Quantitative Propagation of Chaos in Mean-Field NNs: Sufficient Conditions and Lower Bounds

Margalit Glasgow and Joan Bruna

May 22, 2026

Abstract

This manuscript builds upon the work of Glasgow et al. [5], which identifies a set of conditions guaranteeing propagation of chaos in two-layer neural networks for long timescales. First, we identify several novel sufficient conditions for proving long-time propagation of chaos, and demonstrate how these conditions can go beyond the traditional Grönwall approach. Next we provide several counter-examples showing why the approach in Glasgow et al. [5] cannot be used to yield $\mathrm{poly}(t)/\sqrt{m}$ propagation of chaos with significantly weaker assumptions.

This manuscript is a working draft and may be updated.

1 Introduction

We recall several definitions from Glasgow et al. [5]. Let $\mathcal{S}$ be the space in which the neurons lie (typically $\mathbb{R}^d$ or $\mathbb{S}^{d-1}$), and for some distribution $\rho \in \mathcal{P}(\mathcal{S})$ over neurons let

$$f_\rho(x) := \mathbb{E}_{w\sim\rho}\sigma(w^\top x), \tag{1.1}$$

where σ is some activation function. Let $\mathcal{D} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R})$ be the data distribution, and define the square loss

$$L(\rho) := \mathbb{E}_{(x,y)\sim\mathcal{D}}(f_\rho(x) - y)^2. \tag{1.2}$$

We study the evolution of ρ_t under (projected) Wasserstein gradient flow on this loss function, that is, where each particle evolves according to the equation

$$\frac{d}{dt}w = \begin{cases} \nabla_w \nu(w, \rho_t) & \text{standard gradient flow} \\ P_w \nabla_w \nu(w, \rho_t) & \text{projected gradient flow} \end{cases} \tag{1.3}$$

$$\nu(w, \rho_t) := \frac{\delta L(\rho_t)}{\delta \rho_t}(w). \tag{1.4}$$

Here P_w is the projection orthogonal to w in the case of projected gradient flow on $\mathbb{S}^{d-1}$.

We denote the mean-field distribution at time t by $\rho_t^{\mathrm{MF}} \in \mathcal{P}(\mathcal{S})$, with the initialization $\rho_0^{\mathrm{MF}} = \rho_0$. Let $\xi_t(w) \in \mathcal{S}$ denote the *characteristic* of a particle initialized at w and evolved under the mean-field dynamics:

$$\tfrac{d}{dt}\xi_t(w) = \nu(\xi_t(w), \rho_t^{\mathrm{MF}}) \qquad \xi_0(w) = w\ .$$

We denote by $\hat{\rho}_t^m$ the finite m-particle discretization of this flow. We initialize $\hat{\rho}_0^m = \frac{1}{m}\sum_{i=1}^m \delta_{w_i}$, where $w_i \sim \rho_0$ i.i.d. for each $i \in [m]$. Each particle $w \in \mathcal{S}$ in the discretization dynamics evolves according to the

finite-particle velocity field $\nu(w, \hat{\rho}_t^m)$. This defines an ODE in $\mathcal{S}^{\otimes m}$, whose characteristics are now denoted by $\hat{\xi}_t(w_i)$, and solve

$$\frac{d}{dt}\hat{\xi}_t(w_i) = \nu_{\mathcal{D}}(\hat{\xi}_t(w_i), \hat{\rho}_t^m) \qquad \hat{\xi}_0(w_i) = w_i \ , \ i \in [m] \ .$$
$$\hat{\rho}_t^m := \frac{1}{m}\sum_{i=1}^m \delta_{\hat{\xi}_t(w_i)}.$$

To prove quantitative, non-asymptotic weak propagation of chaos (PoC), we must bound with high probability

$$\|f_{\rho_t^{\mathrm{MF}}} - f_{\hat{\rho}_t^m}\|_{\mathcal{D}}^2 := \mathbb{E}_{x\sim\mathcal{D}}(f_{\rho_t^{\mathrm{MF}}}(x) - f_{\hat{\rho}_t^m}(x))^2.$$

More precisely, we can ask for the exact rate of in terms of t, m, and the problem instance $(\rho_0, \sigma, \mathcal{D})$:

Question 1.

Q1.1 *Can we find some $\mathcal{F}(t, \rho_0, \sigma, \mathcal{D})$ such that with high probability (as $m \to \infty$),*

$$\|\rho_t^{\mathrm{MF}} - \hat{\rho}_t^m\|_{\mathcal{D}}^2 \le \frac{\mathcal{F}(t, \rho_0, \sigma, \mathcal{D})}{\tilde{\Omega}(m)} \tag{1.5}$$

Under what conditions on $(\rho_0, \sigma, \mathcal{D})$ is $\mathcal{F}(t, \rho_0, \sigma, \mathcal{D})$ bounded polynomially in t?

Q1.2 *We may also hope to find $\mathcal{F}(t, \rho_0, \sigma, \mathcal{D})$ that is not only sufficient, but also **necessary**. That is, can we show the existence of $(\rho_0, \sigma, \mathcal{D})$ such that for arbitrarily large m and t,*

$$\|\rho_t^{\mathrm{MF}} - \hat{\rho}_t^m\|_{\mathcal{D}}^2 \gtrapprox \frac{\mathcal{F}(t, \rho_0, \sigma, \mathcal{D})}{m}. \tag{1.6}$$

Remark 1. *A priori it is not obvious that the correct rate should be inversely linear in m. In fact, Chen et al. [2] shows that under some strong conditions, asymptotically in t, we can expect a faster rate in m. Conversely, some of uniform-in-time PoC results from [4] yield slower than $1/m$ rates. However, the standard Gronwall argument along with our result in [5] suggest that a $\tilde{\Theta}(1/m)$ rate is most standard.*

Notation. $\mathcal{P}(\Omega)$ denotes the space of probability distributions over Ω. For a vector $w \in \mathbb{R}^d$, we let $\|w\|$ denote its 2-norm, and for a matrix $M \in \mathbb{R}^d \times \mathbb{R}^d$, we let $\|M\|$ denote its operator norm.

We will use lower-case letters (f, g, h) to denote functions defined on $\mathbb{R}^d$. For $f \in L^2(\mathbb{R}^d, \mathcal{D})$, we write $\|f\|_{\mathcal{D}}^2 := \mathbb{E}_x|f(x)|^2$, and omit the subscript when the context is clear.

We use Greek letters (Δ, ξ, etc) to denote vector-valued functions $\mathcal{S} \to \mathbb{R}^d$, and upper-case letters to denote matrix-valued functions $\mathcal{S} \to \mathbb{R}^{d\times d}$ or $\mathcal{S} \times \mathbb{S}^{d-1} \to \mathbb{R}^{d\times d}$. When $\hat{\mu}$ is an empirical measure of the form $\hat{\mu} = \frac{1}{m}\sum_i \delta_{w_i}$, we will use the shorthand $\Lambda(i) = \Lambda(w_i)$, and denote $\mathbb{E}_i\Lambda(i) := \frac{1}{m}\sum_i \Lambda(w_i)$.

For $H \in L^2(\mathcal{S}\times\mathcal{S}, \mu^2, \mathbb{R}^{d\times d})$, $D \in L^2(\mathcal{S}, \mu, \mathbb{R}^{d\times d})$ and $\Lambda, \in L^2(\mathcal{S}, \mu, \mathbb{R}^d)$, we use $H\Lambda(w) := \mathbb{E}_{w'\sim\mu}H(w, w')\Lambda(w')$, and $(D\Lambda)(w) := D \odot \Lambda(w) = D(w)\Lambda(w)$. We let $\|\Lambda\|_{\mu,p} := (\mathbb{E}_{w\sim\mu}\|\Lambda(w)\|^p)^{1/p}$, with the default that $p = 2$ if p is omitted. Further, we let $\|D\|_\mu$ and $\|H\|_\mu$ denote the operator norms $\|D\|_\mu := \sup_{w\in\mathcal{S}}\|D(w)\|$, and $\|H\|_\mu := \sup_{\|\Lambda\|\le 1}\|H\Lambda\|$. In all these norms, we omit the subscript μ when it is clear from context. Occasionally, we will write $\|V\|_{p\to q}$ to denote $\sup_{\|X\|_p\le 1}\|VX\|_q$, where the measures used in these norms will be obvious from context.

Throughout this paper, we use the asymptotic notation $O_C(X)$ to denote X times some constant that depends arbitrarily on C. Whenever a term of the form C (usually with some subscript) appears, this term is referring to a constant, meaning that its value does not depend on m or d (which we may take to infinity).

We write "with high probability" when the probability approaches 1 as m goes to infinity. This probability is always taken over the neural network initialization $\{w_i\}_{i\in[m]}$.

1.1 Coupling Approach.

A standard approach in establishing PoC is to couple $\hat{\rho}_t^m$ with an auxiliary empirical measure $\bar{\rho}_t^m$. Define $\bar{\rho}_t^m$ to be the distribution initialized at $\hat{\rho}_0^m$, but that evolves according to the dynamics $\nu(\cdot, \rho_t^{\mathrm{MF}})$. That is, $\bar{\rho}_t^m = \frac{1}{m}\sum_{i=1}^m \delta_{\xi_t(w_i)}$. Note that $\bar{\rho}_t^m$ is equivalent in distribution to a random sample of m particles drawn iid from ρ_t^{MF}.

Define the coupling error at neuron w_i by

$$\Delta_t(i) := \hat{\xi}_t(w_i) - \xi_t(w_i) \in \mathbb{R}^d,\ i \in [m]\,, \tag{1.7}$$

Viewing Δ_t as an element of $L^2([m], \mathrm{unif}([m]), \mathbb{R}^d)$, Lemma 5 of Glasgow et al. [5] shows that with high probability over $\hat{\rho}_0^m$, under mild regularity conditions, it is possible to track the dynamics of Δ_t by the following ODE:

$$\frac{d}{dt}\Delta_t = D_t \odot \Delta_t - H_t\Delta_t + \boldsymbol{\epsilon}_t, \tag{$\star$}$$

Here D_t and H_t are respectively the diagonal and matrix-valued kernel operators

$$D_t(i) := \nabla_{\xi_t(w_i)}\nu(\xi_t(w_i), \rho_t^{\mathrm{MF}}) \in \mathbb{R}^{d\times d} \tag{1.8}$$

$$H_t(i,j) := K'(\xi_t(w_i), \xi_t(w_j)) \in \mathbb{R}^{d\times d} \tag{1.9}$$

$$K'(w,w') := \nabla_w\nabla_{w'}\mathbb{E}_{x\sim\mathcal{D}}\sigma(x^\top w)\sigma(x^\top w') \tag{1.10}$$

We call D_t the local hessian, and H_t the interaction hessian. Note crucially that K' is a PSD kernel and thus H_t is PSD. Further, $\boldsymbol{\epsilon}_t$ is a small error term on the scale $\frac{\log(m)}{\sqrt{m}} + \|\Delta_t\|_4^2$; the exact scale of $\boldsymbol{\epsilon}_t$ is governed by certain constant C_{reg} governing regularity conditions on σ and $\mathcal{D}$.

Define $\Psi_{t,s}$ to be the solution map of ($\star$) without the source term $\boldsymbol{\epsilon}_t$, that is $\Psi_{s,s} = I$, and $\frac{d}{dt}\Psi_{t,s} = (D_t - H_t)\Psi_{t,s}$. Note that controlling $\Psi_{t,s}$ suffices to characterize the growth of Δ_t, since Duhamel's principle gives that

$$\Delta_t = \Psi_{t,0}\Delta_0 + \int_{s=0}^t \Psi_{t,s}\boldsymbol{\epsilon}_s ds.$$

In this way, its is straightforward to see that quantitative control of $\|\Psi_{t,s}\|$ can give quantitative propagation of chaos guarantees. In particular, with high probability, for all $t \le \frac{m^{1/4}}{\sup_{r\le s\le t}\|\Psi_{s,r}\|}$, we have

$$\|f_{\rho_t^{\mathrm{MF}}} - f_{\hat{\rho}_t^m}\|_{\mathcal{D}} \le \frac{\log(m)}{m} + O_{C_{\mathrm{reg}}}\big(\|\Delta_t\|^2\big) \le \sup_{s\le t}\|\Psi_{t,s}\| O_{C_{\mathrm{reg}}}\left(\frac{t^2\log^2(m)}{m}\right) \tag{1.11}$$

Here the first inequality is from Lemma 1 in [5], while the second inequality holds by an induction argument on the size of $\boldsymbol{\epsilon}_t$.

Remark 2. *While bounding $\|\Psi_{t,0}\|$ is a* sufficient *condition to bound $\|\rho_t^{\mathrm{MF}} - \hat{\rho}_t^m\|_{\mathcal{D}}$, it is a priori possible that the $\boldsymbol{\epsilon}_s$ are highly structured and that such a bound is not required. However, we would find this very surprising.*

In light of the connections between $\sup_s \|\Psi_{t,s}\|$ and the the PoC rate $\|f_{\rho_t^{\mathrm{MF}}} - f_{\hat{\rho}_t^m}\|$, it is natural to investigate Question 1 through the lens of conditions on (D_t, H_t). To phrase this more cleanly, we will state this question for more general L^2 spaces:

Question 2. *Let $\mathcal{H} = L^2(\mathcal{X}, \rho, B)$. Let $H_t : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(B)$ and $D_t : \mathcal{X} \to \mathcal{L}(B)$ be kernels satisfying*

$$\|H(x,y)\|_{\mathrm{op}} \le 1, \qquad \|D(x)\|_{\mathrm{op}} \le 1 \quad \textit{for all } s,t \in \mathcal{X}.$$

They induce bounded operators $H_t, D_t : \mathcal{H} \to \mathcal{H}$ defined by

$$(H_t\Delta)(x) := \mathbb{E}_{t\sim\rho}\big[H(x,y)\Delta(y)\big], \qquad (D\Delta)(x) := D(x)\Delta(x), \quad \Delta \in \mathcal{H}.$$

Suppose H_t is PSD. Let $\Psi_{t,s}$ be the solution map of $\frac{d}{dt}\Delta_t = (D_t - H_t)\Delta_t$, ie. $\Delta_t = \Psi_{t,s}\Delta_s$.

Q2.1 *What additional conditions on (D_t, H_t) suffice to show that the solution map is polynomially bounded, that is, $\sup_{s\le t}\|\Psi_{t,s}\|_{2\to 2} \le \mathrm{poly}(t)$?*

Q2.2 *Can we show that such conditions are necessary to bound $\sup_{s\le t}\|\Psi_{t,s}\|_{2\to 2}$?*

Note that an answer to **Q2.2** does *not* immediately imply an answer to **Q1.2**, since not all systems (D_t, H_t) can be instantiated via instances of gradient flow on 2-layer neural networks. Nevertheless, we believe understanding the tightest conditions on (D_t, H_t) which can bound $\|\Psi_{t,s}\|$ will be useful in approaching Question 1.

1.2 A set of sufficient conditions from [5]

The work [5] proved that under a certain set of assumptions on D_t and H_t, we can bound $\|f_{\rho_t^{\mathrm{MF}}} - f_{\hat{\rho}_t^m}\|_{\mathcal{D}}$. We give a gist of these assumptions below, and describe them in slightly more generality.

We define $J_{t,s}$ to be the solution map of $\frac{d}{dt}\Delta_t = D_t\Delta_t$, that is, $\frac{d}{dt}J_{t,s} = D_tJ_{t,s}$ with $J_{s,s} = I$. Note that $J_{t,s}$ is a diagonal operator.

1. **Worst-Case stability**. We have a polynomial bound on $\|J_{t,s}\|_\infty$, that is,

$$J_{\max}(t,s) := \sup_{s\le q\le r\le t}\|J_{r,q}\|_\infty \le \mathrm{poly}(t).$$

   We will sometimes abbreviate $J_{\max}(t) := J_{\max}(t,0)$.

2. **Local Strong Convexity**: For each i, $D_t(i) \preceq -C_{\mathrm{LSC}}(t)$ eventually. [5] in particular requires that $C_{\mathrm{LSC}}(t) = C_{\mathrm{LSC}}\sqrt{L(\rho_t^{\mathrm{MF}})}$, for some constant C_{LSC}, and that this condition hold as soon as $\|\xi^\infty(w_i) - \xi_t(w_i)\|$ is sufficiently small, where $\xi^\infty(w_i) = \lim_{t\to\infty}\xi_t(w_i)$. A related (though incomparable) local strong convexity condition (used eg. in [2, 3]) requires that $\lim_{t\to\infty}\max_i D_t(i) \preceq -C_{\mathrm{LSC}}$.[1]

3. **Incoherence / Average Stability.** This assumption considers the interaction of $J_{t,s}$ and H_s, and assumes a bound on $J_{t,s}H_s$. Glasgow et al. [5] in particular assumed the following variant:

$$\|\mathbf{1}_{B_\tau^t}\cdot J_{t,s}\cdot H_s\Delta\|_1 \le \|\Delta\|_1\frac{\log(1/\tau)}{t} \qquad B_\tau^t = \{i : \|\xi_t(w_i) - \xi^\infty(w_i)\| \le \tau\} \tag{1.12}$$

   More generally, we will use the terminology *incoherence* assumptions to refer to a bound on $J_{t,s}H_s$, its derivative $D_tJ_{t,s}H_s$, or the operator norm of $J_{t,s}$ or $D_tJ_{t,s}$, restricted to the span of H_s.

[1] See Theorem 3.7 in [2] where the analog of D_t is $\nabla\nabla V$. See Assumption A5 in [3], where the analog of D_t is the object H.

Under the above conditions, and various additional regularity and symmetry conditions, Theorem 1 [5] showed that for any $\delta > 0$, one could achieve the weak PoC bound

$$\|f_{\hat{\rho}_t^m} - f_{\rho_t^{\mathrm{MF}}}\|^2 \leq \frac{O_{C_{\mathrm{reg}},\delta}(1)\,\mathrm{poly}(d) t^2 J_{\max}(t)^2}{m}, \tag{1.13}$$

for $t \leq \{\inf_s : L(\rho_s^{\mathrm{MF}}) \leq \delta\}$. Glasgow et al. [5] also showed that these conditions were met in the setting of learning well-specified single-index models (SIM) with a high information exponent, for which the training time grows polynomially in the dimension d. Yet this work leaves open many questions:

Q3.1 The local strong convexity (LSC) implies that distribution $\rho_\infty^{\mathrm{MF}}$ is atomic, namely, the neurons all converge to a discrete set in $\mathcal{S}$. It is possible to remove the local strong convexity, to cover settings such as misspecified single or multi-index models?

Q3.2 Without explicit strongly convex regularization, the LSC assumption also precludes the PoC result from holding uniformly in time, because the dependence on the loss δ is exponential. Can we attain uniform-in-time PoC that hold for $L(\rho_t^{\mathrm{MF}}) \to 0$?

Q3.3 The incoherence assumption holds in the SIM setting because the neurons are dispersed at initialization, meaning H_t (which captures the neuron-to-neuron interactions) is small. In problems with saddle-to-saddle structure, the neurons may cluster near the saddles, leading to worse incoherence. Is such an incoherence assumption necessary? Does it suffice just to bound $J_{\max}(t)$?

1.3 Contributions and Outline.

This manuscript gives several results which give progress on the questions mentioned in the introduction. The subsequent work of the authors in [4] addresses **Q3.1**, yielding uniform-in-time guarantees whenever the convergence rate is like t^{-2} or faster.

Sufficient Conditions: Question Q2.1.

- Section 2 approaches this problem by looking at the Volterra formulation

$$\Delta_t = J_{t,0}\Delta_0 - \int_{s=0}^{t} J_{t,s} H_s \Delta_s ds, \tag{1.14}$$

which is attained from ($\star$) via Duhamel's principle. This formulation allows us to leverage the commutativity of the D_t to avoid the Grönwall-type bound: indeed, we have $\|J_{t,s}\| = \exp\left(\lambda_{max}\left(\int_{r=s}^t D_r dr\right)\right) \leq \exp\left(\int_{r=s}^t \lambda_{max}(D_r) dr\right)$, where the inequality may be exponentially loose. In Proposition 2, we build upon existing techniques in the Volterra equation literature to attain novel bounds for vector-valued Volterra equations. In Corollaries 4 and 5, we use this proposition to bound $\Psi_{t,s}$ using various "incoherence" assumptions.

The results in this section are particularly useful in the case that H_t is a low-rank matrix, which indeed arises in the case of neural networks when the problem has many symmetries (eg. rotational symmetries).[2] When H_t is low-rank, we show that the projection of Δ_t onto this low-dimensional space can be captured by a low-dimensional Volterra equation. Thus, we lose the simple ODE structure of ($\star$), but simplify the problem by reducing the dimensionality. We give a simple low-rank example showing how our results can improve exponentially over the Grönwall bound $\|\Psi_{t,s}\| \leq \exp\left(\int_{r=s}^t \lambda_{max}(D_r) dr\right)$.

[2]For example, in the SIM example studied in [5], H_t has a diagonal block-structure where each block is constant-rank. Thus the system can be reduced to studying each of these blocks.

- Section 3 looks at the question from a different perspective, by building instead on the Lyapunov equation to obtain a variational characterization of $\|\Psi_{t,s}\|$ that optimizes over instantaneous PSD metrics. Explicit bounds are then obtained by relaxing the associated variational principle to smaller spaces, such as diagonal metrics, leading to explicit gains relative to the naive logarithmic norm control (and Gronwall).

Lower Bounds: Question Q2.2. In Section 4, we show a class of counterexamples where where $J_{\max}(t)$ is uniformly bounded, but $\|\Psi_{t,0}\|$ grows exponentially, or stretched-exponentially, in t. In these examples, $H_t \equiv H$ is rank-one matrix, but the incoherence is significantly weaker than the assumption in [5], or in the sufficient conditions in our present results. These counterexamples also shed light on **Q3.3** above by showing that the $J_{\max}(t)$ condition alone does *not* suffice, even with extremely simple H_t.

2 Sufficient Conditions I: Volterra Approach

2.1 New Bounds for Vector-Valued Volterra Equations.

In this section, we consider the vector-valued Volterra equation

$$X_t = a_t - \int_{s=0}^{t} C(t,s) X_s ds, \tag{2.1}$$

where $C(t,t) \succeq 0$. The following classical bound from Burton [1] gives a sufficient condition to bound $\|X_t\|$:

Proposition 1 (Burton [1]). *Consider a Volterra equation $X_t = a_t - \int_{s=0}^{t} C(t,s)X_s ds$, and suppose either:*

$$C(t,t) \succeq \int_{s=0}^{t} \left\| \frac{d}{dt} C(t,s) \right\| ds + \int_{u=t}^{\infty} \left\| \frac{d}{du} C(u,t) \right\| du, \tag{2.2}$$

or

$$C(t,t) \succeq \int_{s=0}^{t} \left\| \frac{d}{ds} C(t,s) \right\| ds + \int_{u=t}^{\infty} \left\| \frac{d}{dt} C(u,t) \right\| du. \tag{2.3}$$

Then

$$\|X_t\| \leq \|a_0\| + \int_{s=0}^{t} \|\dot{a}_s\| ds. \tag{2.4}$$

The proof of this proposition uses the potential function $V_t(X) = \|X_t\|^2 + \int_{s=0}^{t} \|X_s\|^2 \int_{u=t}^{\infty} \|\frac{d}{du} C(u,s)\| du ds$, or, for the second condition, reparameterize the Volterra equations in terms of $Y_t := \int_{s=0}^{t} X_s ds$.

The limitation of this proposition, however, is that the condition is hard to meet when $C(t,t)$ has small eigenvalues. Indeed, in our Volterra formulation in (1.14), we have $C(t,s) = J_{t,s}H_s$, and $C(t,t) = H_t$. Thus, the above proposition cannot take advantage of the fact that $C(t,s)$ may be small in the same eigenspaces that $C(t,t)$ is small.

We prove the following propostion which resolves this limitation and strengthens Proposition 1. The Lyapunov function we use to prove this Proposition is inspired by the one from [1].

Proposition 2. *In the setting of* (2.1)*, let* $B_t := C(t,t)$*, and suppose that* $\|B_t\| \le 1$*, and*

$$\alpha_L(t) \ge \sup_{\|X\|\le 1} \int_{s=0}^{t} \left\| X^\top B_t^{-1/2} C_t(t,s) B_s^{-1/2} \right\| ds \tag{2.5}$$

$$\alpha_R(t) \ge \sup_{\|X\|\le 1} \int_{u=t}^{\infty} \left\| B_u^{-1/2} C_u(u,t) B_t^{-1/2} X \right\| du \tag{2.6}$$

$$\beta_t := (\alpha_L(t) + \alpha_R(t) - 2)_+ \tag{2.7}$$

Then

$$\|X_t\| \le \exp\left(\int_{r=0}^{t} \beta_r dr\right) \|a_0\| + \int_{s=0}^{t} \exp\left(\int_{r=s}^{t} \beta_r dr\right) \|\dot{a}_s\| ds. \tag{2.8}$$

Proof. Use the potential function

$$V_t(X) = \|X_t\|^2 + \int_{s=0}^{t} \|X_s\|_{B_s} \int_{u=t}^{\infty} \|B_u^{-1/2} C_u(u,s) X_s\| du ds. \tag{2.9}$$

We have

$$\frac{d}{dt} V_t(X) = 2X_t^\top \left(-C(t,t)X_t - \int_{s=0}^{t} C_t(t,s) X_s ds + \dot{a}_t \right) \tag{2.10}$$

$$+ \|X_t\|_{B_t} \int_{u=t}^{\infty} \left\| B_u^{-1/2} C_u(u,t) X_t \right\| du - \int_{s=0}^{t} \|X_s\|_{B_s} \left\| B_t^{-1/2} C_t(t,s) X_s \right\| ds. \tag{2.11}$$

Now using Claim 3 below with $Y = B_t^{1/2} X_t$, $A_s = B_t^{-1/2} C_t(t,s) B_s^{-1/2}$, and $Y_s = B_s^{1/2} X_s$, we have

$$|2X_t^\top \int_{s=0}^{t} C_t(t,s) X_s ds| \le \|X_t\|_{B_t} \int_{s=0}^{t} \|X_t^\top B_t^\top B_t^{-1/2} C_t(t,s) B_s^{-1/2}\| ds + \int_{s=0}^{t} \|X_s\|_{B_s} \left\| B_t^{-1/2} C_t(t,s) X_s \right\| ds \tag{2.12}$$

$$\le \alpha_L(t) \|X_t\|_{B_t}^2 + \int_{s=0}^{t} \|X_s\|_{B_s} \left\| B_t^{-1/2} C_t(t,s) X_s \right\| ds. \tag{2.13}$$

Thus

$$\frac{d}{dt} V_t(X) \le -2\|X_t\|_{B_t}^2 + 2X_t^\top \dot{a}_t + \alpha_L(t) \|X_t\|_{B_t}^2 + \|X_t\|_{B_t} \int_{u=t}^{\infty} \left\| B_u^{-1/2} C_u(u,t) X_t \right\| du \tag{2.14}$$

$$\le -2\|X_t\|_{B_t}^2 + 2X_t^\top \dot{a}_t + \alpha_L(t) \|X_t\|_{B_t}^2 + \alpha_R(t) \|X_t\|_{B_t}^2 \tag{2.15}$$

$$\le (\alpha_L(t) + \alpha_R(t) - 2)_+ V_t(X) + 2X_t^\top \dot{a}_t, \tag{2.16}$$

where the last line holds by assumption of the proposition.

It follows that

$$\frac{d}{dt} \sqrt{V_t(X)} \le \frac{1}{2} \beta_t \sqrt{V_t(X)} + \|\dot{a}_t\|. \tag{2.17}$$

Thus using Duhamel, we have that

$$\sqrt{V_t(X)} \le \exp\left(\frac{1}{2} \int_{r=s}^{t} \beta_r dr\right) \sqrt{V_0(X)} + \int_{s=0}^{t} \exp\left(\frac{1}{2} \int_{r=s}^{t} \beta_r dr\right) \|\dot{a}_s\| ds. \tag{2.18}$$

Since $V_0(X) = \|X_0\|^2$, and $V_t(X) \ge \|X_t\|^2$, this yields the conclusion.

Claim 3. *For any vector Y, vectors Y_s, and matrices A_s, and interval I, we have*

$$|2\int_I Y^\top A_s Y_s ds| \le \|Y\| \int_I \|A_s^\top Y\| ds + \int_I \|Y_s\| \|A_s Y_s\| ds. \tag{2.19}$$

Proof. We have $Y^\top A_s Y_s \le \|Y\| \|A_s Y_s\|$, and $Y^\top A_s Y_s \le \|A_s^\top Y\| \|Y_s\|$, so

$$Y^\top A_s Y_s \le \sqrt{\|Y\| \|A_s^\top Y\|} \sqrt{\|A_s Y_s\| \|Y_s\|}. \tag{2.20}$$

Using Cauchy Shwartz and then AM-GM, we have

$$2|\int_I Y^\top A_s Y_s ds| \le 2\sqrt{\int_I \|Y\| \|A_s^\top Y\| ds} \sqrt{\int_I \|A_s Y_s\| \|Y_s\| ds} \tag{2.21}$$

$$\le \int_I \|Y\| \|A_s^\top Y\| ds + \int_I \|A_s Y_s\| \|Y_s\| ds, \tag{2.22}$$

which proves the claim. □

□

2.2 Implications for Bounding $\|\Psi_{t,s}\|$.

We prove the following corollaries.

Corollary 4. *In setting of Question 2, we have each of the following.*

1. *If for all $t \in [S,T]$,*

$$\int_{s=S}^t \left\| H_t^{-1/2} D_t J_{t,s} H_s^{1/2} \right\| ds + \int_{u=t}^T \left\| H_u^{-1/2} D_u J_{u,t} H_t^{1/2} \right\| du \le 2, \tag{2.23}$$

   then for all $S \le s \le t \le T$, we have $\|\Psi_{t,s}\| \le 1 + \sup_{\|X\|\le 1} \int_{r=s}^t \|D_r J_{r,s} X\| dr$.

2. *If for all $t \in [S,T]$, we have*

$$\int_{s=S}^t \|D_t J_{t,s}\| ds + \int_{u=t}^T \|D_u J_{u,t}\| du \tag{2.24}$$

$$+ \int_{s=S}^t \left\| H_t^{1/2} (\dot{H_t^{1/2}}) J_{t,s} \right\| ds + \int_{u=t}^T \left\| H_u^{-1/2} (\dot{H_u^{1/2}}) J_{u,t} \right\| du \le 2, \tag{2.25}$$

   then for all $S \le s \le t \le T$, we have $\|\Psi_{t,s}\| \le 3 + 9(t-s) + 9(t-s) \int_{r=s}^t \|\dot{H}_r^{1/2}\| dr$. Note that if all the H_t commute, the condition above reduces to

$$\int_{s=S}^t \|D_t J_{t,s}\| ds + \int_{u=t}^T \|D_u J_{u,t}\| du + \int_{s=S}^t \frac{1}{2} \left\| \dot{H}_t J_{t,s} \right\| ds + \int_{u=t}^T \frac{1}{2} \left\| \dot{H}_u J_{u,t} \right\| du \le 2. \tag{2.26}$$

Proof. Recall that in the setting of Question 2, we have

$$\Delta_t = J_{t,s} \Delta_s - \int_{r=s}^t J_{t,r} H_r \Delta_r dr. \tag{2.27}$$

For item (1), we apply Proposition 2 directly to (2.27) with $C(t,s) := J_{t,s}H_s$, such that $C(t,t) = H_t$, and $C_t(t,s) = D_t J_{t,s} H_s$. Here we have $X_t = \Delta_t$, and $a_t = J_{t,s}\Delta_s$, so $\dot{a}_r = D_r J_{r,s}\Delta_s$, and thus by Proposition 2,

$$\|\Delta_t\| \le \|\Delta_s\| + \int_{r=s}^{t} \|D_r J_{r,s}\Delta_s\| dr \tag{2.28}$$

$$\le \|\Delta_s\| \left(1 + \sup_{\|X\|\le 1} \int_{r=s}^{t} \|D_r J_{r,s} X\| dr\right). \tag{2.29}$$

For item (2), we let $X_t := H_t^{1/2}\Delta_t$ and apply Proposition 2 to the system

$$X_t = H_t^{1/2} J_{t,s}\Delta_s - \int_{r=s}^{t} H_t^{1/2} J_{t,r} H_r^{1/2} X_r dr, \tag{2.30}$$

with $C(t,s) := H_t^{1/2} J_{t,s} H_s^{1/2}$, such that $C(t,t) = H_t$, and $C_t(t,s) = H_t^{1/2} D_t J_{t,s} H_s^{1/2} + \left(\dot{H_t^{1/2}}\right) J_{t,s} H_s^{1/2}$. Here $a_t = H_t^{1/2} J_{t,s}\Delta_s$, so $\dot{a}_r = H_r^{1/2} D_r J_{r,s}\Delta_s + \dot{H}_r^{1/2} J_{r,s}\Delta_s$. Thus by Proposition 2,

$$\|H_t^{1/2}\Delta_t\| \le \|H_s^{1/2}\Delta_s\| + \int_{r=s}^{t} \left(\|H_r^{1/2} D_r J_{r,s}\Delta_s\| + \|\dot{H}_r^{1/2} J_{r,s}\Delta_s\|\right) dr \tag{2.31}$$

$$\le \|\Delta_s\| \left(1 + \int_{r=s}^{t} \|D_r J_{r,s}\| dr + \int_{r=s}^{t} \|\dot{H}_r^{1/2} J_{r,s}\| dr\right) \tag{2.32}$$

$$\le \|\Delta_s\| \left(3 + \int_{r=s}^{t} \|\dot{H}_r^{1/2} J_{r,s}\| dr\right). \tag{2.33}$$

Here the second line follows because $\|H_r\| \le 1$ by definition in Question 2, and the third line follows by the condition in this item (2).

It follows, by also applying the bound above to any $r \in [s,t]$ that

$$\|\Delta_t\| \le \|J_{t,s}\Delta_s\| + \int_{r=s}^{t} \|J_{t,r} H_r \Delta_r\| dr \tag{2.34}$$

$$\le \|J_{t,s}\Delta_s\| + \int_{r=s}^{t} \|J_{t,r}\| \|\Delta_s\| \left(3 + \int_{q=s}^{r} \|\dot{H}_q^{1/2} J_{q,s}\| dq\right) dr \tag{2.35}$$

$$\le \|\Delta_s\| \left(3 + 9(t-s) + 9(t-s) \int_{r=s}^{t} \|\dot{H}_r^{1/2}\| dr\right) \tag{2.36}$$

Here in the last line we used that for any $r \in [s,t]$, we have $\|J_{t,r}\| = \|J_{r,r} + \int_{q=r}^{t} D_q J_{q,r} dq\| \le 1 + \int_{q=r}^{t} \|D_q J_{q,r}\| dq \le 3$.

□

Sometimes we need to massage things even more to use. The following corollary allows us to exploit cases where H_t is low rank, that is, we have some decomposition $H_t = U_t \Sigma_t U_t^\top$, where $U_t U_t^\top = I_k$ for some k independent of t.

Corollary 5. *In the setting of Question 2, we have the following two results.*

1. *Suppose for all $t \in [S,T]$, we have the decomposition $H_t = U\Sigma_t U^\top$, where $U^\top U = I$, and*

$$\int_{s=S}^{t} \left\|\Sigma_t^{-1/2} U^\top D_t J_{t,s} U \Sigma_s^{1/2}\right\| ds + \int_{u=t}^{T} \left\|\Sigma_u^{-1/2} U^\top D_u J_{u,t} U \Sigma_t^{1/2}\right\| du \le 2. \tag{2.37}$$

Then for any $S \le s \le t \le T$, we have

$$\|\Psi_{t,s}\| \le J_{max}(t,s)\left(1 + (t-s)\left(1 + \sup_{\|X\|\le 1}\int_{r=s}^{t} \|U^\top D_r J_{r,s} X\| dr\right)\right). \tag{2.38}$$

2. *Suppose for all $t \in [S,T]$, we have the decomposition $H_t = U_t \Sigma_t U_t^\top$, where $U_t^\top U_t = I$, and*

$$\int_{s=S}^{t} \left\| U_t^\top D_t J_{t,s} U_s \right\| ds + \int_{u=t}^{T} \left\| U_u^\top D_u J_{u,t} U_t \right\| du \tag{2.39}$$

$$+ \int_{s=S}^{t} \left\| \Sigma_t^{-1/2}\left(\Sigma_t^{1/2\dot{}} U_t^\top\right) J_{t,s} U_s \right\| ds + \int_{u=t}^{T} \left\| \Sigma_u^{-1/2}\left(\Sigma_t^{1/2\dot{}} U_t^\top\right) J_{u,t} U_t \right\| du \le 2. \tag{2.40}$$

Then for any $S \le s \le t \le T$, we have

$$\|\Psi_{t,s}\| \le J_{max}(t,s)\left(1 + (t-s)\left(1 + \sup_{\|X\|\le 1}\int_{r=s}^{t} \|U^\top D_r J_{r,s} X\| dr + J_{max}(t,s)\int_{r=s}^{t} \|\left(\Sigma_r^{1/2\dot{}} U_r^\top\right)\| dr\right)\right). \tag{2.41}$$

Further, if all the H_t commute, and thus we have the decomposition $H_t = U \Sigma_t U^\top$, where $U^\top U = I$, and $\dot{\Sigma}_t$ commutes with Σ_t, the above result holds whenever we have the condition

$$\int_{s=S}^{t} \left\| U^\top D_t J_{t,s} U \right\| ds + \int_{u=t}^{T} \left\| U^\top D_u J_{u,t} U \right\| du \tag{2.42}$$

$$+ \int_{s=S}^{t} \frac{1}{2}\left\| \dot{\Sigma}_t U^\top J_{t,s} U \right\| ds + \int_{u=t}^{T} \frac{1}{2}\left\| \dot{\Sigma}_u U^\top J_{u,t} U \right\| du \le 2. \tag{2.43}$$

Proof. For item (1) we let $X_t := U^\top \Delta_t$ and apply Proposition 2 to the system

$$X_t = U^\top J_{t,s} \Delta_s - \int_{r=s}^{t} U^\top J_{t,r} U \Sigma_r X_r dr, \tag{2.44}$$

with $C(t,s) := U^\top J_{t,s} U \Sigma_s$, such that $C(t,t) = \Sigma_t$, and $C_t(t,s) = U^\top D_t J_{t,s} U \Sigma_s$. Here $a_r = U^\top J_{r,s} \Delta_s$ such that $\dot{a}_r = U^\top D_r J_{r,s} \Delta_s$. Then Proposition 2 yields for any $r \ge S$:

$$\|U^\top \Delta_r\| \le \|U^\top \Delta_s\| + \int_{q=s}^{r} \|U^\top D_q J_{q,s} \Delta_s\| dq \tag{2.45}$$

$$\le \|\Delta_s\|\left(1 + \sup_{\|X\|\le 1}\int_{q=s}^{r} \|U^\top D_q J_{q,s} X\| dq\right). \tag{2.46}$$

Now

$$\|\Delta_t\| \le \|J_{t,s}\Delta_s\| + \int_{r=s}^{t} \|J_{t,r} U \Sigma_r U^\top \Delta_r\| dr \tag{2.47}$$

$$\le \|J_{t,s}\Delta_s\| + \int_{r=s}^{t} \|J_{t,r}\| \|U^\top \Delta_r\| dr \tag{2.48}$$

$$\le \|J_{t,s}\| \|\Delta_s\| + \int_{r=s}^{t} \|J_{t,r}\| \left(1 + \sup_{\|X\|\le 1}\int_{q=s}^{r} \|U^\top D_q J_{q,s} X\| dq\right) \|\Delta_s\| dr \tag{2.49}$$

$$\le \|\Delta_s\|\left(\|J_{t,s}\| + \int_{r=s}^{t} \|J_{t,r}\| dr \left(1 + \sup_{\|X\|\le 1}\int_{r=s}^{t} \|U^\top D_r J_{r,s} X\| dr\right)\right) \tag{2.50}$$

$$\le \|\Delta_s\| J_{\max}(t,s)\left(1 + (t-s)\left(1 + \sup_{\|X\|\le 1}\int_{r=s}^{t} \|U^\top D_r J_{r,s} X\| dr\right)\right). \tag{2.51}$$

For item (2), we let $X_t := \Sigma_t^{1/2} U_t^\top \Delta_t$ and apply Proposition 2 to the system

$$X_t = \Sigma_t^{1/2} U_t^\top J_{t,s} \Delta_s - \int_{r=s}^t \Sigma_t^{1/2} U_t^\top J_{t,r} U_r \Sigma_r^{1/2} X_r dr, \tag{2.52}$$

with $C(t,s) := \Sigma_t^{1/2} U_t^\top J_{t,s} U_s \Sigma_s^{1/2}$, such that $C(t,t) = \Sigma_t$, and $C_t(t,s) = \Sigma_t^{1/2} U_t^\top D_t J_{t,s} U_s \Sigma_s^{1/2} + \left(\dot{\Sigma_t^{1/2}} U_t^\top\right) J_{t,s} U_s \Sigma_s^{1/2}$. Here $a_r = \Sigma_r^{1/2} U_r^\top J_{r,s} \Delta_s$, so $\dot{a}_r = \Sigma_r^{1/2} U_r^\top D_r J_{r,s} \Delta_s + \left(\dot{\Sigma_r^{1/2}} U_r^\top\right) J_{r,s} \Delta_s$. Thus Proposition 2 yields that for all $r \in [S,T]$,

$$\|\Sigma_r^{1/2} U_r^\top \Delta_r\| \le \|\Sigma_s^{1/2} U_s^\top \Delta_s\| + \int_{q=s}^r \left(\|\Sigma_q^{1/2} U_q^\top D_q J_{q,s} \Delta_s\| + \|\left(\dot{\Sigma_q^{1/2}} U_q^\top\right) J_{q,s} \Delta_s\| \right) dq \tag{2.53}$$

$$\le \|\Delta_s\| + \|\Delta_s\| \sup_{\|X\| \le 1} \int_{q=s}^r \left(\|U_q^\top D_q J_{q,s} X\| + J_{\max}(t,s) \|\left(\dot{\Sigma_q^{1/2}} U_q^\top\right)\| \right) dq \tag{2.54}$$

$$\tag{2.55}$$

Now

$$\|\Delta_t\| \le \|J_{t,s}\Delta_s\| + \int_{r=s}^t \|J_{t,r} U_r \Sigma_r U_r^\top \Delta_r\| dr \tag{2.56}$$

$$\le \|J_{t,s}\Delta_s\| + \int_{r=s}^t \|J_{t,r}\| \|\Sigma_r^{1/2} U_r^\top \Delta_r\| dr \tag{2.57}$$

$$\le \|J_{t,s}\| \|\Delta_s\| + \int_{r=s}^t \|J_{t,r}\| \left(1 + \sup_{\|X\| \le 1} \int_{q=s}^r \|U_q^\top D_q J_{q,s} X\| dq + J_{\max}(t,s) \int_{q=s}^r \|\left(\dot{\Sigma_q^{1/2}} U_q^\top\right)\| dq \right) \|\Delta_s\| dr \tag{2.58}$$

$$\le \|\Delta_s\| J_{\max}(t,s) \left(1 + (t-s) \left(1 + \sup_{\|X\| \le 1} \int_{r=s}^t \|U^\top D_r J_{r,s} X\| dr + J_{\max}(t,s) \int_{r=s}^t \|\left(\dot{\Sigma_r^{1/2}} U_r^\top\right)\| dr \right) \right). \tag{2.59}$$

□

2.3 Simple example where Corollary 5 improves exponentially over the Grönwall Bound

Suppose $X_t \in \mathbb{R}^d$, and let $D_t(i) := \mathbf{1}(t \in [i, i+1])$, and $H_t \equiv \mathbf{1}\mathbf{1}^\top / d$, such that $\|H_t\| = 1$. Then the Grönwall bound gives

$$\|\Psi_{T,0}\| \le \exp\left(\int_{s=0}^T \|D_s\| ds \right) = \exp(\min(d, T)). \tag{2.60}$$

Now let us apply Corollary 5(2), with $U_t = \mathbf{1}/\sqrt{d}$, $\Sigma_t = 1$. The condition in the corollary requires that for any $t \le T$, we have

$$\int_{s=0}^t \left\| U^\top D_t J_{t,s} U \right\| ds + \int_{u=t}^T \left\| U^\top D_u J_{u,t} U \right\| du \le 2. \tag{2.61}$$

Now $D_t J_{t,s}(i) = \mathbf{1}(t \in [i, i+1]) \exp(t - \max(s,i))$, so $|U^\top D_t J_{t,s} U| \le \frac{\exp(t - \max(s, \lfloor t \rfloor))}{d}$. It follows that

$$\int_{s=0}^t \left\| U^\top D_t J_{t,s} U \right\| ds \le \frac{1}{d} \int_{s=0}^t \exp(t - \max(s, \lfloor t \rfloor)) ds \le \frac{t}{d} \exp(t - \lfloor t \rfloor) ds \le \frac{te}{d} \tag{2.62}$$

Similarly,

$$\int_{u=t}^{T} \left\| U^\top D_u J_{u,t} U \right\| du \le \frac{1}{d} \int_{u=t}^{T} \exp(u - \max(t, \lfloor u \rfloor)) du \le \frac{(T-t)e}{d}. \tag{2.63}$$

Thus for $T \le 2d/e$, the condition holds, and the corollary yields the bound

$$\|\Psi_{T,0}\| \le J_{\max}(T) \left(1 + T \left(1 + \sup_{\|X\| \le 1} \int_{r=0}^{T} \|U^\top D_r J_{r,s} X\| dr \right) \right) \tag{2.64}$$

$$\le e(1+3T). \tag{2.65}$$

Note that we can apply this bound twice yielding

$$\|\Psi_{d,0}\| \le \|\Psi_{d,2d/e}\| \|\Psi_{2d/e,0}\| \le e^2(1+6d/e)^2 \le 37d^2. \tag{2.66}$$

3 Lyapunov Variational Formula

We describe in this section an alternative formulation that complements the previous control of $\|\Psi_{t,s}\|$. This time we exploit the ODE structure rather than the Volterra integral form.

3.1 Exact Variational Principle

Using a Lyapunov transformation, we can characterize the operator norm $\|\Psi_{t,s}\|$ as follows:

Proposition 6 (Variational Characterization of Propagator Norm)**.** *We have*

$$\|\Psi_{t,s}\| = \inf \left\{ \lambda_{\max}(M_s); M_t = I, M_u \succeq 0; \dot{M}_u + (D_u - H_u) M_u + M_u (D_u - H_u) \preceq 0 \, \forall \, u \in (s,t) \right\}. \tag{3.1}$$

Proof. Let us first show that LHS $\le$ RHS. We write $A_u = D_u - H_u$. Consider any admissible time-varying metric $(M_u)_u$, and suppose x_u solves the ODE $\dot{x}_u = A_u x_u$, so $x_t = \Psi_{t,s} x_s$. Let $l_u := x_u^\top M_u x_u$. Its time derivative is

$$\dot{l}_u = x_u^\top \dot{M}_u x_u + x_u^\top A_u M_u x_u + x_u^\top M_u A_u x_u \tag{3.2}$$

$$= x_u^\top \left(\dot{M}_u + A_u M_u + M_u A_u \right) x_u \tag{3.3}$$

$$\le 0 \, , \tag{3.4}$$

thus

$$\|x_t\|^2 = l_t \le l_s \le \lambda_{\max}(M_s) \|x_s\|^2 \, . \tag{3.5}$$

To show that the inequality is tight, consider the metric M_u^* given by the propagator $M_u^* = \Psi_{t,u}^\top \Psi_{t,u}$. Since $\Psi_{t,u}$ solves $\partial_u \Psi_{t,u} = -\Psi_{t,u} A_u$, we have

$$\dot{M}_u^* = -A_u \Psi_{t,u}^\top \Psi_{t,u} - \Psi_{t,u}^\top \Psi_{t,u} A_u \, , \tag{3.6}$$

and $M_t^* = I$, so M_u^* is admissible, and trivially we have $\|\Psi_{t,u}\| = \lambda_{\max}(M_s^*)$. □

3.2 Diagonal Relaxation

While the previous variational principle gives the exact propagator norm, it is not immediately obvious how to leverage it in our setting. A first natural starting point is to consider a relaxation using diagonal metrics of the form $M_u = \exp(-2R_u)$ with $R_u = \mathrm{diag}(r_1(u), \dots, r_d(u))$. For convenience, we can reparameterize the previous bound as follows. We set the boundary condition at $R(s) = 0$, and define

$$\gamma_R(u) := \lambda_{\max}(D_u - \dot{R}_u - H_u \odot C_u) \,, \tag{3.7}$$

with $(C_u)_{ij} = \cosh(r_u(i) - r_u(j))$.

The previous variational bound now reads

$$\|\Psi_{t,s}\| \le \exp\left(\inf_{R(s)=0} \mathcal{C}_{t,s}(R)\right) \,, \text{ with} \tag{3.8}$$

$$\mathcal{C}_{t,s}(R) = \max_i r_i(t) + \int_s^t \gamma_R(u) du \,. \tag{3.9}$$

By picking $R \equiv 0$ we obtain the naive Gronwall bound. In the other extreme, when H is diagonal, then $\mathcal{G}_R(u) = H_u$ and thus if we choose R such that $\dot{R} = D_u - H_u$, we recover the exact propagator that fully exploits commutation.

3.3 Delocalization

For each $u \in [s,t]$ we decompose $H_u = H_u^{st} + H_u^{wk}$, where the 'strong' component is spanned by the first P eigenfunctions: $H_u^{st} = V_u^\top \Lambda_u V_u$ and $H_u^{wk} = W_u^\top \tilde{\Lambda} W_u$. We assume that the dominant subspace is delocalised: $\max_i (V_u^\top V_u)_{ii} \le P\mu/d$ with $\mu = \Theta(1)$ and assume that $\|H_u^{wk}\| \le \delta$ for all u.

For any vector $r \in \mathbb{R}^d$, define $\bar{r} = \frac{1}{d}\sum_i r_i$ and $\tilde{r}_i = r_i - \bar{r}$, as well as $\Delta = \max_i r_i - \min_i r_i$, and finally

$$\kappa_s(r) := \sinh^2(\Delta/2) \,, \tag{3.10}$$

$$\kappa_a(r) := \frac{1}{d}\sum_i \sinh^2(\tilde{r}_i) \,. \tag{3.11}$$

For any $H \succeq 0$ and $r \in \mathbb{R}^d$, if we set $M_{ij} = H_{ij}\cosh(r_i - r_j)$, we first verify that

$$\lambda_{\min}(M) \ge \lambda_{\min}(H)\cosh^2(\Delta/2) - \lambda_{\max}(H)\sinh^2(\Delta/2) \,. \tag{3.12}$$

By exploiting the delocalization in H^{st}, we verify that

$$\lambda_{\min}(M) \ge -\delta\kappa_s(r) - \|H\|\min(P\mu\kappa_a(r), \kappa_s(r)) \tag{3.13}$$

Therefore, we have the surrogate upper bound

$$\log\|\Psi_{t,s}\| \le \inf_{R(s)=0}\left\{\max_i r_i(t) + \int_s^t \lambda_{\max}[D_u - \dot{R}_u]du + P\mu\sup_u\|H_u\|\int_s^t \kappa_a(R_u)du + \delta\int_s^t \kappa_s(R_u)du\right\} \tag{3.14}$$

The feature of this is that under this delocalization hypothesis, this objective no longer depends on H, and it becomes a filtering problem of the trajectories $D_u(i)$, $i = 1 \dots d$.

Sanity Check: Counter-example Consider the system with $H = \alpha u u^\top$, where $u = (1/\sqrt{d}, \dots, 1/\sqrt{d})$, and $D_s(i) = \mathbf{1}(\lfloor s/\log d \rfloor = i)$ for $s \in [0, d\log d := t]$. In this case our previous surrogate cost becomes

$$\inf_{R(0)=0} \left\{ \max_i r_i(t) + \int_0^t \max_i (D_u(i) - \dot{R}_u(i)) du + \int_0^t \kappa_a(R_u) du \right\}.$$

The true operator norm $\log \|\Psi\|$ is of order $\log d + O(\alpha d \log d)$.

By choosing the gauge $\dot{r}_k(u) = 1$ in the active interval $u \in I_k$ and $\dot{r}_k(u) = -1/(d-1)$ during the others, we obtain a propagator bound of the form $\tilde{C}_R = (1 + \alpha d^2/4)\log d$, so it becomes effective when $\alpha = O(1/d^2)$ instead of the optimal $O(1/d)$ rate.

Grownall gets $\log \|\Psi\| \le d \log d$ irrespective of α, and the diagonal gauge $\dot{R} = D$ gets also $O(\alpha d^2 \log d)$.

Modifying the counter-example Suppose now that we preserve the property $0 \le D_s(i) \le 1$ and $\int_0^t D_s(i) ds = \log d$, but break the active set into several intervals rather than a single one. In other words, we shrink the interval $(0, t)$ into $(0, t/L)$ and then replicate it L times. In this case, the first two terms satisfy $\max_i r_i(t) = 0$ and $\int \max_i (d_i(u) - \dot{r}_i(u)) du = O(\log d)$, and the regularization term is now $\kappa_a(r) \simeq d^{1/L}$, so

$$\alpha \int \kappa_a(r_u) du \simeq \alpha d^{1+L^{-1}} \log d$$

As a result, we can now accommodate $\alpha \gtrsim d^{-1-L^{-1}}$.

Another potential modification: Set $D_s(i) = \mathbf{1}(\lfloor s/\log d \rfloor = i) - \mathbf{1}(\lfloor s/\log d \rfloor = i+1)$ for $s \in [0, d\log d := t]$. In that case gronwall is also of order $d \log d$, and choosing the diagonal gauge leads to $\alpha d \log d$, which reaches the threshold $\alpha = O(1/d)$.

This gives a potential blueprint for the sufficient condition: we want the 'active curvature regions' $\Omega_i = \{s; D_i(s) > 0\}$ to be fragmented into short intervals, so that when we integrate the gauges $\dot{r}_i(u) = D_i(s)\mathbf{1}(s \in I_i) - \beta$, the running sums are such that $\kappa_a(r_u) = O(\log d/T)$

4 Counterexamples

A natural question to ask is whether the assumption $J_{\max}(t) \le \mathrm{poly}(t)$ is sufficient to bound $\|\Psi_{t,s}\|$. The following proposition shows that even if $J_{\max}(t) \le e$ for all t, it is possible to have $\|\Psi_{t,s}\|$ grow exponentially in t. This is clear by plugging in $a = 0$ in the proposition.

By choosing larger values of a, we see that even with $J_{\max}(t) \le e$ *and* an additional incoherence assumption, we can still have $\|\Psi_{t,s}\|$ grow exponentially in $\mathrm{poly}(t)$.

Proposition 7. *For any $0 \le a < 1$, there exists a system (D_t, H_t) where for all $t \ge 1$:*

1. *The D_t are diagonal and $\|D_t\| \le 1$,*
2. *The H_t all commute, and $\|H_t\| \le 1$,*
3. *$J_{max}(t) \le e$.*
4. *$\|J_{t,s} H_s\| \le \frac{e}{\sqrt{2}} s^{-a}$, and $\|\frac{d}{dt} J_{t,s} H_s\| = \|D_t J_{t,s} H_s\| \le \frac{e}{2}(ts)^{-a}$.*

and there exists an unbounded sequence of (s,t) with $1 \le s \le t$ where

$$\|\Psi_{t,s}\| \ge 0.5 \exp\big(0.05(t^{1-a})\big). \tag{4.1}$$

Note that this example can be trivially modified to start at $t = 0$.

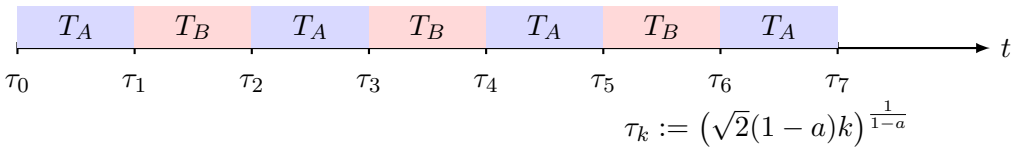

Figure 1: The time axis is partitioned into alternating buckets $[\tau_k, \tau_{k+1})$, with $T_A = \bigcup_{k \text{ even}} [\tau_k, \tau_{k+1})$ and $T_B = \bigcup_{k \text{ odd}} [\tau_k, \tau_{k+1})$. On each bucket, the weighted length $\frac{1}{\sqrt{2}} \int_{\tau_k}^{\tau_{k+1}} t^{-a}\, dt$ is exactly 1, so the contribution of D_t over successive T_A and T_B buckets cancels.

Proof. Consider the following construction, with $D_t, H_t \in \mathbb{R}^{2\times 2}$. Let

$$H_t = \frac{t^{-a}}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{4.2}$$

$$D_t = \frac{t^{-a}}{\sqrt{2}} \begin{cases} D & t \in T_A \\ -D & t \in T_B, \end{cases} \tag{4.3}$$

where $D = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, and

$$t_A = \cup_{k\in\mathbb{N}} \left[\left(\sqrt{2}(1-a)k\right)^{\frac{1}{1-a}}, \left(\sqrt{2}(1-a)(k+1)\right)^{\frac{1}{1-a}} \right) \tag{4.4}$$

$$t_B = \cup_{k\in\mathbb{N}} \left[\left(\sqrt{2}(1-a)(k+1)\right)^{\frac{1}{1-a}}, \left(\sqrt{2}(1-a)(k+2)\right)^{\frac{1}{1-a}} \right). \tag{4.5}$$

Now it is immediate to verify the conditions (1) and (2) for $t \geq 1$.

Let $\tau_k = k^{1+a}$. For (3), since the D_t commute, we have

$$J_{\tau_{k+1},\tau_k} = \exp\left(\frac{1}{\sqrt{2}} D(-1)^k \int_{t=\tau_k}^{q^{k+1}} t^{-a} dt \right) \tag{4.6}$$

$$= \exp\left(\frac{1}{\sqrt{2}} D(-1)^k \frac{1}{1-a} \left(\tau_{k+1}^{1-a} - \tau_k^{1-a}\right) \right) \tag{4.7}$$

$$= \exp\left(D(-1)^k \right) \tag{4.8}$$

$$\tag{4.9}$$

Thus, it is easy to verify that the contribution of a k odd interval and a k even interval exactly cancel, yielding $J_{\max}(t) \leq e$. The bounds in (4) immediately following by bounding $J_{t,s} H_s \leq J_{\max}(t) \|H_s\|$, and $D_t J_{t,s} H_s \leq \|D_t\| J_{\max}(t) \|H_s\|$.

Now we need to show that the propagator $\Psi_{t,s}$ of the full system $D_t - H_t$ is large. We will consider only $\Psi_{\tau_\ell,\tau_{\ell_0}}$ for even ℓ, and ℓ_0 the smallest even number exceeding $\frac{1}{\sqrt{2}(1-a)}$. Note that $\tau_{\ell_0} \geq 1$.

Note that the propagator $\Psi_{t,s}$ satisfies the semigroup (or cocycle) property

$$\Psi_{t,r}\Psi_{r,s} = \Psi_{t,s} \qquad \text{for all } s \leq r \leq t.$$

Thus we have

$$\Psi_{\tau_\ell,\tau_{\ell_0}} = \prod_{k=\ell_0}^{\ell-1} \Psi_{\tau_{k+1},\tau_k} \tag{4.10}$$

where the product is ordered from right to left. Thus it suffices to understand the Ψ_{τ_{k+1},τ_k}. Now reusing the integral calculation, and observing that the $D_t - H_t$ commute within this interval, we have

$$\Psi_{\tau_{k+1},\tau_k} = \exp\Big(M_{(-1)^k}\Big), \tag{4.11}$$

$$\tag{4.12}$$

where

$$M_0 = \begin{pmatrix} 0 & -1 \\ -1 & -2 \end{pmatrix}, \qquad M_{-1} = \begin{pmatrix} -2 & -1 \\ -1 & 0 \end{pmatrix}. \tag{4.13}$$

Thus for even ℓ, we have

$$\overline{M}^{\ell/2}, \tag{4.14}$$

where

$$\overline{M} = \exp(M_0)\exp(M_{-1}) = \begin{pmatrix} 0.344... & -0.625... \\ -0.625... & 0.344... \end{pmatrix}. \tag{4.15}$$

We can compute that $\overline{M}$ has a top eigenvalue λ of value at least 1.26. It follows that

$$\|\overline{M}^{(\ell-\ell_0)/2}\| \geq \lambda^{(\ell-\ell_0)/2} \geq \lambda^{\frac{1}{2\sqrt{2}(1-a)}(t^{1-a}-2)} \geq \exp\big(0.05(t^{1-a}-2)\big) \geq 0.5\exp(0.05t^{1-a}). \tag{4.16}$$

□

References

[1] Theodore Allen Burton. *Liapunov Functionals for Integral Equations*. Trafford Publishing Bloomington, IN, USA, 2008.

[2] Zhengdao Chen, Grant Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. *Advances in Neural Information Processing Systems*, 33:22217–22230, 2020.

[3] Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 2022.

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

[5] Margalit Glasgow, Denny Wu, and Joan Bruna. Mean-field analysis of polynomial-width two-layer neural network beyond finite time horizon. *Proceedings of Machine Learning Research vol*, 1:79, 2025.