbf{A}), \eta_c = \eta_c(\mathbf{A}), \lambda_c = \lambda_c(\mathbf{A}) < \eta_c^{-1}$ such that:*

(i) *For every $\eta > 0$, $\lambda \in [0, \eta^{-1})$, and $k \le (\log(1 + C\eta))^{-1}$,*

$$\|\theta_k - \widetilde{\theta}_k\|_{L^\infty} \;\le\; 3\big(\|\theta_0 - \widetilde{\theta}_0\|_{L^\infty} + \mathcal{E}_k\big). \tag{C.4}$$

(ii) *For every $\eta < \eta_c$, $\lambda \in (\lambda_c, \eta^{-1})$, and every $k \in \mathbb{N}$,*

$$\|\theta_k - \widetilde{\theta}_k\|_{L^\infty} \;\le\; \|\theta_0 - \widetilde{\theta}_0\|_{L^\infty} + \mathcal{E}_k. \tag{C.5}$$

Proof. Set $\Delta_l \coloneqq \|\theta_l - \widetilde{\theta}_l\|_{L^\infty}$. Subtracting the two OGD updates and applying Theorem B.4,

$$\Delta_{l+1} \le (1-\eta\lambda)\Delta_l + \eta C\big(\Delta_l + \mathrm{W}_1(\mu_0^l, \widetilde{\mu}_0^l) + |x_0^l - \widetilde{x}_0^l| + |y_0^l - \widetilde{y}_0^l|\big) \le (1 - \eta(\lambda - C))\Delta_l + \eta C\,\mathcal{E}_k. \tag{C.6}$$

Statement (i). For $\lambda \ge 0$, $1 - \eta(\lambda - C) \le 1 + \eta C$, and iterating yields

$$\Delta_k \le (1+\eta C)^k \Delta_0 + \eta C\,\mathcal{E}_k \sum_{i=0}^{k-1}(1+\eta C)^i \le (1+\eta C)^k(\Delta_0 + \mathcal{E}_k).$$

For $k \le (\log(1+\eta C))^{-1}$, $(1+\eta C)^k \le e < 3$, giving (C.4).

Statement (ii). Choose $\lambda_c = 2C$ and $\eta_c = (2C)^{-1}$. Then $1 - \eta(\lambda - C) \in (0, 1)$, and iterating (C.6) gives, for every $k \in \mathbb{N}$,

$$\Delta_k \le (1-\eta C)^k \Delta_0 + \eta C\,\mathcal{E}_k \sum_{i=0}^{k-1}(1-\eta C)^i \le \Delta_0 + \mathcal{E}_k.$$

□

C.3 Empirical OGD: mean and concentration bounds

We now specialize Theorem C.2 to the case in which $\widetilde{\theta}_k = \widehat{\theta}_k$ is driven by the empirical contextual measure

$$\widehat{\mu}_0^k \;\coloneqq\; \frac{1}{n}\sum_{i=1}^n \delta_{z^{(k,i)}}, \qquad z^{(k,i)} \overset{\text{iid}}{\sim} \mu_0^k, \tag{C.7}$$

with the same initialization θ_0, the same (x_0^k, y_0^k), and the same update rule.

Theorem C.3 (Mean bound for the empirical OGD)**.** *Adopt the assumptions of Theorem 3.1, and let $d \ge 3$. There exist $C = C(\mathbf{A}), \eta_c = \eta_c(\mathbf{A}), \lambda_c = \lambda_c(\mathbf{A}) < \eta_c(\mathbf{A})^{-1}$ such that:*

(i) *For $\eta > 0$, $\lambda \in [0, \eta^{-1})$, we let $k_c \coloneqq (\log(1 + C\eta))^{-1}$. Then,*

$$\mathbb{E}\sup_{k\le k_c}\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \le 3n^{-1/d}. \tag{C.8}$$

(ii) *For $\eta < \eta_c$ and $\lambda \in (\lambda_c, \eta^{-1})$, every $k \in \mathbb{N}$ satisfies*

$$\mathbb{E}\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \le n^{-1/d}. \tag{C.9}$$

If μ_0^k is supported on a d'-dimensional submanifold of $\mathbb{R}^d$, the rate $n^{-1/d}$ is replaced by $n^{-1/d'}$.

Proof. Set $\Delta_l = \|\theta_l - \widehat{\theta}_l\|_{L^\infty}$ and apply Theorem B.4 as in (C.6). Since the only data discrepancy is the contextual measure,

$$\Delta_{l+1} \le (1 - \eta(\lambda - C))\Delta_l + \eta C \, \mathrm{W}_1(\mu_0^l, \widehat{\mu}_0^l). \tag{C.10}$$

Notice that $\Delta_0 = 0$.

Statement (i). For $\lambda \ge 0$, $1 - \eta(\lambda - C) \le 1 + \eta C$, and iterating yields

$$\sup_{k \le k_c} \Delta_k \le \eta C \sum_{l=0}^{k_c - 1} (1 + \eta C)^{k_c - 1 - l} \mathrm{W}_1(\mu_0^l, \widehat{\mu}_0^l).$$

Take expectations and apply Lemma A.3 to each $\mathbb{E}\mathrm{W}_1(\mu_0^l, \widehat{\mu}_0^l)$, we find that

$$\mathbb{E} \sup_{k \le k_c} \Delta_k \le \eta C n^{-1/d} \sum_{l=0}^{k_c - 1} (1 + \eta C)^{k_c - 1 - l} \le (1 + \eta C)^{k_c} n^{-1/d}$$

Because $k_c = (\log(1 + \eta C))^{-1}$, $(1 + \eta C)^k \le e < 3$, we then obtain (C.8).

Statement (ii). Choose $\lambda_c = 2C$ and $\eta_c = (2C)^{-1}$. Then $1 - \eta(\lambda - C) \in (0, 1)$, and iterating (C.6) gives, for every $k \in \mathbb{N}$,

$$\Delta_k \le \eta C \sum_{l=0}^{k-1} (1 - \eta C)^{k-1-l} \mathrm{W}_1(\mu_0^l, \widehat{\mu}_0^l).$$

Take expectations and apply Lemma A.3, we similarly find that

$$\mathbb{E}\Delta_k \le \eta C n^{-1/d} \sum_{l=0}^{k-1} (1 - \eta C)^{k-1-l} \le (1 - (1 - \eta C)^k) n^{-1/d} \le n^{-1/d}.$$

□

Theorem C.4 (Concentration of the empirical OGD)**.** *Under the assumptions and notations of Theorem C.3, there exists a constant $\alpha = \alpha(\mathbf{A})$ such that:*

(i) *For $\eta > 0$, $\lambda \in [0, \eta^{-1})$, we let $k_c \coloneqq (\log(1 + C\eta))^{-1}$. Then, for every $\varepsilon > 0$,*

$$\mathbb{P}\big(\big| \sup_{k \le k_c} \|\theta_k - \widehat{\theta}_k\|_{L^\infty} - \mathbb{E} \sup_{k \le k_c} \|\theta_k - \widehat{\theta}_k\|_{L^\infty} \big| > \varepsilon \big) \le 2 \exp(-\eta^{-1} \alpha n \varepsilon^2). \tag{C.11}$$

(ii) *Let $\eta < \eta_c$, $\lambda \in (\lambda_c, \eta^{-1})$. Then, for every $k \in \mathbb{N}$ and $\varepsilon > 0$,*

$$\mathbb{P}\big(\big| \, \|\theta_k - \widehat{\theta}_k\|_{L^\infty} - \mathbb{E}\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \big| > \varepsilon \big) \le 2 \exp(-\eta^{-1} \alpha n \varepsilon^2). \tag{C.12}$$

Proof. Fix $k \in \mathbb{N}$ and view $f \coloneqq \|\theta_k - \widehat{\theta}_k\|_{L^\infty}$ as a function of the $N = kn$ i.i.d. samples $\{\{z^{(\ell,i)}\}_{i=1}^n\}_{\ell=0}^{k-1}$. The randomness enters only through $\widehat{\theta}_k$.

Bounded differences. Replacing one sample $z^{(\ell,i)}$ by an independent copy $\overline{z}^{(\ell,i)}$ produces a new initial measure $\overline{\mu}_0^\ell$ and new empirical OGD trajectories $\{\overline{\theta}_l\}_{l > \ell}$. We bound

$$\Big| \sup_{k \le k_c} \|\theta_k - \widehat{\theta}_k\|_{L^\infty} - \sup_{k \le k_c} \|\theta_k - \overline{\theta}_k\|_{L^\infty} \Big| \le \sup_{k \le k_c} \|\widehat{\theta}_k - \overline{\theta}_k\|_{L^\infty}.$$

For $l = \ell$, since the perturbed trajectories agree before step ℓ, applying Theorem B.4 to the single OGD step gives

$$\|\widehat{\theta}_{\ell+1} - \overline{\theta}_{\ell+1}\|_{L^\infty} \le \eta C \, \mathrm{W}_1(\widehat{\mu}_0^\ell, \overline{\mu}_0^\ell) \le \eta C \cdot \tfrac{2R}{n}.$$

For $l > \ell$, both trajectories see the same data and Theorem B.4 yields the contraction

$$\|\widehat{\theta}_{l+1} - \overline{\theta}_{l+1}\|_{L^\infty} \le (1 - \eta(\lambda - C)) \|\widehat{\theta}_l - \overline{\theta}_l\|_{L^\infty}.$$

Statement (i). Iterating gives the McDiarmid coefficient

$$c_i^\ell \coloneqq \sup_{k \le k_c} \|\widehat{\theta}_k - \overline{\theta}_k\|_{L^\infty} \le \; (1 + \eta C)^{k_c - 1 - \ell} \, \tfrac{2RC\eta}{n}.$$

Hence,

$$\sum_{\ell=0}^{k_c-1}\sum_{i=1}^{n}\left(c_i^\ell\right)^2 \le \frac{4R^2C^2}{n}\eta^2\sum_{\ell=0}^{k_c-1}(1+\eta C)^{2\ell}. \tag{C.13}$$

For $\lambda \in [0, \eta^{-1})$ and $k_c = (\log(1+\eta C))^{-1}$, the right-hand side of (C.13) is bounded by

$$\frac{4R^2}{n}C\eta(1+\eta C)^{2k_c} \le \frac{36R^2}{n}C\eta. \tag{C.14}$$

McDiarmid's inequality (Lemma A.4) yields (C.11).

Statement (ii). Iterating gives the McDiarmid coefficient

$$c_i^\ell := \|\widehat{\theta}_k - \overline{\theta}_k\|_{L^\infty} \le \ (1-\eta(\lambda - C))^{k-1-\ell}\,\tfrac{2RC\eta}{n}.$$

Hence,

$$\sum_{\ell=0}^{k-1}\sum_{i=1}^{n}\left(c_i^\ell\right)^2 \le \frac{4R^2C^2}{n}\eta^2\sum_{\ell=0}^{k-1}(1-\eta(\lambda-C))^{2\ell}. \tag{C.15}$$

For $\eta < \eta_c$ and $\lambda > \lambda_c = 2C$, the geometric series in (C.15) is bounded by

$$\frac{4R^2C^2}{n}\eta^2\frac{1}{1-(1-\eta(\lambda-C))^2} = \frac{4R^2C^2}{n}\eta^2\frac{1}{(2-\eta(\lambda-C))\eta(\lambda-C)} \le \frac{4R^2C\eta}{n}. \tag{C.16}$$

McDiarmid's inequality yields (C.12). □

Corollary C.5 (High-probability bounds)**.** *Under the assumptions of Theorem C.4, for every $\beta > 0$:*

(i) *In the finite-horizon regime, with probability at least $1 - 2e^{-\beta}$,*

$$\sup_{k \le k_c}\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \le 3n^{-1/d} + C_1\eta^{1/2}\beta^{1/2}n^{-1/2}.$$

(ii) *In the uniform-in-time regime, with probability at least $1 - 2e^{-\beta}$,*

$$\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \le n^{-1/d} + C_1\eta^{1/2}\beta^{1/2}n^{-1/2}.$$

Proof. The proof is a direct combination of Theorem C.3 and Theorem C.4. □

C.4 Proof of Theorem 2.2

Choose $\beta = \eta^{-1}n^{1-2/d}$ in Corollary C.5. In statement (i), $C_1\eta^{1/2}\beta^{1/2}n^{-1/2} = C_1n^{-1/d}$, so $\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \le Cn^{-1/d}$ with probability at least $1 - 2\exp(-\eta^{-1}n^{1-2/d})$, which is (2.8). In statement (ii), $C_1\eta^{1/2}\beta^{1/2}n^{-1/2} = Cn^{-1/d}$, and the probability bound becomes $1 - 2\exp(-\eta^{-1}n^{1-2/d})$, which is (2.10).

For (2.9) and (2.11) in Theorem 2.2, they follow from Theorem D.4 if we choose β such that $2e^{-\beta} = \delta$. □

D Sharp Concentration Rate Under Assumption B

In this section, we prove the sharp $n^{-1/2}$ rate in Theorem 2.1 and 2.2. For any $(x_0, \mu_0) \in \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d)$, we use (x_s, μ_s) to denote the solution to the contextual flow (1.1) with initial condition (x_0, μ_0). The solutions to the adjoint system in Theorem 3.1 are denoted as (p_s, ϕ_s), representing the token adjoint and the measure adjoint respectively. For the random empirical measure $\widehat{\mu}_0$ sampled from μ_0, i.e.,

$$\widehat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n}\delta_{z^{(i)}}, \qquad z^{(i)} \overset{\text{iid}}{\sim} \mu_0,$$

we use $(\widehat{x}_s, \widehat{\mu}_s)$ to denote the solution to the contextual flow (1.1) with initial condition $(x_0, \widehat{\mu}_0)$ and the parameter path $\widehat{\theta}(s)$. Similarly, we use $(\widehat{p}_s, \widehat{\phi}_s)$ to denote the token adjoint and the measure adjoint respectively for the adjoint system in Theorem 3.1 with $(x_0, \widehat{\mu}_0, \widehat{\theta}(s))$. We will finally choose $\widehat{\theta}(s)$ as those $\widehat{\theta}_k$ in Theorem 2.2. In this section, we only need to keep in mind that $\widehat{\theta}(s)$ is a random parameter path depending on $\widehat{\mu}_0$ implicitly.

Throughout this section, we use Assumption D.1, which can be implied directly from Assumption B.

Assumption D.1. Let

$$\mathscr{A} := \left\{ \mathcal{V},\ \tfrac{\delta \mathcal{V}}{\delta \mu},\ D_x \mathcal{V},\ D_\theta \mathcal{V},\ \mathbb{W}_\mu \mathcal{V} \right\} \tag{D.1}$$

denote the functional collections in Assumption A. For any $x, z \in \mathbb{R}^d, \mu, \nu \in \mathcal{P}(\mathbb{R}^d), \theta \in \mathbb{R}^d$ satisfying Item A1, and for any $A \in \mathscr{A}$, we assume that

$$\|A(x, z, \mu; \theta) - A(x, z, \nu; \theta)\| \leq M_4 \sum_{h=1}^{H} \left| \int_{\mathbb{R}^d} f_h(x, z, y; \theta)\, \mathrm{d}(\mu - \nu)(y) \right|, \tag{D.2}$$

for some constant $M_4 > 0$ and $H \in \mathbb{Z}_+$. Also, there is an $L_4 > 0$, such that for each $h \in [\![1, H]\!]$, $f_h(x, y, z; \theta)$ is a Lipschitz function or Lipschitz map, and for any $x, z, y, \widetilde{x}, \widetilde{z}, \widetilde{y} \in \mathbb{R}^d$ and any $\theta, \widetilde{\theta} \in \mathbb{R}^p$, we have that

$$|f_h(x, z, y; \theta)| \leq L_4 \left(|x| + |z| + |y| + |\theta| + 1 \right), \tag{D.3}$$

and

$$\left| f_h(x, z, y; \theta) - f_h(\widetilde{x}, \widetilde{z}, \widetilde{y}; \widetilde{\theta}) \right| \leq L_4 \left(|x - \widetilde{x}| + |z - \widetilde{z}| + |y - \widetilde{y}| + |\theta - \widetilde{\theta}| \right). \tag{D.4}$$

Here, for $A \in \mathscr{A}$ without the z component, i.e., $\mathcal{V},\ D_x \mathcal{V},\ D_\theta \mathcal{V}$, we omit those z components in the above (D.3) and (D.4).

Theorem D.2. *Adopt the settings and assumptions in Theorem 2.1 and additionally adopt Assumption D.1. There exists a constant $C = C(\mathbf{A}, M_4, L_4, H)$ such that*

$$\mathbb{E}\Big[\sup_{s \in [0,1]} |x_s - \widehat{x}_s| \Big] \ \leq\ C(n^{-1/2} + \mathbb{E}\|\theta - \widehat{\theta}\|_{L^1}). \tag{D.5}$$

What's more, if $\widehat{\theta}(s) = \theta(s)$, then for every $\beta > 0$, with probability at least $1 - 2e^{-\beta}$,

$$\sup_{s \in [0,1]} \left| x_s - \widehat{x}_s \right| \ \leq\ C\,(1 + \beta^{1/2}) n^{-1/2}. \tag{D.6}$$

Proof. Similar to the proof of Theorem A.2, (D.6) follows directly once we obtain (D.5). We then only show (D.5).

Recall that

$$\begin{cases} \dot{x}_s = \mathcal{V}(x_s, \mu_s; \theta(s)), \quad \dot{\widehat{x}}_s = \mathcal{V}(\widehat{x}_s, \widehat{\mu}_s; \widehat{\theta}(s)), \\ \partial_s \mu_s + \nabla \cdot \big(\mu_s\, \mathcal{V}(\cdot, \mu_s; \theta(s)) \big) = 0, \quad \partial_s \widehat{\mu}_s + \nabla \cdot \big(\widehat{\mu}_s\, \mathcal{V}(\cdot, \widehat{\mu}_s; \widehat{\theta}(s)) \big) = 0, \\ x_s|_{s=0} = x_0, \quad \widehat{x}_s|_{s=0} = x_0, \\ \mu_s|_{s=0} = \mu_0, \quad \widehat{\mu}_s|_{s=0} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \delta_{z^{(i)}}. \end{cases}$$

We consider an auxiliary flow solving the system

$$\begin{cases} \dot{\widetilde{x}}_s = \mathcal{V}(\widetilde{x}_s, \widetilde{\mu}_s; \theta(s)), \\ \partial_s \widetilde{\mu}_s + \nabla \cdot \big(\widetilde{\mu}_s\, \mathcal{V}(\cdot, \widetilde{\mu}_s; \theta(s)) \big) = 0, \\ \widetilde{x}_s|_{s=0} = x_0, \\ \widetilde{\mu}_s|_{s=0} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \delta_{z^{(i)}}. \end{cases}$$

By Theorem 4.1 applied to $(\widetilde{\mu}_s, \widehat{\mu}_s)$ and $(\widetilde{x}_s, \widehat{x}_s)$, we take supremum on (4.1) and (4.2) and then take expectations, we have that for some $C = C(\mathbf{A})$,

$$\mathbb{E} \sup_{s\in[0,1]} \mathrm{W}_1(\widetilde{\mu}_s, \widehat{\mu}_s) \le C\mathbb{E}\|\theta - \widehat{\theta}\|_{L^1} \quad \text{and} \quad \mathbb{E}\Big[\sup_{s\in[0,1]} |\widetilde{x}_s - \widehat{x}_s|\Big] \le C\mathbb{E}\|\theta - \widehat{\theta}\|_{L^1}.$$

Hence, we only need to show that for some $C = C(\mathbf{A}, M_4, L_4, H)$,

$$\mathbb{E}\Big[\sup_{s\in[0,1]} |x_s - \widetilde{x}_s|\Big] \le Cn^{-1/2}.$$

Consider the characteristic ODEs associated with (x_s, μ_s) and $(\widetilde{x}_s, \widetilde{\mu}_s)$: let $\Phi_s, \widetilde{\Phi}_s : \mathbb{R}^d \to \mathbb{R}^d$ solve

$$\begin{cases} \partial_s \Phi_s(x) = \mathcal{V}(\Phi_s(x), \mu_s; \theta(s)), \\ \Phi_0(x) = x, \end{cases} \qquad \begin{cases} \partial_s \widetilde{\Phi}_s(x) = \mathcal{V}(\widetilde{\Phi}_s(x), \widetilde{\mu}_s; \theta(s)), \\ \widetilde{\Phi}_0(x) = x. \end{cases} \tag{D.9}$$

Then $x_s = \Phi_s(x_0)$, $\widetilde{x}_s = \widetilde{\Phi}_s(x_0)$, and $\mu_s = (\Phi_s)_\#(\mu_0)$, $\widetilde{\mu}_s = (\widetilde{\Phi}_s)_\#(\widetilde{\mu}_0)$. Also, $\widetilde{\mu}_s$ actually has the form

$$\widetilde{\mu}_s = \frac{1}{n}\sum_{i=1}^n \delta_{\widetilde{\Phi}_s(z^{(i)})}.$$

We define the auxiliary empirical measure as

$$\overline{\mu}_s \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{\Phi_s(z^{(i)})}. \tag{D.10}$$

Then, by Assumption A,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s}|x_s - \widetilde{x}_s| &\le |\mathcal{V}(x_s, \mu_s; \theta(s)) - \mathcal{V}(\widetilde{x}_s, \widetilde{\mu}_s; \theta(s))| \\ &\le |\mathcal{V}(x_s, \mu_s; \theta(s)) - \mathcal{V}(x_s, \widetilde{\mu}_s; \theta(s))| + |\mathcal{V}(x_s, \widetilde{\mu}_s; \theta(s)) - \mathcal{V}(\widetilde{x}_s, \widetilde{\mu}_s; \theta(s))| \\ &\le |\mathcal{V}(x_s, \mu_s; \theta(s)) - \mathcal{V}(x_s, \widetilde{\mu}_s; \theta(s))| + L|x_s - \widetilde{x}_s| \\ &\le \underbrace{|\mathcal{V}(x_s, \mu_s; \theta(s)) - \mathcal{V}(x_s, \overline{\mu}_s; \theta(s))|}_{:=A_s} + \underbrace{|\mathcal{V}(x_s, \overline{\mu}_s; \theta(s)) - \mathcal{V}(x_s, \widetilde{\mu}_s; \theta(s))|}_{:=B_s} \\ &\quad + L|x_s - \widetilde{x}_s|. \end{aligned} \tag{D.11}$$

A direct Grönwall argument shows that

$$\sup_{s\in[0,1]} |x_s - \widetilde{x}_s| \le e^L \int_0^1 A_s + B_s \,\mathrm{d}s,$$

and thus

$$\mathbb{E}\Big[\sup_{s\in[0,1]} |x_s - \widetilde{x}_s|\Big] \le e^L \int_0^1 \mathbb{E}A_s + \mathbb{E}B_s \,\mathrm{d}s$$

We then estimate $\mathbb{E}A_s$ and $\mathbb{E}B_s$ in respectively. For $\mathbb{E}A_s$, we notice that, by Assumption D.1,

$$\begin{aligned} A_s &= |\mathcal{V}(x_s, \mu_s; \theta(s)) - \mathcal{V}(x_s, \overline{\mu}_s; \theta(s))| \le M_4 \sum_{h=1}^H \left| \int_{\mathbb{R}^d} f_h(x_s, y; \theta(s)) \,\mathrm{d}(\mu_s - \overline{\mu}_s)(y) \right| \\ &= M_4 \sum_{h=1}^H \left| \int_{\mathbb{R}^d} f_h(x_s, y; \theta(s)) \,\mathrm{d}(\Phi_s)_\#(\mu_0 - \overline{\mu}_0)(y) \right| \\ &= M_4 \sum_{h=1}^H \left| \int_{\mathbb{R}^d} f_h(x_s, \Phi_s(y); \theta(s)) \,\mathrm{d}(\mu_0 - \overline{\mu}_0)(y) \right|. \end{aligned}$$

For each $h \in [\![1, H]\!]$, we define random variables

$$\xi_{h,j} \coloneqq f_h(x_s, \Phi_s(z^{(j)}); \theta(s)), \quad j = 1, 2, \ldots, n,$$

and $\overline{\xi}_h \coloneqq \frac{1}{n}\sum_{j=1}^n \xi_{h,j}$. We know that $z^{(1)}, \dots, z^{(n)} \overset{\text{iid}}{\sim} \mu_0$. Thus,

$$\mathbb{E}\left|\int_{\mathbb{R}^d} f_h(x_s, \Phi_s(y); \theta(s))\, \mathrm{d}(\mu_0 - \overline{\mu}_0)(y)\right| = \mathbb{E}\left|\mathbb{E}(\overline{\xi}_h) - \overline{\xi}_h\right| \le \sqrt{\mathrm{Var}\left(\overline{\xi}_h\right)} = n^{-1/2}\sqrt{\mathrm{Var}\left(\xi_{h,1}\right)}, \tag{D.12}$$

and we have that $\mathrm{Var}\left(\xi_{h,1}\right) \le C$ for some $C = C(\mathbf{A}, L_4)$ independent of d according to Assumption D.1, Item A1, and Theorem A.1. Thus, for some $C = C(\mathbf{A}, M_4, L_4, H)$, $\mathbb{E}A_s \le Cn^{-1/2}$.

Then, we estimate $\mathbb{E}B_s$. By Assumption D.1,

$$\begin{aligned}
B_s &= \left|\mathcal{V}(x_s, \overline{\mu}_s; \theta(s)) - \mathcal{V}(x_s, \widetilde{\mu}_s; \theta(s))\right| \le M_4 \sum_{h=1}^H \left|\int_{\mathbb{R}^d} f_h(x_s, y; \theta(s))\, \mathrm{d}(\overline{\mu}_s - \widetilde{\mu}_s)(y)\right| \\
&= \frac{M_4}{n}\sum_{h=1}^H \left|\sum_{i=1}^n \left(f_h(x_s, \Phi_s(z^{(i)}); \theta(s)) - f_h(x_s, \widetilde{\Phi}_s(z^{(i)}); \theta(s))\right)\right| \\
&\le \frac{M_4 H L_4}{n}\sum_{i=1}^n \left|\Phi_s(z^{(i)}) - \widetilde{\Phi}_s(z^{(i)})\right|.
\end{aligned}$$

Denote $W_s = \frac{1}{n}\sum_{i=1}^n \left|\Phi_s(z^{(i)}) - \widetilde{\Phi}_s(z^{(i)})\right|$. Clearly, we have that

$$\mathrm{W}_1\left(\overline{\mu}_s, \widetilde{\mu}_s\right) \le W_s.$$

Then, similar to (D.11), we have that, for each $i \in [\![1, n]\!]$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}s}\left|\Phi_s(z^{(i)}) - \widetilde{\Phi}_s(z^{(i)})\right| \le \left|\mathcal{V}(\Phi_s(z^{(i)}), \mu_s; \theta(s)) - \mathcal{V}(\widetilde{\Phi}_s(z^{(i)}), \widetilde{\mu}_s; \theta(s))\right| \\
&\le \left|\mathcal{V}(\Phi_s(z^{(i)}), \mu_s; \theta(s)) - \mathcal{V}(\Phi_s(z^{(i)}), \overline{\mu}_s; \theta(s))\right| + \left|\mathcal{V}(\Phi_s(z^{(i)}), \overline{\mu}_s; \theta(s)) - \mathcal{V}(\widetilde{\Phi}_s(z^{(i)}), \widetilde{\mu}_s; \theta(s))\right| \\
&\le \left|\mathcal{V}(\Phi_s(z^{(i)}), \mu_s; \theta(s)) - \mathcal{V}(\Phi_s(z^{(i)}), \overline{\mu}_s; \theta(s))\right| + L_0\left(\left|\Phi_s(z^{(i)}) - \widetilde{\Phi}_s(z^{(i)})\right| + \mathrm{W}_1\left(\overline{\mu}_s, \widetilde{\mu}_s\right)\right) \\
&\le \left|\mathcal{V}(\Phi_s(z^{(i)}), \mu_s; \theta(s)) - \mathcal{V}(\Phi_s(z^{(i)}), \overline{\mu}_s; \theta(s))\right| + L_0\left(\left|\Phi_s(z^{(i)}) - \widetilde{\Phi}_s(z^{(i)})\right| + W_s\right)
\end{aligned}$$

where we applied (A3). Thus,

$$\sup_{s\in[0,1]} W_s \le \frac{e^{2L_0}}{n}\sum_{i=1}^n \int_0^1 \left|\mathcal{V}(\Phi_s(z^{(i)}), \mu_s; \theta(s)) - \mathcal{V}(\Phi_s(z^{(i)}), \overline{\mu}_s; \theta(s))\right| \mathrm{d}s$$

Similar to the estimates for $\mathbb{E}A_s$, use Assumption D.1, we have that

$$\mathbb{E}\sup_{s\in[0,1]} W_s \le Cn^{-1/2}, \tag{D.13}$$

for some $C = C(\mathbf{A}, M_4, L_4, H)$. The only mild modification is that we need to condition on $z^{(i)}$, and then only the random variables $f_h(\Phi_s(z^{(i)}), \Phi_s(z^{(j)}); \theta(s))$, with $j \ne i$ are i.i.d.. But the contribution of the term $f_h(\Phi_s(z^{(i)}), \Phi_s(z^{(i)}); \theta(s))$ in (D.12) is n^{-1}, so we can still get the $n^{-1/2}$ rate in the expectation. Thus, we have that $\mathbb{E}B_s \le M_4 H L_4 \mathbb{E}W_s \le Cn^{-1/2}$.

Combine the estimates for $\mathbb{E}A_s$ and $\mathbb{E}B_s$, we obtain (D.5). □

We remark that one important by-product in the estimates of the term $\mathbb{E}B_s$ is that

$$\mathbb{E}\sup_{s\in[0,1]} \mathrm{W}_1\left(\overline{\mu}_s, \widehat{\mu}_s\right) \le \mathbb{E}\sup_{s\in[0,1]} W_s + \mathbb{E}\sup_{s\in[0,1]} \mathrm{W}_1(\widetilde{\mu}_s, \widehat{\mu}_s) \le C(n^{-1/2} + \mathbb{E}\|\theta - \widehat{\theta}\|_{L^1}), \tag{D.14}$$

which will be essential in the proof of the following Theorem D.3 for the expectations of the adjoint system in Theorem 3.1. That is, (3.1), (3.2), and (3.3), also satisfy this sharp $n^{-1/2}$-rate.

Theorem D.3. *Adopt the settings and assumptions in Theorem 3.1 and additionally adopt Assumption B. There exists a constant $C = C(\mathbf{A}, L_4)$ such that*

$$\mathbb{E}\left[\sup_{s\in[0,1]}|p_s - \widehat{p}_s|\right] \le C(n^{-1/2} + \mathbb{E}\|\theta - \widehat{\theta}\|_{L^1}), \tag{D.15}$$

and

$$\mathbb{E}\left[\sup_{s\in[0,1],z\in B_R(0)}|\nabla\phi_s(z) - \nabla\widehat{\phi}_s(z)|\right] \le C\left(d^{1/2}n^{-1/2} + \mathbb{E}\|\theta - \widehat{\theta}\|_{L^1}\right). \tag{D.16}$$

Proof. Adopt the $(\widetilde{x}_s, \widetilde{\mu}_s)$ defined in the proof of Theorem D.2. First, by Proposition B.1, we can assume that both $p_s, \widehat{p}_s$ are bounded by a $C = C(\mathbf{A})$. For (D.15), by (3.2) and those Lipschitz assumptions in Assumption A, we have that

$$\begin{aligned}|p_s - \widehat{p}_s| \le |p_1 - \widehat{p}_1| + \int_s^1 \Big\{&\big|[D_x\mathcal{V}(x_\tau, \mu_\tau; \theta(\tau)) - D_x\mathcal{V}(x_\tau, \overline{\mu}_\tau; \theta(\tau))]^\top p_\tau\big| \\ &+ C\big(\mathrm{W}_1(\overline{\mu}_\tau, \widehat{\mu}_\tau) + |x_\tau - \widehat{x}_\tau| + |p_\tau - \widehat{p}_\tau| + |\theta(\tau) - \widehat{\theta}(\tau)|\big)\Big\}\,\mathrm{d}\tau,\end{aligned}$$

for some $C = C(\mathbf{A})$. Then, (D.15) follows from combining (D.5) and a similar argument for the estimates of the terms $\mathbb{E}A_s, \mathbb{E}B_s$ in the proof of Theorem D.2.

Next, similar to the proof of Proposition B.3, we define $\Gamma_s(z) \coloneqq \widehat{\phi}_s(z) - \phi_s(z)$. Subtracting the two equations (3.3) and rearranging,

$$\partial_s\Gamma_s(z) + \nabla_z(\Gamma_s(z))\cdot\mathcal{V}(z, \widehat{\mu}_s; \widehat{\theta}(s)) = h(s,z) + g(s,z),$$

with $\Gamma_1 \equiv 0$ and source terms

$$h(s,z) = -\int_{\mathbb{R}^d}\nabla(\Gamma_s(\xi))\cdot\tfrac{\delta\mathcal{V}}{\delta\mu}[\xi, \widehat{\mu}_s; \widehat{\theta}(s)](z)\mathrm{d}\widehat{\mu}_s(\xi),$$

and

$$\begin{aligned}g(s,z) = &-\nabla\phi_s(z)\cdot\big(\mathcal{V}(z, \widehat{\mu}_s; \widehat{\theta}(s)) - \mathcal{V}(z, \mu_s; \theta(s))\big) \\ &- \big(\widehat{p}_s\cdot\tfrac{\delta\mathcal{V}}{\delta\mu}[\widehat{x}_s, \widehat{\mu}_s; \widehat{\theta}(s)](z) - p_s\cdot\tfrac{\delta\mathcal{V}}{\delta\mu}[x_s, \mu_s; \theta(s)](z)\big) \\ &- \int_{\mathbb{R}^d}\nabla\phi_s(\xi)\cdot\big(\tfrac{\delta\mathcal{V}}{\delta\mu}[\xi, \widehat{\mu}_s; \widehat{\theta}(s)](z)\,\mathrm{d}\widehat{\mu}_s(\xi) - \tfrac{\delta\mathcal{V}}{\delta\mu}[\xi, \mu_s; \theta(s)](z)\,\mathrm{d}\mu_s(\xi)\big).\end{aligned} \tag{D.17}$$

We used similar notations $\Gamma_s(z)$ and $g(s,z)$ as in the proof of Proposition B.3.

Similar to the proof of Proposition B.2, we let $\widehat{\Psi}_\tau(z)$ be the solution from depth s of $\partial_\tau\widehat{\Psi}_\tau = \mathcal{V}(\widehat{\Psi}_\tau, \widehat{\mu}_\tau; \widehat{\theta}(\tau))$ with $\widehat{\Psi}_s(z) = z, \forall z \in \mathbb{R}^d$, we have that

$$\Gamma_s(z) \;=\; -\int_s^1 h(\tau, \widehat{\Psi}_\tau(z)) + g(\tau, \widehat{\Psi}_\tau(z))\,\mathrm{d}\tau. \tag{D.18}$$

Denote

$$\|\Gamma\|_R \coloneqq \sup_{s\in[0,1],z\in B_R(0)} e^{\alpha(s-1)}|\nabla\Gamma_s(z)|.$$

Because $\operatorname{supp}\widehat{\mu}_s \subseteq B_R(0)$ according to (A1), we have that $\|h(s, \widehat{\Psi}_s(z))\|_R \le C\|\Gamma\|_R$ for some $C = C(\mathbf{A})$ using those Lipschitz assumptions on $\mathcal{V}$ in Assumption A. Thus, by (D.18)

$$\begin{aligned}\|\Gamma\|_R &\le \sup_{s\in[0,1],z\in B_R(0)} e^{\alpha(s-1)}\int_s^1 e^{\alpha(1-\tau)}C\|\Gamma\|_R + C|\nabla g(\tau, \widehat{\Psi}_\tau(z))|\,\mathrm{d}\tau \\ &\le \frac{C}{\alpha}\|\Gamma\|_R + \sup_{s\in[0,1],z\in B_R(0)}\int_s^1 C|\nabla g(\tau, \widehat{\Psi}_\tau(z))|\,\mathrm{d}\tau.\end{aligned}$$

Thus, if we take $\alpha = 2C$, and take expectations on the above inequality, we get that

$$\begin{aligned}\mathbb{E}\|\Gamma\|_R &\le 2C\mathbb{E}\sup_{s\in[0,1],z\in B_R(0)}\int_s^1 |\nabla g(\tau,\widehat{\Psi}_\tau(z))|\,\mathrm{d}\tau\\ &\le 2C\int_0^1 \mathbb{E}\sup_{z\in B_{R'}(0)}|\nabla g(\tau,z)|\,\mathrm{d}\tau,\end{aligned}$$

where in the second inequality, $R' = R'(\mathbf{A})$ obtained in (A.1) of Theorem A.1 to control the norm of $|\widehat{\Psi}_\tau(z)|$ for $z \in B_R(0)$. After bounding the Hessian factors by the $C^{1,1}$ estimate, the remaining empirical kernels are Lipschitz in z under Assumptions A and B. Hence, similar to the proof for Theorem D.2, we combine with the Dudleys entropy integral among the class of Lipschitz functions $\{|\nabla g(\tau,z)|\}_{z\in B_{R'}(0)}$ (see for example [VDVW96, Chapter 2] and [Kos08, Chapter 2]), we have a $C = C(\mathbf{A})$, such that

$$\mathbb{E}\sup_{z\in B_{R'}(0)}|\nabla g(\tau,z)| \le Cd^{1/2}n^{-1/2} + \mathbb{E}|\theta(\tau)-\widehat{\theta}(\tau)| + C\mathbb{E}\|\theta-\widehat{\theta}\|_{L^1}.$$

For example, by the form of $g(\tau,z)$ in (D.17), we need to estimate the term

$$\begin{aligned}&\mathbb{E}\sup_{z\in B_{R'}(0)}\left\|D_x\mathcal{V}(z,\widehat{\mu}_\tau;\widehat{\theta}(\tau)) - D_x\mathcal{V}(z,\mu_\tau;\theta(\tau))\right\|\\ &\le \mathbb{E}\sup_{z\in B_{R'}(0)}\|D_x\mathcal{V}(z,\overline{\mu}_\tau;\theta(\tau)) - D_x\mathcal{V}(z,\mu_\tau;\theta(\tau))\| + C\left(\mathbb{E}\mathrm{W}_1\left(\overline{\mu}_\tau,\widehat{\mu}_\tau\right) + \mathbb{E}|\theta(\tau)-\widehat{\theta}(\tau)|\right),\end{aligned}$$

where $C = C(\mathbf{A})$. Combine Assumption D.1 for $D_x\mathcal{V}$ and (D.14), together with the Dudley's entropy integral, we can bound the above term by $C(\sqrt{d/n} + \mathbb{E}|\theta(\tau)-\widehat{\theta}(\tau)| + \mathbb{E}\|\theta-\widehat{\theta}\|_{L^1})$. Other terms in $g(\tau,z)$ in (D.17) can be estimated similarly, combining with (D.15) and (D.5). This finishes the proof for (D.16). □

Theorem D.4. *Adopt the settings and assumptions in Theorem 2.2 and additionally adopt Assumption B. There exist constants C, η_c, λ_c depending on $\mathbf{A}, L_4$, such that the following hold.*

(i) Small or no ridge penalty. *For $\eta > 0$, $\lambda \in [0,\eta^{-1})$, we let $k_c := (\log(1+C\eta))^{-1}$. Then,*

$$\sup_{k\le k_c}\left\|\theta_k - \widehat{\theta}_k\right\|_{L^\infty} \;\le\; C\left(1+\eta^{1/2}\beta^{1/2}\right)d^{1/2}n^{-1/2}, \tag{D.19}$$

with probability at least $1-2e^{-\beta}$.

(ii) Large ridge penalty. *If $\eta < \eta_c$ and $\lambda \in (\lambda_c,\eta^{-1})$, then for any $k \in \mathbb{N}$,*

$$\left\|\theta_k - \widehat{\theta}_k\right\|_{L^\infty} \;\le\; C\left(1+\eta^{1/2}\beta^{1/2}\right)d^{1/2}n^{-1/2}, \tag{D.20}$$

with probability at least $1-2e^{-\beta}$.

Proof. According to the proof of Corollary C.5 we only need to modify Theorem C.3 and show that

$$\mathbb{E}\|\theta_k - \widehat{\theta}_k\|_{L^\infty} \le 3d^{1/2}n^{-1/2},$$

in both the small or no ridge penalty case and the large ridge penalty case. We only discuss the small ridge penalty case in the following, and the case for the large ridge penalty is very similar to Theorem C.3.

Set $\Delta_l := \|\widehat{\theta}_l - \theta_l\|_{L^\infty}$. Subtracting the two OGD updates (2.3), similar to (C.6), we first take $\sup_{s\in[0,1]}$ then take $\sup_{k\le k_c}$, and finally apply expectations. Combine Theorem D.2, (D.14), Theorem D.3, and the arguments in Theorem C.3 to expand those error terms, we can obtain that

$$\mathbb{E}\sup_{k\le k_c}\Delta_k \le \eta C d^{1/2}n^{-1/2}\sum_{l=0}^{k_c-1}(1+\eta C)^{k_c-1-l} \le (1+\eta C)^{k_c}d^{1/2}n^{-1/2}. \tag{D.21}$$

We remark some details when subtracting the two OGD updates (2.3) for each $l = 0, 1, \ldots, k_c$. Similar to the arguments in Theorem C.3, the main error term comes from

$$\begin{aligned}
&\mathbb{E}\left\|\frac{\delta\mathcal{L}(\widehat{\theta}_l;\widehat{\mu}_0^l,x_0^l,y_0^l)}{\delta\theta}(s)-\frac{\delta\mathcal{L}(\theta_l;\mu_0^l,x_0^l,y_0^l)}{\delta\theta}(s)\right\|_{L^\infty}\\
&\le \underbrace{\mathbb{E}\left\|D_\theta\mathcal{V}(\widehat{x}_s^l,\widehat{\mu}_s^l;\widehat{\theta}_l(s))^\top\widehat{p}_s^l-D_\theta\mathcal{V}(x_s^l,\mu_s^l;\theta_l(s))^\top p_s^l\right\|_{L^\infty}}_{\text{Term 1}}\\
&\quad+\underbrace{\mathbb{E}\left\|\int_{\mathbb{R}^d}D_\theta\mathcal{V}(z,\widehat{\mu}_s^l;\widehat{\theta}_l(s))^\top\nabla\widehat{\phi}_s^l(z)\,\mathrm{d}\widehat{\mu}_s^l(z)-D_\theta\mathcal{V}(z,\mu_s^l;\theta_l(s))^\top\nabla\phi_s^l(z)\,\mathrm{d}\mu_s^l(z)\right\|_{L^\infty}}_{\text{Term 2}}.
\end{aligned}$$

The Term 1 on the right hand side can be estimated by combining (D.5), (D.14), and Theorem D.3. For the Term 2, we can decompose it by

$$\begin{aligned}
&\mathbb{E}\left\|\int_{\mathbb{R}^d}D_\theta\mathcal{V}(z,\widehat{\mu}_s^l;\widehat{\theta}_l(s))^\top\nabla\widehat{\phi}_s^l(z)\,\mathrm{d}\widehat{\mu}_s^l(z)-D_\theta\mathcal{V}(z,\mu_s^l;\theta_l(s))^\top\nabla\phi_s^l(z)\,\mathrm{d}\mu_s^l(z)\right\|_{L^\infty}\\
&\le C\Bigg(\mathbb{E}\Delta_l+\mathbb{E}\sup_{s\in[0,1],z\in B_R(0)}\left\|D_\theta\mathcal{V}(z,\overline{\mu}_s^l;\theta_l(s))-D_\theta\mathcal{V}(z,\mu_s^l;\theta_l(s))\right\|\\
&\quad+\mathbb{E}\left[\sup_{s\in[0,1],z\in B_R(0)}|\nabla\phi_s^l(z)-\nabla\widehat{\phi}_s^l(z)|\right]\\
&\quad+\mathbb{E}\sup_{s\in[0,1]}\left|\int_{\mathbb{R}^d}D_\theta\mathcal{V}(z,\mu_s^l;\theta_l(s))^\top\nabla\phi_s^l(z)\,\mathrm{d}(\overline{\mu}_s^l-\mu_s^l)(z)\right|+\mathbb{E}\sup_{s\in[0,1]}\mathrm{W}_1(\overline{\mu}_s^l,\widehat{\mu}_s^l)\Bigg),
\end{aligned}$$

for some $C = C(\mathbf{A})$. Similar to the proofs of Theorems B.4 and C.2, the Lipschitz bound in (A1) propagates uniformly along OGD in the respective regimes. The Dudley argument in the proof of Theorem D.3 therefore applies jointly in depth and space, and together with (D.5) and (D.14), bounds Term 2 by $C(\mathbb{E}\Delta_l + d^{1/2}n^{-1/2})$ for some $C = C(\mathbf{A}, L_4)$.

After obtaining (D.21), the remaining steps are exactly the same as the proofs in Theorem C.3. □

E Verification of Assumption A and B for transformer architectures

This appendix verifies that the standard transformer building blocks—self-attention layers (1.2) and MLP layers—satisfy Assumptions (A3), (A4), and (B), with constants that depend only on the support radius R in (A1), the parameter bound M, and (for MLPs) the activation regularity. In particular all constants are independent of the embedding dimension d and of the parameter dimension p.

E.1 Self-attention blocks

Recall that for a single-head self-attention block, the velocity field (1.2) is

$$\mathcal{V}(x,\mu;\theta) \;=\; \frac{1}{Z(x)}\int_{\mathbb{R}^d}e^{\langle Qx,Ky\rangle}Vy\,\mathrm{d}\mu(y),\qquad Z(x)=\int_{\mathbb{R}^d}e^{\langle Qx,Kz\rangle}\mathrm{d}\mu(z),\tag{E.1}$$

with $\theta = (Q, K, V) \in \mathbb{R}^{3d^2}$. We first establish uniform boundedness along the contextual flow.

Proposition E.1 (Boundedness of the transformer flow)**.** *Adopt* (A1) *for θ and μ_0. Let μ_s solve* (1.1)*. Then for all $s \in [0, 1]$,*

$$\operatorname{supp}\mu_s \subseteq B_{e^M R}(0).$$

Proof. Let $R_\tau := \sup\{|x| \mid x \in \operatorname{supp}\mu_\tau\}$, then $R_0 \le R$. First, because $\forall s \in [0,1]$, $|\theta(s)| \le M$, the Frobenius norms $\|Q\|_F, \|K\|_F, \|V\|_F$ are bounded by M. Thus $\|Q\|, \|K\|, \|V\|$ are also bounded by M. Hence, $\forall x \in \mathbb{R}^d$, $\forall \tau \in [0,1]$,

$$|\mathcal{V}(x,\mu_\tau;\theta)| \le \frac{1}{Z(x)}\int_{\mathbb{R}^d}\exp\left(\langle Qx,Ky\rangle\right)\|V\||y|\,\mathrm{d}\mu_\tau(y) \le MR_\tau$$

Fix a $\tau > 0$, let Ψ be the characteristic flow map of $\mathcal{V}(z, \mu_s; \theta(s))$ from τ to 1, i.e.,

$$\partial_s \Psi_s(z) = \mathcal{V}(\Psi_s(z), \mu_s; \theta(s)), \quad \Psi_\tau(z) = z.$$

Let $z_\tau \in \mathbb{R}^d$ be a point that attains R_τ, then

$$\partial_s |\Psi_s(z_\tau)|\Big|_{s=\tau} \leq |(\partial_s \Psi_s)(z_\tau)|\Big|_{s=\tau} = |\mathcal{V}(z_\tau, \mu_\tau; \theta(\tau))| \leq M R_\tau.$$

Because this is true for any z_τ that attains R_τ, we must have that

$$\partial_s R_s|_{s=\tau} \leq M R_\tau.$$

Hence, by a Grönwall argument, $R_s \leq e^{Ms} R_0$, $\forall s \in [0, 1]$.

□

We may therefore verify Assumptions (A3) and (A4) on $B_R(0)$, treating R as the support radius (after the substitution $R \leftarrow e^M R$ if needed). For notational simplicity we keep R throughout.

Notation. Set

$$\alpha(x, y) \coloneqq \frac{e^{\langle Qx, Ky \rangle}}{Z(x)}, \qquad \mathcal{U}(x) \coloneqq \int_{\mathbb{R}^d} \alpha(x, y)\, y \, \mathrm{d}\mu(y),$$

$$\mathcal{M}(x) \coloneqq \int_{\mathbb{R}^d} \alpha(x, y)\, (y - \mathcal{U}(x))(y - \mathcal{U}(x))^\top \, \mathrm{d}\mu(y),$$

so that $\mathcal{V}(x, \mu; \theta) = V\mathcal{U}(x)$ and $\|\mathcal{M}(x)\| \leq 4R^2$ when $x \in B_R(0)$ and $\operatorname{supp} \mu \subseteq B_R(0)$.

Proposition E.2 (Lipschitz continuity of self-attention)**.** *Adopt* (A1)*. There exist $M_0 = M_0(M, R)$ and $L_0 = L_0(M, R)$ such that $\mathcal{V}$ and $\delta\mathcal{V}/\delta\mu$ in* (E.1) *satisfy* (A3) *for $x, z \in B_R(0)$.*

Proof. Lipschitz in x. A direct computation gives

$$D_x \alpha(x, y) \;=\; \alpha(x, y) \big(y - \mathcal{U}(x)\big)^\top K^\top Q, \tag{E.2}$$

so that

$$D_x \mathcal{U}(x) \;=\; \mathcal{M}(x) K^\top Q, \tag{E.3}$$

and consequently

$$D_x \mathcal{V}(x, \mu; \theta) \;=\; V \mathcal{M}(x) K^\top Q, \qquad \|D_x \mathcal{V}\| \leq 4M^3 R^2. \tag{E.4}$$

Lipschitz in μ. A direct computation gives

$$\tfrac{\delta \mathcal{V}}{\delta \mu}[x, \mu; \theta](z) \;=\; \alpha(x, z)\, V \big(z - \mathcal{U}(x)\big),$$

and since $\nabla_z \alpha(x, z) = \alpha(x, z)\, K^\top Q x$,

$$\mathbb{W}_\mu \mathcal{V}(z) = \alpha(x, z) \big[V(z - \mathcal{U}(x))(K^\top Q x)^\top + V\big], \qquad \|\mathbb{W}_\mu \mathcal{V}\| \leq e^{2M^2R^2} M(2M^2R^2 + 1). \tag{E.5}$$

Lipschitz in $\theta = (Q, K, V)$. In the following, we use the notations

$$D_V \mathcal{V} \cdot [\Delta V] \coloneqq \lim_{t \to 0} \frac{\mathcal{V}(x, \mu; (Q, K, V + t\Delta V)) - \mathcal{V}(x, \mu; (Q, K, V))}{t},$$

where $\Delta V \in \mathbb{R}^{d \times d}$ is any $d \times d$ matrix. We similarly define $D_Q \mathcal{V} \cdot [\Delta Q]$ and $D_K \mathcal{V} \cdot [\Delta K]$. A direct computation gives

$$D_Q \alpha(x, y) \cdot [\Delta Q] = \alpha(x, y) \big(\langle \Delta Q\, x, Ky \rangle - \langle \Delta Q\, x, K\mathcal{U}(x) \rangle\big) = \alpha(x, y)\, \langle \Delta Q\, x,\, K(y - \mathcal{U}(x)) \rangle. \tag{E.6}$$

Thus, the directional derivatives of $\mathcal{V}$ in θ are

$$D_V \mathcal{V} \cdot [\Delta V] = \Delta V\, \mathcal{U}(x), \tag{E.7}$$

$$D_Q \mathcal{V} \cdot [\Delta Q] = V \mathcal{M}(x) K^\top \Delta Q\, x, \tag{E.8}$$

$$D_K \mathcal{V} \cdot [\Delta K] = V \mathcal{M}(x) (\Delta K)^\top Q\, x, \tag{E.9}$$

with $\|D_V \mathcal{V}\| \leq R$, $\|D_Q \mathcal{V}\|, \|D_K \mathcal{V}\| \leq 4M^2 R^3$, hence $\|D_\theta \mathcal{V}\| \leq R(1 + 8M^2R^2)$. The linear-growth bound for $\mathcal{V}$ on $B_R(0)$ follows from boundedness on $B_R(0)$. The same arguments applied to $\delta\mathcal{V}/\delta\mu$ yield the corresponding bounds. □

Proposition E.3 (Smoothness of $D_x\mathcal{V}$)**.** *Adopt* (A1)*. There exists* $L_1 = L_1(M, R)$ *such that* $D_x\mathcal{V}$ *in* (E.1) *satisfies* (A4) *(a) for* $x \in B_R(0)$.

Proof. We use the notation

$$D_x^2\mathcal{V} \cdot [h] \coloneqq \lim_{t\to 0} \frac{D_x\mathcal{V}(x+th,\mu;\theta) - D_x\mathcal{V}(x,\mu;\theta)}{t},$$

where $h \in \mathbb{R}^d$ is any d-dimensional vector. Thus, the second-order spatial derivative is

$$D_x^2\mathcal{V} \cdot [h] = V\int \alpha(x,y)\big[(y-\mathcal{U}(x))^\top K^\top Q h\big](y-\mathcal{U}(x))(y-\mathcal{U}(x))^\top \,\mathrm{d}\mu(y)\, K^\top Q,$$

so $\|D_x^2\mathcal{V}\| \le 8M^5R^3$.

The variation of $D_x\mathcal{V}$ in μ is computed using

$$\frac{\delta\alpha(x,y)}{\delta\mu}(z) = -\alpha(x,y)\,\alpha(x,z), \quad \frac{\delta\mathcal{U}(x)}{\delta\mu}(z) = \alpha(x,z)(z-\mathcal{U}(x)). \tag{E.10}$$

A direct computation then shows that

$$\frac{\delta(D_x\mathcal{V})}{\delta\mu}(z) = V\,\alpha(x,z)\big[(z-\mathcal{U}(x))\,(z-\mathcal{U}(x))^\top - \mathcal{M}(x)\big]\,K^\top Q,$$

and thus

$$\begin{aligned}
&\mathbb{W}_\mu(D_x\mathcal{V})(z)\cdot[h]\\
&= V\,\alpha(x,z)\Big[\big((z-\mathcal{U}(x))\,(z-\mathcal{U}(x))^\top - \mathcal{M}(x)\big)\langle K^\top Q x,\, h\rangle\\
&\quad + (z-\mathcal{U}(x))\,h^\top + h\,(z-\mathcal{U}(x))^\top\Big]K^\top Q.
\end{aligned}$$

Hence, $\|\mathbb{W}_\mu(D_x\mathcal{V})(z)\| \le 4M^3Re^{2M^2R^2}(1+2M^2R^2)$.

The variation in θ is computed similarly. By the fact that $\int_{\mathbb{R}^d}\alpha(x,y)\,(y-\mathcal{U}(x))\,\mathrm{d}\mu(y) = \mathcal{U}(x) - \mathcal{U}(x) = 0$, a direct computation gives that

$$D_V(D_x\mathcal{V})\cdot[\Delta V] = \Delta V\,\mathcal{M}(x)\,K^\top Q,$$

and

$$\begin{aligned}
D_Q(D_x\mathcal{V})\cdot[\Delta Q] = V\left[\int \alpha(x,y)\,\langle \Delta Q\, x,\, K(y-\mathcal{U}(x))\rangle\,(y-\mathcal{U}(x))\,(y-\mathcal{U}(x))^\top\, d\mu(y)\right]K^\top Q\\
+ V\,\mathcal{M}(x)\,K^\top \Delta Q,
\end{aligned} \tag{E.11}$$

and, similarly,

$$\begin{aligned}
D_K(D_x\mathcal{V})\cdot[\Delta K] = V\left[\int \alpha(x,y)\,\langle Qx,\, \Delta K(y-\mathcal{U}(x))\rangle\,(y-\mathcal{U}(x))\,(y-\mathcal{U}(x))^\top\, d\mu(y)\right]K^\top Q\\
+ V\,\mathcal{M}(x)\,(\Delta K)^\top Q.
\end{aligned}$$

Collecting terms gives $\|D_\theta(D_x\mathcal{V})\| \le 12M^2R^2(1+2M^2R^2)$. □

Proposition E.4 (Smoothness of $\mathbb{W}_\mu\mathcal{V}$)**.** *Adopt* (A1)*. There exists* $L_2 = L_2(M, R)$ *such that* $\mathbb{W}_\mu\mathcal{V}$ *in* (E.1) *satisfies* (A4) *(b) for* $x, z \in B_R(0)$.

Proof. We differentiate $\mathbb{W}_\mu\mathcal{V}(z) = \alpha(x,z)[V(z-\mathcal{U}(x))(K^\top Qx)^\top + V]$ with respect to x, z, μ, and θ.

According to (E.2) and (E.3), a direct computation shows that

$$\begin{aligned}
D_x(\mathbb{W}_\mu\mathcal{V})(z)\cdot[h] = \alpha(x,z)\Big[&\langle Q^\top K(z-\mathcal{U}(x)),\, h\rangle\big(V(z-\mathcal{U}(x))x^\top Q^\top K + V\big)\\
&- V\mathcal{M}(x)K^\top Q h\, x^\top Q^\top K + V(z-\mathcal{U}(x))\,h^\top Q^\top K\Big],
\end{aligned}$$

and

$$D_z(\mathbb{W}_\mu\mathcal{V})(z)\cdot[h]=\alpha(x,z)\Big[x^\top Q^\top Kh\big(V(z-\mathcal{U}(x))x^\top Q^\top K+V\big)+Vh\,x^\top Q^\top K\Big].$$

Thus,

$$\|D_x(\mathbb{W}_\mu\mathcal{V})(z)\|\le e^{2M^2R^2}4M^3R(1+2M^2R^2),$$

and

$$\|D_z(\mathbb{W}_\mu\mathcal{V})(z)\|\le e^{2M^2R^2}2M^3R(1+M^2R^2).$$

For the derivatives with respect to μ, we have that, by (E.2) and (E.10),

$$\frac{\delta(\mathbb{W}_\mu\mathcal{V})(z)}{\delta\mu}(w)=-\alpha(x,z)\,\alpha(x,w)\left[V(z-\mathcal{U}(x))x^\top Q^\top K+V+V(w-\mathcal{U}(x))\,x^\top Q^\top K\right]$$

and

$$\begin{aligned}&\mathbb{W}_\mu((\mathbb{W}_\mu\mathcal{V})(z))(w)\cdot[h]\\&=-\alpha(x,z)\,\alpha(x,w)\Big[\langle Q^\top K(w-\mathcal{U}(x)),h\rangle\big[V(z-\mathcal{U}(x))x^\top Q^\top K+V\\&\quad+V(w-\mathcal{U}(x))\,x^\top Q^\top K\big]+Vh\,x^\top Q^\top K\Big].\end{aligned}$$

Thus, we get

$$\|\mathbb{W}_\mu((\mathbb{W}_\mu\mathcal{V})(z))(w)\|\le e^{4M^2R^2}4M^3R(1+M^2R^2).$$

Finally, we prove the Lipschitz continuity of $\mathbb{W}_\mu\mathcal{V}$ in $\theta=(Q,K,V)$. A direct computation shows that

$$D_V(\mathbb{W}_\mu\mathcal{V})(z)\cdot[\Delta V]=\alpha(x,z)\Big[(\Delta V)(z-\mathcal{U}(x))x^\top Q^\top K+(\Delta V)\Big].$$

By (E.6),

$$\begin{aligned}D_Q(\mathbb{W}_\mu\mathcal{V})(z)\cdot[\Delta Q]=\alpha(x,z)\Big[&\langle\Delta Q\,x,K(z-\mathcal{U}(x))\rangle\big(V(z-\mathcal{U}(x))x^\top Q^\top K+V\big)\\&-V\mathcal{M}(x)K^\top(\Delta Q)\,x\,x^\top Q^\top K+V(z-\mathcal{U}(x))\,x^\top(\Delta Q)^\top K\Big],\end{aligned}$$

and similarly,

$$\begin{aligned}D_K(\mathbb{W}_\mu\mathcal{V})(z)\cdot[\Delta K]=\alpha(x,z)\Big[&\langle Q\,x,\Delta K(z-\mathcal{U}(x))\rangle\big(V(z-\mathcal{U}(x))x^\top Q^\top K+V\big)\\&-V\mathcal{M}(x)(\Delta K)^\top Q\,x\,x^\top Q^\top K+V(z-\mathcal{U}(x))\,x^\top Q^\top(\Delta K)\Big].\end{aligned}$$

Therefore,

$$\begin{aligned}\|D_\theta(\mathbb{W}_\mu\mathcal{V}_\theta)\|&\le 3\max\{\|D_V(\mathbb{W}_\mu\mathcal{V}_\theta)\|,\|D_Q(\mathbb{W}_\mu\mathcal{V}_\theta)\|,\|D_K(\mathbb{W}_\mu\mathcal{V}_\theta)\|\}\\&\le 3e^{2M^2R^2}(1+2M^2R^2)^2.\end{aligned}$$

□

Proposition E.5 (Smoothness of $D_\theta\mathcal{V}$)**.** *Adopt* (A1)*. There exists $L_3=L_3(M,R)$ such that $D_\theta\mathcal{V}$ in* (E.1) *satisfies* (A4) *(c) for $x\in B_R(0)$.*

Proof. By the commutativity of mixed partial derivatives, the bounds on $D_x(D_\theta\mathcal{V})$ and $\mathbb{W}_\mu(D_\theta\mathcal{V})$ follow from Propositions E.3 and E.4. It suffices to bound $D_\theta(D_\theta\mathcal{V})$.

From (E.7), we have that

$$\begin{aligned}&D_K((D_V\mathcal{V})\cdot[\Delta V])\cdot[\Delta K]=\Delta V\,\mathcal{M}(x)\,(\Delta K)^\top Qx,\\&D_Q((D_V\mathcal{V})\cdot[\Delta V])\cdot[\Delta Q]=\Delta V\,\mathcal{M}(x)\,K^\top\Delta Qx,\\&D_V(D_V\mathcal{V})=0.\end{aligned}$$

Thus $\|D_\theta(D_V\mathcal{V})\| \le 12MR^3$.

From (E.8),

$$D_V((D_Q\mathcal{V})\cdot[\Delta Q])\cdot[\Delta V] = \Delta V\,\mathcal{M}(x)\,K^\top \Delta Q x,$$

$$D_Q((D_Q\mathcal{V})\cdot[\Delta Q])\cdot[\Delta Q'] = V\bigg[\int \alpha(x,y)\,\langle \Delta Q'\,x,\,K(y-\mathcal{U}(x))\rangle \cdot (y-\mathcal{U}(x))(y-\mathcal{U}(x))^\top\,d\mu(y)\bigg]K^\top\Delta Q x,$$

$$D_K((D_Q\mathcal{V})\cdot[\Delta Q])\cdot[\Delta K] = V\bigg[\int \alpha(x,y)\,\langle Q\,x,\,\Delta K(y-\mathcal{U}(x))\rangle \cdot (y-\mathcal{U}(x))(y-\mathcal{U}(x))^\top \mathrm{d}\mu(y)\bigg]K^\top\Delta Q x + V\mathcal{M}(x)(\Delta K)^\top\Delta Q x.$$

Hence, $\|D_\theta(D_Q\mathcal{V})\| \le 8MR^3(1+2M^2R^2)$.

Similarly,

$$D_V((D_K\mathcal{V})\cdot[\Delta K])\cdot[\Delta V] = \Delta V\,\mathcal{M}(x)\,(\Delta K)^\top Q x,$$

$$D_K((D_K\mathcal{V})\cdot[\Delta K])\cdot[\Delta K'] = V\bigg[\int \alpha(x,y)\,\langle Q\,x,\,\Delta K'(y-\mathcal{U}(x))\rangle \cdot (y-\mathcal{U}(x))\,(y-\mathcal{U}(x))^\top\,d\mu(y)\bigg](\Delta K)^\top Q x,$$

$$D_Q((D_K\mathcal{V})\cdot[\Delta K])\cdot[\Delta Q] = V\bigg[\int \alpha(x,y)\,\langle \Delta Q\,x,\,K(y-\mathcal{U}(x))\rangle \cdot (y-\mathcal{U}(x))\,(y-\mathcal{U}(x))^\top \mathrm{d}\mu(y)\bigg](\Delta K)^\top Q x + V\mathcal{M}(x)(\Delta K)^\top\Delta Q x.$$

Thus, $\|D_\theta(D_K\mathcal{V})\| \le 8MR^3(1+2M^2R^2)$.

Combining the above estimates, we get that $\|D_\theta(D_\theta\mathcal{V})\| \le 24MR^3(1+2M^2R^2)$. □

Proposition E.6 (Kernel form of $\mathcal{V}$)**.** *Adopt* (A1)*. There exists $L_4 = L_4(M,R)$ such that $\mathcal{V}$ in* (E.1) *satisfies Assumption B for $m = d+1$ and all $x \in B_R(0)$.*

Proof. Let $w = (w_1, w_2, \ldots, w_d, w_{d+1}) \in \mathbb{R}^{d+1}$ with $w_{d+1} \neq 0$, we define

$$\mathcal{F}(w) \coloneqq \frac{(w_1, w_2, \ldots, w_d)}{w_{d+1}},$$

and for $\theta = (Q, K, V)$, define

$$F(x,y;\theta) \coloneqq \Big(e^{\langle Qx, Ky\rangle} V y,\; e^{\langle Qx, Ky\rangle}\Big) \in \mathbb{R}^{d+1}.$$

When $x, y \in B_R(0)$, there is a $C = C(\mathbf{A})$, such that $|F(x,y;\theta)| \le C$ and $e^{\langle Qx,Ky\rangle} > C^{-1} > 0$. Hence, because all the derivatives of $\mathcal{F}(w)$ are smooth, bounded, and dimensional free, when $|w| \le C$ and $w_{d+1} > C^{-1}$, $\mathcal{F}$ satisfies the assumptions in Assumption B with a dimensional free constant L_4. Also, by analogous arguments in verifying Assumptions (A3)–(A4) in this Section E.1, F also satisfies Assumption B with an L_4 only depending on R, M in Assumption (A1). □

E.2 MLP blocks

For an MLP block, the velocity field is

$$\mathcal{V}(x;\theta) \;=\; \mathrm{W}_1\,\sigma(W_2 x + b), \qquad \theta = (W_1, W_2, b) \in \mathbb{R}^{2d^2+d}, \tag{E.12}$$

which depends only on x and θ (not on μ).

We impose the following regularity on the activation.

Assumption E.7 (Activation regularity)**.** The activation $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is twice continuously differentiable, and there exists $\alpha > 0$ such that $|\sigma(x)| \le \alpha(|x|+1)$, $\|D\sigma(x)\| \le \alpha$, and $\|D^2\sigma(x)\| \le \alpha$ for all $x \in \mathbb{R}^d$.

Remark E.8. Assumption E.7 is satisfied by the GELU activation used in modern LLMs. The ReLU activation $\sigma_{\mathrm{ReLU}}(t) = \max\{t, 0\}$ is not C^2, but its piecewise-linear structure permits all results in Section 2 to go through after a separate analysis that avoids the classical ODE well-posedness theory with Lipschitz coefficients. As discussed after Assumption A, the assumption is needed only on the compact sets supplied by Proposition E.9; we state it globally for notational simplicity.

Proposition E.9 (Boundedness of the MLP flow)**.** *Adopt* (A1) *for θ and μ_0 and Assumption E.7. There exists $C = C(M, R, \alpha)$ such that* $\operatorname{supp}\mu_s \subseteq B_C(0)$ *for all $s \in [0,1]$.*

Proof. The proof is similar to Proposition E.1. Setting $R_\tau = \sup\{|x| : x \in \operatorname{supp}\mu_\tau\}$, the linear-growth bound $|\mathcal{V}(x;\theta)| \le \alpha M(M|x| + M + 1)$ together with Grönwall yields the conclusion. □

Proposition E.10 (Lipschitz continuity of the MLP block)**.** *Adopt* (A1) *and Assumption E.7. There exist M_0, L_0 depending on (M, R, α) such that $\mathcal{V}$ in* (E.12) *satisfies* (A3) *for $x \in B_R(0)$.*

Proof. The spatial Jacobian

$$D_x\mathcal{V}(x;\theta) = W_1(D\sigma(W_2x + b))W_2 \tag{E.13}$$

satisfies $\|D_x\mathcal{V}\| \le \alpha M^2$. The directional derivatives in θ are

$$\begin{aligned} D_{W_1}\mathcal{V}\cdot[\Delta W_1] &= (\Delta W_1)\sigma(W_2x + b),\\ D_{W_2}\mathcal{V}\cdot[\Delta W_2] &= W_1(D\sigma(W_2x + b))(\Delta W_2)x,\\ D_b\mathcal{V}\cdot[\Delta b] &= W_1(D\sigma(W_2x + b))(\Delta b), \end{aligned} \tag{E.14}$$

with bounds $\alpha(MR + M + 1)$, $M\alpha R$, $M\alpha$ respectively. □

Proposition E.11 (Smoothness of $D_x\mathcal{V}$ for MLP)**.** *Adopt* (A1) *and Assumption E.7. There exists $L_1 = L_1(M, R, \alpha)$ such that $D_x\mathcal{V}$ in* (E.12) *satisfies* (A4) *(a) for $x \in B_R(0)$.*

Proof. By (E.13), we directly see that

$$\|D_x^2\mathcal{V}\cdot[h]\| \le \alpha M^2|W_2h| \le \alpha M^3|h|.$$

Thus, $\|D_x^2\mathcal{V}\| \le \alpha M^3$.

Also,

$$\|D_{W_1}(D_x\mathcal{V})\cdot[\Delta W_1]\| \le \|\Delta W_1(D\sigma(W_2x + b))W_2\| \le \|\Delta W_1\|\alpha M,$$

and

$$\|D_{W_2}(D_x\mathcal{V})\cdot[\Delta W_2]\| \le \|W_1\|\cdot(\alpha\cdot|\Delta W_2x|)\cdot\|W_2\| + \|W_1\|\alpha\|\Delta W_2\| \le M\alpha(MR+1)\|\Delta W_2\|,$$

and

$$\|D_b(D_x\mathcal{V})\cdot[\Delta b]\| \le \|W_1\|\cdot(\alpha|\Delta b|)\cdot\|W_2\| \le \alpha M^2|\Delta b|.$$

Hence, $\|D_\theta(D_x\mathcal{V})\| \le \alpha M(MR + M + 2)$. □

Proposition E.12 (Smoothness of $D_\theta\mathcal{V}$ for MLP)**.** *Adopt* (A1) *and Assumption E.7. There exists $L_3 = L_3(M, R, \alpha)$ such that $D_\theta\mathcal{V}$ in* (E.12) *satisfies* (A4) *(c) for $x \in B_R(0)$.*

Proof. Recall the terms constituting $D_\theta \mathcal{V}$ in (E.14). A direct computation consequently shows that

$$D_{W_1}(D_{W_1}\mathcal{V}) = 0, \quad D_{W_2}(D_{W_1}\mathcal{V} \cdot [\Delta W_1]) \cdot [\Delta W_2] = \Delta W_1 (D\sigma(W_2 x + b))\Delta W_2 x$$
$$D_b(D_{W_1}\mathcal{V} \cdot [\Delta W_1]) \cdot [\Delta b] = \Delta W_1 (D\sigma(W_2 x + b))\Delta b.$$

Thus, $\|D_\theta D_{W_1}\mathcal{V}\| \le \alpha(R+1)$.

Similarly,

$$\begin{aligned}\|D_{W_2}(D_{W_2}\mathcal{V} \cdot [\Delta W_2]) \cdot [\Delta W_2']\| &\le \|W_1\|\|(D^2\sigma(W_2 x + b))\||\Delta W_2 x||\Delta W_2' x| \\ &\le \alpha M R^2 \|\Delta W_2\|\|\Delta W_2'\|,\end{aligned}$$

and

$$\|D_b(D_{W_2}\mathcal{V} \cdot [\Delta W_2]) \cdot [\Delta b]\| \le \|W_1\|\|(D^2\sigma(W_2 x + b))\||\Delta W_2 x||\Delta b| \le M\alpha R \cdot \|\Delta W_2\||\Delta b|.$$

Thus, $\|D_\theta D_{W_2}\mathcal{V}\| \le \alpha(MR^2 + MR + R)$.

Finally,

$$\|D_b(D_b\mathcal{V} \cdot [\Delta b]) \cdot [\Delta b']\| \le \|W_1\|\|(D^2\sigma(W_2 x + b))\| \cdot |\Delta b||\Delta b'| \le M\alpha \cdot |\Delta b||\Delta b'|.$$

Hence, $\|D_\theta D_b \mathcal{V}\| \le \alpha(M + MR + 1)$.

Combining the above estimates, $\|D_\theta D_\theta \mathcal{V}\| \le C(M, R, \alpha)$. The bounds on $D_x(D_\theta \mathcal{V})$ follow from Proposition E.11 by the commutativity of mixed partial derivatives. This finishes the proof for Proposition E.12. □

Conclusion. Combining Propositions E.2–E.5 for self-attention layers and Propositions E.10–E.12 for MLP layers, every transformer block satisfies Assumptions (A3) and (A4) on the compact sets supplied by (A1) (after Propositions E.1 and E.9), with constants depending only on (M, R) and on the activation regularity α, and in particular independent of the embedding dimension d and the parameter dimension p. All results in Section 2 therefore apply to standard transformer architectures with absolute constants.

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: Yes. The main claims are clearly stated in the abstract and in Section 1 (Introduction), including the key methodological contributions and the problem settings. These claims are directly supported by the theoretical results in Section 3, Section 4, and Section 5.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Some of the proofs used the Grönwall type arguments and hence resulted in exponentially large constants when time increases. We discussed this briefly after Theorem 2.2, in the part *Role of regularization*.

   Guidelines:

   - The answer [N/A] 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: The assumptions are included in Section 5, Assumption A. The proofs are included in Appendix A, Appendix B, Appendix C, and Appendix D. We have short proof sketches after theorems.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not include experiments.

Guidelines:

- The answer [N/A] 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: [N/A]

   Justification: The paper does not include experiments requiring code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] 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: Yes. We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms to it in every respect. The paper presents purely theoretical contributions; it does not involve human subjects, private datasets, or applications with foreseeable negative societal impacts.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: This paper presents theoretical contributions in machine learning, supported by mathematical proofs. There are no direct applications or foreseeable societal impacts (positive or negative) of the work performed.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release any data, models, or code that could pose a high risk for misuse. It consists solely of theoretical analysis.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper presents only theoretical contributions. It does not rely on any external code, datasets, or models.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper is purely theoretical and does not introduce or release any new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper presents only theoretical contributions with mathematical proofs. It does not involve crowdsourcing or any research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper presents only theoretical contributions with mathematical proofs and does not involve any research with human subjects.

    Guidelines:

- The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The mathematical results and proofs in this paper are derived by the authors. It does not involve LLMs as any important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.