Provable State Estimation with Recurrent Models

Vincent Andrieu[1], Pauline Bernard[2], Lucas Brivadis[3], Laurent Praly[2] and Christian Wolf[4]

[1]LAGEPP, CNRS, UCBL 1 [2]CAS, Mines Paris - PSL
[3]L2S, CNRS [4]Naver Labs Europe

Abstract

Estimating the state of a dynamical system from observations is a crucial task in robotics and control. Despite their empirical success, recurrent models often lack formal guarantees on convergence and robustness. We bridge this gap by analyzing these architectures from the perspective of nonlinear observer theory. By framing RNNs as a time-discretization of continuous latent dynamical systems (Neural ODEs), we demonstrate that enforcing a contraction condition on the latent flow guarantees the existence of a smooth, injective map of the true physical state into the latent space. We prove that this architecture acts as a *tunable* observer: by increasing a single scalar gain that scales the latent vector field, the estimation error can be driven to an arbitrarily small neighborhood of zero at an exponentially fast rate. We further establish quantitative bounds on the MLP complexity required to decode this latent state, and extend our formal guarantees to the discrete-time residual RNN implementation. We validate the theory on a real mobile robotic task, where a provably contracting model outperforms a vanilla GRU with substantially more stable training. We also report that unconstrained GRUs converge to contracting dynamics on this task, suggesting that contraction is an emergent property of well-trained sequence models.

Keywords: State estimation, RNNs, Contracting systems, Nonlinear KKL observers.

1 Introduction

We consider the classical problem of estimating the state of a dynamical system, given some online measurements of the observations $o(t)$. This is motivated by applications in robotics control, system monitoring, fault diagnosis, and data assimilation and has been addressed independently in different fields, related to *Observer Design* [6] in control theory and to *State Representation Learning* [29] in ML and embodied AI. In many methods, the quality of the latent state representation is determined by the downstream application building on it, and in this case, it is often learned together ("*end-to-end*") with the downstream task with RL [24] or imitation learning [14]; alternatively through pre-training, i.e. through next token prediction [43], with contrastive learning [47, 49] or recently with joint embedding predictive architectures [3].

In this work, we suppose the existence of a target variable $s(t)$ of interest linked to the state of the system (e.g. the pose of a robot or a hidden quantity of a biological reactor) and are interested in its estimation. A data-driven model for this task is required to learn, both, a latent state representation and a mapping from the latent state to the variable of interest. Natural questions then arise: *Under which conditions does this neural architecture guarantee that the estimate $\hat{s}(t)$ asymptotically converges to the true target $s(t)$? Is it possible to establish formal guarantees regarding the performance and tunability of a trained model?*

To answer these questions, we suppose that the relation between the observations and the target variable is generated by a deterministic dynamical system. More precisely, we assume the existence

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

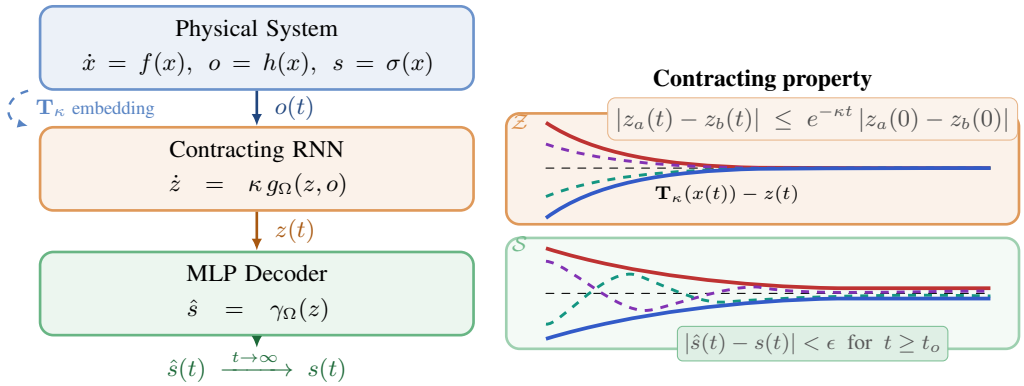

Figure 1: **Provable state estimation:** a contracting Neural ODE with latent dynamics $\dot{z} = \kappa g_\Omega(z, o)$ processes observations $o(t)$ from the physical system with unobserved $x(t)$, of which elements $s(t)$ are of interest. A decoder γ_Ω maps the latent state to an estimate $\hat{s}(t)$. The existence of an injective state-to-latent map $\mathbf{T}_\kappa : \mathcal{X} \to \mathcal{Z}$ is guaranteed (Theorems 3.4-3.7). A contraction property ensures that two latent trajectories z_a, z_b initialized anywhere in $\mathcal{Z}$ exponentially forgets initial conditions. The estimate $\hat{s}(t)$ reaches an ϵ-neighborhood of the true $s(t)$ after a finite observation time, and increasing κ accelerates convergence at the cost of larger network capacity (Theorem 4.2).

of a physical[1] state $x(t) \in \mathbb{R}^n$, governed by an unknown ODE, that determines both $o(t)$ and $s(t)$:

$$\frac{dx}{dt}(t) = f(x(t)), \quad o(t) = h(x(t)), \quad s(t) = \varsigma(x(t)), \tag{1}$$

where f, h, and ς are continuous functions. Moreover, we assume this state is somehow minimal, in the sense that it can be dynamically reconstructed from the knowledge of o – it is *observable* in the control theory sense [6]. Our method thus relies on reconstructing the full physical state itself, looking for an estimate $\hat{x}(t) \approx x(t)$, from which an estimate $\hat{s}(t)$ of $s(t)$ can then be recovered. Without loss of generality, we thus set $s(t) = x(t)$ in this paper[2]. When the dependence on the initial condition is required, we denote the unique solution at time t starting from x_0 as $x(t) = \Phi_f(x_0, t)$. This point of view formulates the machine learning problem as an observer design problem.

From RNNs to Neural ODEs — *Recurrent Neural Networks* (RNNs) and their variants (LSTM [20], GRU [10]) have been shown to be powerful architectures for solving machine learning problems on sequential data. They process sequential observations $o_k \in \mathbb{R}^p$ according to the following recurrence relation:

$$z_{k+1} = \varphi_\Omega(z_k, o_{k+1}), \tag{2}$$

where $z_k \in \mathbb{R}^m$ denotes the hidden (or latent) state of the network, φ_Ω denotes the non-linear update map, and Ω is a vector of trainable weights. The output of the network, which gives an estimate of a target variable s_k, is usually computed by a readout map, often a Multi-Layer Perceptron (MLP):

$$\hat{s}_k = \gamma_\Omega(z_k). \tag{3}$$

Reformulated with residual connection, relation (2) can be rewritten as:

$$z_{k+1} = z_k + \delta \cdot g_\Omega(z_k, o_{k+1}), \tag{4}$$

where $\delta > 0$ is a step-size parameter. This formulation highlights the incremental nature of the state update. When the sampling period of the observations tends to zero ($\delta \to 0$), this recurrence converges to a continuous-time differential equation. This point of view naturally leads to *Neural ODEs*[3] [9], where the latent state and the output evolve continuously in time:

$$\frac{dz}{dt}(t) = g_\Omega(z(t), o(t)), \quad \hat{s}(t) = \gamma_\Omega(z(t)). \tag{5}$$

[1] We call this state *physical* as it is usually obtained through the application of physics laws (e.g. motion laws giving x as positions and velocities, electromagnetic laws giving x as fluxes and currents, chemical laws giving x as concentrations, etc.), as opposed to the *latent* state (later introduced in this paper) that is purely abstract. Note, though, that the knowledge of f, h, σ is not necessary in this paper, so those maps could also have an abstract nature. What is important is that s and o are linked via a state that has a kind of *markovian* property, namely, its instantaneous evolution only depends on its current value.

[2] Reconstructing only $\varsigma(x)$ when x itself is not observable is the object of *functional observability* [11, 15, 36], also addressed with KKL observers [48]; we leave it for future work (see Section 7).

[3] In the original work [9] the ODE was not necessarily applied to problems involving dynamical systems but rather built an internal dynamical system running over the layers of a ResNet-like neural network.

Analyzing the continuous-time limit allows one to use classical tools from dynamical systems and control theory to study stability and convergence properties. As we will show, these properties remain true for practical discrete implementations if the sampling step is small enough.

Contributions — While classical observer theory usually needs explicit knowledge of the functions f and h, this paper develops a theoretical framework to certify the representational capabilities of data-driven RNN-MLP estimators. We prove that, under mild assumptions, there exists a choice of the trainable weights Ω that solves the estimation task with guaranteed exponential convergence.

To obtain these guarantees, we explicitly isolate a tunable scalar gain parameter $\kappa > 0$ in the latent dynamics that acts as a global time-scale separation gain. More precisely, we write the estimator as

$$\frac{dz}{dt}(t) = \kappa g_\Omega(z(t), o(t)), \quad \hat{s}(t) = \gamma_\Omega(z(t)) \tag{6}$$

where $g_\Omega : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^m$ represents the normalized base latent dynamics parameterized by Ω. The following contributions are obtained in this paper.

1. **Existence of the State-to-Latent Map:** We establish a structural result linking deep learning and the Kazantzis-Kravaris-Luenberger (KKL) theory [2]. We prove that if (i) the latent base dynamics generates a contraction, (ii) the latent gain is large enough, (iii) the physical state is differentially observable from o, (iv) the latent dimension is large enough[4], then there exists a smooth injective map $x \mapsto \boldsymbol{T}_\kappa(x)$ which gives a representation of the physical state in the latent space. Contrary to previous existence results with nonlinear latent dynamics [40], the latent vector field may be coupled across latent coordinates, as in GRUs and LSTMs, rather than diagonal; the price is a rank condition on the latent dynamics, which is generic (Appendix K).
2. **Convergence and Tunability:** In an ideal setting where γ_Ω is a left-inverse of $\boldsymbol{T}_\kappa$, we prove that the estimation error provided by the estimator (6) converges exponentially fast to zero whatever the initial condition of the latent dynamics. Moreover, the architecture is formally *tunable*, which means the error can be made arbitrarily small in finite time.
3. **MLP Complexity Bounds:** When the decoder $\boldsymbol{T}_\kappa^{-1}$ is approximated by a neural network γ_Ω, we show that the tunability property is preserved and give a quantitative link between the latent gain κ and the required MLP complexity. We show that while a high gain gives a faster convergence, it may increase the Lipschitz constant of the mapping, which requires a larger network.
4. **Guarantees for Discrete Implementation:** We explicitly extend our continuous-time analysis to the practical case of discrete-time RNNs. We prove that the tunability property is preserved and we give a formal condition linking the required gain κ and the maximum sampling step δ.
5. **Experimental Results**: We showcase the effect of the contracting property on a state estimation problem in robotics with real data and provide evidence for the tunability result in simulation.

2 Related Works

Model-based observers. When a model (f, h) is available, state estimation relies on analytical methods like Kalman, LMI-based or high-gain observers [6]. Among them, the Kazantzis-Kravaris-Luenberger (KKL) observer [26, 2] immerses the physical state dynamics into a stable, observation-driven *linear* filter, through a map solving a partial differential equation (PDE) that is usually analytically intractable. Observer design is also closely related to *synchronization*: an observer is a system that synchronizes with the plant it is driven by [38], and the existence of an attracting map from the plant state to the observer state, as in Theorem 3.4 below, is known as *generalized synchronization*. *Contraction* is an important approach to obtain an observer (see for instance [6, 40]) or a synchronizing system (see for instance [1]).

Data-driven observers and learned dynamical models. Learning dynamical systems from offline datasets is a task inherited from *System Identification* [28]; learned models are then coupled with classical algorithms for analysis [57], simulation [42, 4, 23], state estimation [50, 13], prediction [22] or control [4, 56, 59]. In the KKL line of work, machine learning approximates the immersion map, or rather its inverse, from data (supervised approaches on simulated datasets, autoencoder losses embedding the PDE, or a mix of both [44, 41, 22, 37, 8, 34, 53]), but always with *linear* latent dynamics. *Reservoir computing* [32] and reservoir observers [31] go one step further by driving a

[4] An explicit bound is given depending on the dimension of the physical state x and the observability property from o.

fixed random nonlinear recurrent network with the observations and training a readout: the *echo state property* [21] is essentially a contraction assumption, but neither the injectivity of the induced map nor the tunability of the estimate come with guarantees. The works [12, 33] suggest that contraction-like properties can be enforced during training and enable the design of control laws and observers. The only existence result for KKL observers with nonlinear contracting latent dynamics [40] assumes a *diagonal* latent vector field, each coordinate being an independent scalar contraction driven by the observation. This structure is the pivot of the proof (coordinate-wise regularity, Vandermonde rank argument), and no recurrent architecture in practical use has this form.

Recurrent models like *LSTM* [20] and *GRU* [10] precisely mix nonlinearities across latent coordinates, so that their latent Jacobian is dense. Their memory capacity scaling problem has been addressed by external neural memory [16, 51, 25] before transformers even arrived. Recurrence gained traction again, e.g. through state space models like *S4* [18], *Mamba* [17] and *LRU* [39]. Attention has been combined with recurrence, e.g. *xLSTM* [5], and full-fledged Transformers handling a memory bank are proposed by *MooG* [55], *TTM* [45] and *Kinaema* [46]. We lift the diagonality restriction of [40] and provide existence, injectivity and tunability guarantees for this class under a contraction condition and a generic rank condition, and derive an architecture satisfying the former by design.

3 The Latent Embedding Architecture

To ensure the mathematical rigor, we introduce the following regularity and invariance assumptions.

Assumption 3.1 (Blanket Assumption)**.** *Throughout this paper, we assume the functions f and h in (1) are defined on the entire system state space $\mathbb{R}^n$, are continuously differentiable as many times as needed and the physical state trajectories $x(t)$ obtained from (1) remain in positive times in a compact set $\mathcal{X} \subset \mathbb{R}^n$. We assume also the function g_Ω defining the latent dynamics (6) is as many times continuously differentiable as needed.*

Based on this assumption, we introduce a compact and convex inflation of the set $\mathcal{X}$, denoted $\breve{\mathcal{X}}$, such that $\mathcal{X}$ is contained within the topological interior of $\breve{\mathcal{X}}$. We then define the corresponding extended compact observation set $\mathcal{O} = h(\breve{\mathcal{X}}) \subset \mathbb{R}^p$. This allows us to establish a safe operating region for the latent state. We define the compact set $\mathcal{Z} \subset \mathbb{R}^m$ as a closed ball $\mathcal{Z} = \{z \in \mathbb{R}^m : |z| \leq R_z\}$, where the radius $R_z > 0$ is chosen large enough (specifically $R_z \geq 2 \max_{o \in \mathcal{O}} |g_\Omega(0, o)|$).

3.1 Fading Memory via Contracting Latent Flows

To ensure the RNN acts as a fading memory filter that exponentially forgets its arbitrary initial condition z_0, we restrict its latent vector field to generate contracting flows.

Assumption 3.2 (Monotonicity of the Latent Dynamics)**.** *The base latent vector field g_Ω is uniformly monotonic on $\mathcal{Z}$ for any observation in $\mathcal{O}$. More precisely, its Jacobian matrix satisfies:*

$$\frac{\partial g_\Omega}{\partial z}(z, o) + \left(\frac{\partial g_\Omega}{\partial z}(z, o)\right)^\top \leq -2I_m, \quad \forall (z, o) \in \mathcal{Z} \times \mathcal{O}. \tag{7}$$

For instance, a residual RNN cell with a linear dissipative term, $\dot{z}_i = \kappa(-\alpha_i z_i + \lambda_i \tanh(z_i + U_i o))$, globally satisfies (7) if $\alpha_i - |\lambda_i| \geq 1$. It also covers the case where there exist $P \in \mathbb{R}^{m \times m}$ positive definite and $a > 0$ such that $P \frac{\partial g_\Omega}{\partial z}(z, o) + \left(\frac{\partial g_\Omega}{\partial z}(z, o)\right)^\top P \leq -aP$ for all (z, o), modulo a linear change of coordinates of z and a scaling on κ. This is exploited in Appendix H.2 to propose a novel contracting GRU and handle multi-layered structures. No constraint is put on the *structure* of g_Ω: a stable linear recurrence $g_\Omega(z, o) = Az + Bo$ with $A + A^\top \leq -2I_m$, as in state space models (S4, LRU), is a special case; the nonlinear coupled recurrences of GRUs and LSTMs are covered as well.

This monotonicity directly enforces the fading memory property: it guaranties that the operating region is forward-invariant and that all latent trajectories exponentially converge towards each other, effectively erasing the initial condition.

Proposition 3.3 (Invariant and Contracting Flow)**.** *Under Assumptions 3.1 and 3.2, the compact set $\mathcal{Z}$ is forward-invariant for the latent dynamics in (6), for any input $o(t) \in \mathcal{O}$. Moreover, for any*

$\kappa > 0$ and for any two latent trajectories $z_a(t), z_b(t)$ in $\mathcal{Z}$, driven by the same observation $o(t) \in \mathcal{O}$, we have a contraction:

$$|z_a(t) - z_b(t)| \leq e^{-\kappa(t-t_0)}|z_a(t_0) - z_b(t_0)|, \quad \forall t \geq t_0.$$

The proof is in Section C. The scalar parameter κ acts as a tunable gain that globally scales the vector field. By increasing κ, we accelerate the internal dynamics of the observer relative to the evolution of the input $o(t)$, ensuring a convergence rate proportional to $e^{-\kappa t}$.

3.2 Existence of the State-to-Latent Map

With forgetting its initialization, the latent state becomes entirely driven by the physical system. Specifically, $z(t)$ asymptotically converges to an image of the physical state, denoted $\boldsymbol{T}_\kappa(x(t))$. See Section D for a proof of the result below.

Theorem 3.4 (Existence of the State-to-Latent Map)**.** *Suppose Assumptions 3.1 and 3.2 are satisfied. For any gain κ large enough ($\kappa > 2L_f$, with L_f the Lipschitz constant of f over $\breve{\mathcal{X}}$), there exists a mapping $\boldsymbol{T}_\kappa : \mathcal{X} \to \mathcal{Z}$ such that:*

1. ***Invariance:*** *The physical trajectories are mapped to the latent trajectories:*
$$\boldsymbol{T}_\kappa(x(t)) = z(t), \quad \forall t \geq 0 \quad \text{if } z(0) = \boldsymbol{T}_\kappa(x_0). \tag{8}$$
2. ***Attractivity:*** *For any initialization $z_0 \in \mathcal{Z}$, the latent state asymptotically tracks the image by $\boldsymbol{T}_\kappa$ of the physical state:*
$$|z(t) - \boldsymbol{T}_\kappa(x(t))| \leq e^{-\kappa t}|z_0 - \boldsymbol{T}_\kappa(x_0)|, \quad \forall t \geq 0. \tag{9}$$
3. ***Regularity:*** *$\boldsymbol{T}_\kappa$ is C^1 on $\mathcal{X}$ with Lipschitz partial derivatives.*

3.3 Information Preservation: Injectivity of the State-to-Latent Map

Theorem 3.4 guarantees that the RNN creates a latent representation of x as $z = \boldsymbol{T}_\kappa(x)$. However, for the MLP decoder to be able to reconstruct the physical state via $\hat{x} = \gamma_\Omega(z)$, this representation must preserve the original information. Mathematically, it means that $\boldsymbol{T}_\kappa$ must be *injective*. The injectivity needs two properties: the physical system must give enough information in its observations (Observability), and the RNN must have a sufficient dimension and algebraic structure to process it (Filter Rank Condition).

Assumption 3.5 (Differential Observability)**.** *There exists an integer $\ell \geq 0$ such that the physical state x is uniquely determined by the observation and its ℓ first time-derivatives. More precisely, there exists a constant $c_{obs} > 0$ such that the map $x \mapsto (h(x), \mathcal{L}_f h(x), \ldots, \mathcal{L}_f^\ell h(x))$ is Lipschitz injective on $\mathcal{X}$:*

$$\sum_{i=0}^{\ell} |\mathcal{L}_f^i h(x_a) - \mathcal{L}_f^i h(x_b)| \geq c_{obs}|x_a - x_b|, \quad \forall x_a, x_b \in \mathcal{X}, \tag{10}$$

where $\mathcal{L}_f^i h$ is the i-th Lie derivative of h along the vector field f.

Assumption 3.6 (Filter Rank and Capacity Condition)**.** *The base dynamics g_Ω must satisfy the following conditions:*

1. ***Capacity:*** *The dimension of the latent space satisfies $m \geq p(\ell + 1)$, where $p = \dim(o)$.*
2. ***Input-to-Dynamics Injectivity :*** *For any $z \in \mathcal{Z}$, the map $o \mapsto g_\Omega(z, o)$ is injective on $\mathcal{O}$.*
3. ***Rank Condition:*** *By Assumption 3.2, for any o in $\mathcal{O}$, there exists a unique equilibrium $\varphi_0(o)$ such that $g_\Omega(\varphi_0(o), o) = 0$. Let the Jacobians be evaluated at this equilibrium:*
$$A(o) = \frac{\partial g_\Omega}{\partial z}(\varphi_0(o), o), \quad B(o) = \frac{\partial g_\Omega}{\partial o}(\varphi_0(o), o).$$
The $(m, p(\ell + 1))$ matrix $C_\ell(o)$ defined as:
$$C_\ell(o) = \begin{bmatrix} B(o) & A(o)B(o) & \cdots & A(o)^\ell B(o) \end{bmatrix} \tag{11}$$
is left-invertible for all $o \in \mathcal{O}$.

The rank condition is related to the left-invertibility [30] of the linear control system $\frac{d\xi}{dt} = A(o)\xi + B(o)u$ with o fixed. This is considered to be a mild assumption in nonlinear control, as almost

any linear control system satisfies it [7] (see also Appendix K for a detailed analysis). Under these assumptions, we can show that increasing the gain κ forces the latent state to contain the necessary measurement derivatives, which makes the mapping invertible.

Theorem 3.7 (Lipschitz Injectivity of $\boldsymbol{T}_\kappa$)**.** *Under Assumptions 3.1, 3.2, 3.5, and 3.6, there exist a constant $c_T > 0$ and a minimal gain $\underline{\kappa}$ such that for all $\kappa \geq \underline{\kappa}$, there exists $\boldsymbol{T}_\kappa : \mathcal{X} \to \mathbb{R}^m$ which satisfies the conclusion of Theorem 3.4 and which is Lipschitz injective:*

$$c_T \kappa^\ell |\boldsymbol{T}_\kappa(x_a) - \boldsymbol{T}_\kappa(x_b)| \geq |x_a - x_b|, \quad \forall x_a, x_b \in \mathcal{X}. \tag{12}$$

As a consequence, there exists a globally Lipschitz left-inverse $\boldsymbol{T}_\kappa^{-1} : \boldsymbol{T}_\kappa(\mathcal{X}) \to \mathcal{X}$ mapping the latent state back to the physical state.

The proof is in Section E. Theorem 3.7 is a fundamental result for our architecture. It proves that if we use a contracting Neural ODE with a high gain κ, the image of $\boldsymbol{T}_\kappa$ contains all the necessary information to uniquely compute $x(t)$. Note that what has to be learned is not $\boldsymbol{T}_\kappa$ but its inverse; the role of $\boldsymbol{T}_\kappa$ in our analysis is to establish that this inverse exists, which makes learning the decoder well posed. Theorems 3.4 and 3.7 being sufficient conditions, nothing says that a moderate gain fails; this is the regime our trained models operate in.

Remark 3.8 ($\boldsymbol{T}_\kappa$ as filtered output derivatives)**.** *The map $\boldsymbol{T}_\kappa$ is characterized by the nonlinear KKL equation $\frac{\partial \boldsymbol{T}_\kappa}{\partial x}(x) f(x) = \kappa\, g_\Omega(\boldsymbol{T}_\kappa(x), h(x))$, which follows from the invariance property of Theorem 3.4. Its relationship with the output and its successive time derivatives $(h(x), \mathcal{L}_f h(x), \dots, \mathcal{L}_f^\ell h(x))$ appearing in Assumption 3.5 is best seen in the linear case $\dot{x} = Fx$, $o = Hx$, $g_\Omega(z, o) = Az + Bo$ with A Hurwitz. Then $\boldsymbol{T}_\kappa$ is the unique solution of the Sylvester equation $\boldsymbol{T}_\kappa F = \kappa(A\boldsymbol{T}_\kappa + BH)$ and, for κ large, it expands as*

$$\boldsymbol{T}_\kappa\, x = -\sum_{i \geq 0} \kappa^{-i} A^{-(i+1)} B\, (HF^i x), \tag{13}$$

where $HF^i x$ is the i-th derivative of the output. Hence $\boldsymbol{T}_\kappa$ is not the observability matrix $(H; HF; \dots; HF^\ell)$ itself but a filtered version of it: the i-th output derivative appears at scale κ^{-i}, weighted by the resolvent of the latent dynamics, and left-inverting the sum truncated at $i = \ell$ requires the left-invertibility of $[A^{-1}B, \dots, A^{-(\ell+1)}B]$, i.e., the rank condition of Assumption 3.6. This is the origin of the κ^ℓ factor in Theorem 3.7 and of the requirement $m \geq p(\ell + 1)$, and it explains the interest of the approach: instead of differentiating o ℓ times, which is ill-posed under noise, the latent state carries the same information through a stable filter whose bandwidth is set by κ (Proposition 4.3).

4 Tunability and Decoder Complexity

In Section 3, we established that a sufficiently large gain κ guarantees the existence of a Lipschitz-continuous left-inverse map $\boldsymbol{T}_\kappa^{-1}$. We now study the convergence of the estimation error when we use this inverse map to decode the latent state. Then, we analyze the practical case where this theoretical map is approximated by a neural network.

4.1 Convergence of the Ideal Observer

Let us first consider the ideal case where an exact left-inverse map $\boldsymbol{T}_\kappa^{-1}$ is available. We define the state estimate as:

$$\hat{x}(t) = \boldsymbol{T}_\kappa^{-1}(z(t)). \tag{14}$$

We now prove that this estimator is formally *tunable*. An observer is said to be tunable if, for any desired precision and any given time, we can choose its parameters to guarantee that the estimation error is smaller than this precision after this time.

Theorem 4.1 (Exponential Convergence and Tunability)**.** *Under the assumptions of Theorem 3.7 and given $\underline{\kappa}$ as in Theorem 3.7, there exists $c_T' > 0$ such that for any $\kappa \geq \underline{\kappa}$, the state estimate $\hat{x}(t)$ satisfies:*

$$|\hat{x}(t) - x(t)| \leq c_T' \kappa^\ell e^{-\kappa t} |z_0 - \boldsymbol{T}_\kappa(x_0)|, \quad \forall t \geq 0, \tag{15}$$

for any physical trajectory $x(t)$ and any initial latent state $z_0 \in \mathcal{Z}$ and moreover, the architecture is tunable.

The proof is given in Section F.2. This result highlights a fundamental trade-off. Increasing the gain κ gives a faster exponential decay of the error. However, it may also increase the Lipschitz constant of the inverse map (which scales as κ^ℓ), potentially amplifying the initial transient error before the exponential term dominates. The same factor governs the sensitivity to measurement noise (Proposition 4.3).

4.2 MLP Approximation and Complexity Bound

In practice, the theoretical map $\boldsymbol{T}_\kappa^{-1}$ is not available analytically. Instead, we use a Multi-Layer Perceptron (MLP) as the decoder $\gamma_\Omega \approx \boldsymbol{T}_\kappa^{-1}$, and the practical estimate is given by $\hat{x}(t) = \gamma_\Omega(z(t))$.

Note that the strict left-inverse $\boldsymbol{T}_\kappa^{-1}$ is only defined on the image $\boldsymbol{T}_\kappa(\mathcal{X})$, which is a compact subset of $\mathcal{Z}$. By applying the McShane-Whitney extension theorem component-wise, this Lipschitz-continuous inverse can be extended to a globally Lipschitz function defined on the entire operating region $\mathcal{Z}$, preserving its Lipschitz constant $c_T\kappa^\ell$. The role of the MLP decoder γ_Ω is to approximate this extended map.

Since the extended $\boldsymbol{T}_\kappa^{-1}$ is a Lipschitz-continuous function on a compact set, the Universal Approximation Theorem guarantees that an MLP can approximate it with arbitrary precision. We now quantify the required size of the network with respect to the gain κ to ensure tunability.

Theorem 4.2 (Neural Decoder Complexity)**.** *Suppose the assumptions of Theorem 4.1 hold. For any desired precision $\epsilon > 0$ and any observation time $t_o > 0$, there exists a structural gain $\kappa > 0$ and a ReLU MLP decoder γ_Ω such that the neural estimate satisfies:*

$$|\gamma_\Omega(z(t)) - x(t)| \leq \epsilon, \quad \forall t \geq t_o. \tag{16}$$

Furthermore, the required complexity of the neural network (i.e., its total number of weights $\mathcal{W}$) grows polynomially with the gain κ and the inverse precision $1/\epsilon$. More precisely, there exists a constant $C_{net} > 0$ (depending only on the dimensions n and m) such that:

$$\mathcal{W} \leq C_{net} \left(\frac{\kappa^\ell}{\epsilon}\right)^m \log\left(\frac{\kappa^\ell}{\epsilon}\right). \tag{17}$$

The proof can be found in Section F.3. This theorem provides a formal justification for the RNN-MLP architecture and shows that it remains tunable even when the exact inverse map is replaced by an MLP. However, it also clearly exposes the cost of this tunability: accelerating the continuous dynamics to improve the time-scale separation (increasing κ) makes the state-to-latent mapping more sensitive, requiring a larger network to learn its inverse with sufficient precision.

4.3 Robustness to measurement disturbances

The following input-to-state estimate quantifies the sensitivity of the RNN-MLP estimator to measurement noise, and makes the classical high-gain trade-off explicit.

Proposition 4.3 (Robustness to measurement disturbances)**.** *Suppose the assumptions of Theorem 4.1 hold. Let $d_0 > 0$, let $\bar{L}_{go} = \max\{|\frac{\partial g_\Omega}{\partial o}(z,o)| : z \in \mathcal{Z},\ \mathrm{dist}(o,\mathcal{O}) \leq d_0\}$ and $\bar{d} = \min\{d_0, R_z/(4\bar{L}_{go})\}$. Let the latent dynamics in (6) be driven by the corrupted observation $o(t) + \delta(t)$, where δ is a Lebesgue measurable disturbance (no regularity in time is assumed) such that $\|\delta\|_\infty = \operatorname{ess\,sup}_{t\geq 0} |\delta(t)| \leq \bar{d}$, and let the estimate be $\hat{x}(t) = \gamma_\Omega(z(t))$, where the decoder γ_Ω satisfies $|\gamma_\Omega(z) - \boldsymbol{T}_\kappa^{-1}(z)| \leq \epsilon_{dec}$ for all $z \in \mathcal{Z}$, for some $\epsilon_{dec} \geq 0$ ($\epsilon_{dec} = 0$ for the exact inverse). Then $\mathcal{Z}$ remains forward invariant for the latent dynamics in (6) and, for all $\kappa \geq \underline{\kappa}$ and any initial latent state $z_0 \in \mathcal{Z}$,*

$$|\hat{x}(t) - x(t)| \leq \epsilon_{dec} + c_T'\kappa^\ell e^{-\kappa t}|z_0 - \boldsymbol{T}_\kappa(x_0)| + 2\,c_T'\kappa^\ell \bar{L}_{go}\, d(t), \quad \forall t \geq 0, \tag{18}$$

where $d(t) = \operatorname{ess\,sup}_{s\in[0,t]} e^{-\kappa(t-s)/2}|\delta(s)| \leq \|\delta\|_\infty$.

The proof is given in Section F.4. In particular, the ReLU MLP decoder of Theorem 4.2 satisfies $|\gamma_\Omega(z(t)) - x(t)| \leq \epsilon + 2c_T'\kappa^\ell \bar{L}_{go}\, d(t)$ for all $t \geq t_o$. The term $d(t)$ forgets past disturbances at rate $\kappa/2$: if δ vanishes after a time t_1, then $d(t) \leq e^{-\kappa(t-t_1)/2}\|\delta\|_\infty$. For a persistent disturbance, the noise floor $2c_T'\kappa^\ell \bar{L}_{go}\|\delta\|_\infty$ is the high-gain effect: raising κ accelerates the transient exponentially but raises the asymptotic error polynomially, with a degree set by the observability index ℓ. The transient falls below the floor once $e^{-\kappa t_o} \approx \bar{L}_{go}\|\delta\|_\infty$, so for a target settling time t_o the useful gain

is $\kappa \approx t_o^{-1} \log(1/(\bar{L}_{go}\|\delta\|_\infty))$, beyond which any increase only raises the floor. In practice κ is absorbed into the freely learned weights of g_Ω (as (A, B) in numerical KKL methods); we fix $\kappa = 1$ in the robotics experiments.

5 Discrete-Time Implementation: Residual RNNs

This section formally establishes that the discrete-time implementation of our contracting latent dynamics, which naturally takes the form of a standard Residual Recurrent Neural Network (RNN), preserves the fundamental property of tunability established in the continuous-time case.

System and Discrete Observations. While the theoretical analysis relies on the continuous-time generative model (1), in practice, observations are collected at discrete time steps $t_k = k\delta$, where $\delta > 0$ is the sampling period. We denote the discrete observation sequence as:

$$o_k = h(\Phi_f(x_0, t_k)). \tag{19}$$

Residual RNN as an Euler Discretization. The theoretical continuous-time observer is defined by $\dot{z} = \kappa g_\Omega(z, o)$. Its discrete implementation via an explicit Euler numerical scheme is given by:

$$z_{k+1} = z_k + \delta\kappa g_\Omega(z_k, o_k) = \text{RNN}_\Omega(z_k, o_k). \tag{20}$$

Note that corresponds to the architecture of a Residual RNN, where $z_k \in \mathbb{R}^m$ acts as the hidden state.

Remark 5.1 (Contraction via Bounded Activations)**.** *Consider the base dynamics $g_\Omega(z, o) = -z + \tanh(Wz + Uo + b)$, which is standard in RNN architectures. Since the hyperbolic tangent is bounded by 1 element-wise, its Euclidean norm is bounded by $\sqrt{m}$. Let us define the compact set $\mathcal{Z} = \{z \in \mathbb{R}^m \mid |z| \leq R\}$ with a radius $R > \sqrt{m}$. For any z on the boundary of $\mathcal{Z}$, the linear dissipative term $-z$ strictly dominates the bounded term $\tanh(\dots)$, ensuring that the vector field points inward. The discrete update becomes $z_{k+1} = (1 - \delta\kappa)z_k + \delta\kappa \tanh(Wz_k + Uo_k + b)$. The parameters $\Omega = (W, U, b)$ can then be chosen to ensure the monotonicity condition (7) specifically for $z \in \mathcal{Z}$, which is a less restrictive constraint than global contraction.*

The discrete state estimate is finally decoded by the MLP:

$$\hat{x}_k = \gamma_\Omega(z_k) \approx \boldsymbol{T}_\kappa^{-1}(z_k). \tag{21}$$

Discrete Tunability and Main Result. We now adapt our definition of tunability to the discrete context. Here, the sampling period δ is treated as a design variable. Consequently, the concept of tunability requires the adjustment of this time step to achieve the desired accuracy.

Definition 5.2 (Discrete Tunability)**.** *The Residual RNN architecture is said to be tunable if, for any error threshold $\epsilon > 0$ and any observation time $t_o > 0$, there exist a structural gain $\kappa > 0$ and a maximum sampling step δ^* such that, for all $\delta \leq \delta^*$, the state estimate satisfies:*

$$|\hat{x}_k - x(t_k)| \leq \epsilon, \quad \forall k \text{ such that } t_k \geq t_o, \tag{22}$$

for any physical initial condition $x_0 \in \mathcal{X}$ and any latent initialization $z_0 \in \mathcal{Z}$.

Theorem 5.3 (Tunability of the Residual RNN)**.** *Under the assumptions of Theorem 3.7 (Observability, Latent Monotonicity, and Filter Rank Condition), the Residual RNN defined by (20) with the ideal decoder $\hat{x}_k = \boldsymbol{T}_\kappa^{-1}(z_k)$ is tunable.*

The proof is in Section G. This theorem gives a formal justification for the use of RNNs in state estimation tasks: by choosing an architecture that respects local contraction, we can guarantee that an arbitrary precision is achievable provided the sampling frequency is sufficiently high.

6 Experimental results

We provide experiments on two different types of data: a real robotics application trained from large-scale data, and a simulated system which allow us to verify theoretical claims.

A robotics application: learning pose corrections from real large-scale data. Our objective is to estimate the pose of a mobile robot with respect to its starting position. It is equipped with sensors providing relative odometry $d_t = (\mathrm{x}_t, \mathrm{y}_t, \theta_t, v_t, \omega_t)^\top$ of pose and velocities obtained from wheel encoders and IMUs. We are interested in estimating the pose $\hat{s}_t$ in the coordinate frame of the episode start t=0, which classically can be obtained by kinematic integration of odometry readings over time ("*Dead Reckoning*"), which is subject to drift: $\hat{s}_t^{raw} = \text{Integrate}(\hat{s}_{t-1}^{raw}, d_t)$.



Figure 2: **Robot**: raw *vs* corrected trajectory.

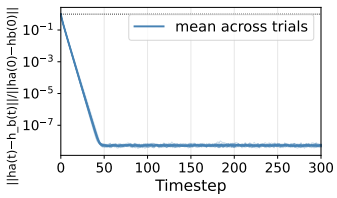

Figure 3: **Empirical contraction** of the classical GRU. The network rapidly forgets its initial state.

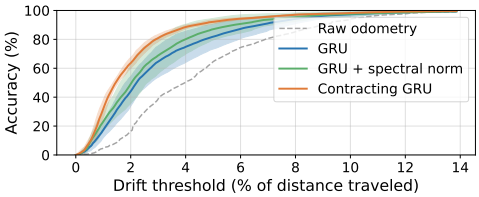

Figure 4: **Pose correction**: all models obtain strong gains but *Contracting GRU* outperforms the unconstrained GRU and is more stable (mean ±std-dev over 5 training runs).

We propose a learning based system to improve the odometry measurements through "privileged information" available during training but not during testing when the robot is deployed. We have collected a large-scale dataset of trajectories obtained with a real mobile service robot (cf. Figure 2) in 4 different buildings on two continents spanning 14 floors, 27km and 18h of navigation which was split into train/val/test, see Table 1 in Section H.3. Beyond odometry the measurements also include localization from a Lidar based *Adaptive Monte Carlo Localization* (AMCL) [54] system. These noisy pose readings s_t^{amcl} are thus available during training but not during testing. The goal is to learn a correction model such that the corrected trajectory $\hat{s}_t^{corr} \approx s_t^{\text{amcl}}$, using only information available at runtime (no map).

The correction network learns a dynamical model of the robot and the odometry error capable of correcting pose measurements d_t and replacing them with improved measurements $d_t + P\hat{y}_t$, where $\hat{y}_t \in \mathbb{R}^3$ is a spatial correction term obtained by a recurrent network maintaining a hidden state z_t, and $P \in \mathbb{R}^{5\times 3}$ is a projection matrix mapping the spatial components to the full odometry state. The corrected trajectory is obtained by kinematic integration:

$$z_{t+1} = \varphi_\Omega(z_t, o_t) \tag{23}$$

$$\hat{y}_t = \gamma_\Omega(z_t), \tag{24}$$

$$\hat{s}_t^{corr} = \text{Integrate}(\hat{s}_{t-1}^{corr}, d_t + P\hat{y}_t), \tag{25}$$

where o_t is a 9-dimensional feature vector containing odometry readings, and command velocities,

$$o_t = \left[\mathrm{x}_t,\ \mathrm{y}_t,\ \theta_t,\ v_t,\ \omega_t,\ v_t^{\text{cmd}},\ \omega_t^{\text{cmd}},\ v_t^{\text{cmd}} - v_t,\ \omega_t^{\text{cmd}} - \omega_t\right] \tag{26}$$

The last two values are *slip features* and encode the instantaneous mismatch between commanded and measured motion. We propose a **trajectory integration loss** that directly penalizes accumulated drift. Trajectories are divided into non-overlapping windows of length $H{=}50$. Within each window, both the corrected and target body-frame displacements are integrated kinematically, and position MSE is computed on the resulting trajectories. Details on losses and training hyper-parameters are provided in Section H.1.

Gated RNN-MLP architectures: we propose to address this problem with two different RNN architectures in their gated state-of-the-art variants: (i) "**GRU**", a vanilla GRU [10] without any constraints on contraction, with 2-layers, hidden dim of 128, followed by a two-layer MLP head (128→64→3). (ii) "**Contracting GRU**", a two layer RNN with an update very much inspired by a GRU but with built-in contracting properties satisfying Assumption 3.2. This is enforced with spectral normalization and choosing appropriate hyper-parameters, see Section H.2 and H.4. (iii) **GRU + spectral norm** which ablates the effect of spectral normalization alone on the vanilla GRU. The filter rank condition of Assumption 3.6 is numerically tested and validated in Section H.5.

The GRU learns contraction. In Figure 3 we provide empirical evidence that on this application the classical GRU without inductive bias on contraction properties does learn a contracting behavior. Two network copies (a, b) are initialized with $z_a(0) \neq z_b(0)$ and driven by the same inputs. We show the normalized state distance $|z_a(t) - z_b(t)|/|z_a(0) - z_b(0)|$ on a log scale for 20 independent trials (light) and their mean. The network forgets its initial condition under any input.

Results. Figure 4 gives a plot of pose estimation accuracy as a function of the relative pose estimation error. An accuracy of 80% at 2.0 means that 80% of poses are correctly detected with an error which is lower than 2% of the trajectory length. We can see that both "GRU" and "Contracting GRU" achieve strong gains, in particular for higher required precision (left part of the plot). However, the contracting GRU does outperform the classical GRU and, importantly, shows a far more stable training behavior when multiple seeds are compared.

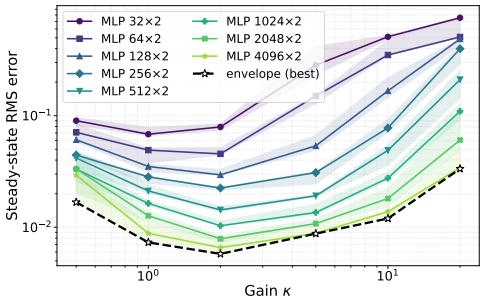

Figure 5: Steady-state RMS vs. gain κ, for 2-layer MLP decoders of varying width w.

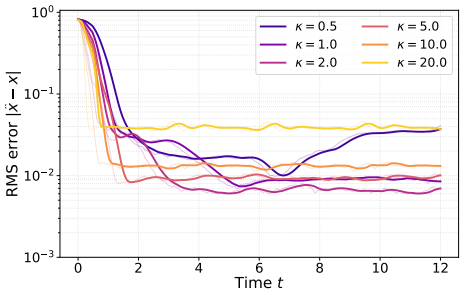

Figure 6: Reconstruction error over time at varying κ, with the widest (=4096) MLP.

Simulated system. To complete the experimental part with a simulated system, we now consider a nonlinear oscillator with unknown parameter:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -x_1 + (x_3 - x_1^2 - x_2^2)x_2, \quad \dot{x}_3 = 0 \tag{27}$$

where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ is the state (encompassing both the state (x_1, x_2) of the oscillator and the unknown positive parameter x_3), the observation is $o = x_1$ and the target variable to be estimated is $s = x_3$. In Appendix J, we define a compact forward-invariant set $\mathcal{X} \subset \mathbb{R}^3$ of trajectories on which system (27) is differentially observable of order $\ell = 3$ (see Assumption 3.5).

Figures 5 and 6 illustrate tunability, i.e. Theorem 4.1. The theoretical guarantee states that *there exist* weights achieving the bound for large κ, so we train a separate observer for each value of $\kappa \in \{0.5, 1, 2, 5, 10, 20\}$ and each decoder width $w \in \{32, 64, 128, 256, 512, 1024, 2048, 4096\}$ and 3 seeds. The capacity sweep (Fig. 5) shows that for every fixed width the error decreases up to $\kappa \approx 2$ and then re-grows, in line with the κ^ℓ decoder-Lipschitz cost predicted by Theorem 4.2; widening the decoder lowers absolute error. The time-resolved view at the widest decoder (Fig. 6) confirms that larger κ accelerates the transient, although a slight steady-state penalty remains at $\kappa = 20$.

7 Discussion, limitations and conclusion

Viewing RNN-MLP estimators as time-discretized continuous systems, we extended nonlinear KKL observer theory from linear filters to contracting latent dynamics with dense Jacobians: the state-to-latent map preserves information under a generic rank condition, the architecture is tunable through a global gain κ, the decoder can be an MLP of quantified size, and these guarantees extend to the discrete-time implementation. On a real pose-estimation task, a provably contracting model outperforms a baseline with increased training stability, and unconstrained GRUs converge to contracting dynamics, suggesting that contraction is an emergent property of well-trained sequence models.

Scope of the assumptions. Differential observability (Assumption 3.5) is a property of the physical system, not of the method: for analytic observable systems it is necessary for the problem to be well posed, since no estimator, learned or model-based, can recover in arbitrarily short time what the observations do not contain. Contraction (Assumption 3.2) holds by construction in the architecture of Appendix H.2, and the filter rank condition (Assumption 3.6) is generic (Appendix K) and cheap to check after training by automatic differentiation (Appendix H.5). Reconstructing only a function $\varsigma(x)$ of a non-observable state would require a functional observability framework [36], developed for KKL observers with linear latent dynamics [48] but not yet with nonlinear ones.

Noise and the curse of the observability index. Our generative model is deterministic. Proposition 4.3 shows that contraction nevertheless yields robustness to bounded measurement disturbances, with an explicit noise floor in κ^ℓ; the same exponent governs the decoder capacity $\mathcal{O}(\kappa^{\ell m})$ in Theorem 4.2. The observability index ℓ thus controls both faces of the trade-off between convergence speed and noise amplification or decoder complexity. A stochastic treatment (e.g. bounding the error variance under stochastic differential equations) is left for future work.

Optimization and training dynamics. Our theorems establish the *existence* of suitable parameters Ω (any architecture enjoying a universal approximation theorem could replace the MLP in Theorem 4.2), not a constructive algorithm: enforcing the rank condition during training is non-trivial, and our evidence is empirical (Appendices H.2, H.5). Training algorithms enforcing these conditions are a natural next step.

References

[1] Vincent Andrieu, Bayu Jayawardhana, and Sophie Tarbouriech. Some results on exponential synchronization of nonlinear systems. *IEEE Transactions on Automatic Control*, 63(4):1213–1219, 2018.

[2] Vincent Andrieu and Laurent Praly. On the existence of a kazantzis–kravaris/luenberger observer. *SIAM Journal on Control and Optimization*, 45(2):432–456, 2006.

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. In *arxiv:2506.09985*, 2025.

[4] Leonard Bauersfeld, Elia Kaufmann, Philipp Foehn, Sihao Sun, and Davide Scaramuzza. Neurobem: Hybrid aerodynamic quadrotor model. *arXiv preprint*, 2021.

[5] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[6] Pauline Bernard, Vincent Andrieu, and Daniele Astolfi. Observer design for continuous-time dynamical systems. *Annual Reviews in Control*, 53:224–248, 2022.

[7] Lucas Brivadis, Vincent Andrieu, Pauline Bernard, and Ulysse Serres. Further remarks on kkl observers. *Systems & Control Letters*, 172:105429, 2023.

[8] Mona Buisson-Fenet, Lukas Bahr, Valery Morgenthaler, and Florent Di Meglio. Towards gain tuning for numerical kkl observers. *IFAC-PapersOnLine*, 56(2):4061–4067, 2023.

[9] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[11] Mohamed Darouach. Existence and design of functional observers for linear systems. *IEEE Transactions on Automatic Control*, 45(5):940–943, 2000.

[12] Laura Boca de Giuli, Alessio La Bella, Marcello Farina, and Riccardo Scattolini. Stability and learning of networked systems with application to physics-informed neural networks. *Automatica*, 189:113005, 2026.

[13] José de Jesús Rubio and Wen Yu. Nonlinear system identification with recurrent neural networks and dead-zone kalman filter algorithm. *Neurocomputing*, 2007.

[14] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. Goal-conditioned imitation learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[15] Tyrone L. Fernando, Hieu M. Trinh, and Les Jennings. Functional observability and the design of minimum order linear functional observers. *IEEE Transactions on Automatic Control*, 55(5):1268–1273, 2010.

[16] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[17] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

[18] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

[19] J. Hale. *Ordinary differential equation*. Krieger Publishing Company, Malabar, FL, second edition, 1980.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computing*, 1997.

[21] Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical Report 148, German National Research Center for Information Technology (GMD), 2001.

[22] Steeven Janny, Vincent Andrieu, Madiha Nadri, and Christian Wolf. Deep KKL: Data-driven output prediction for non-linear systems. In *60th IEEE Conference on Decision and Control (CDC)*, pages 4376–4381, 2021.

[23] Steeven Janny, Madiha Nadri, Julie Digne, and Christian Wolf. Space and time continuous physics simulation from partial observations. In *International Conference on Learning Representations (ICLR)*, 2024.

[24] Steeven Janny, Hervé Poirier, Leonid Antsfeld, Guillaume Bono, Gianluca Monaci, Boris Chidlovskii, Francesco Giuliari, Alessio Del Bue, and Christian Wolf. Reasoning in visual navigation of end-to-end trained agents: a dynamical systems approach. In *CVPR*, 2025.

[25] Łukasz Kaiser and Ilya Sutskever. Neural gpus learn algorithms. In *International Conference on Learning Representations (ICLR)*, 2016.

[26] N. Kazantzis and C. Kravaris. Nonlinear observer design using Lyapunov's auxiliary theorem. *Systems & Control Letters*, 34(5):241–247, 1998.

[27] V. Lakshmikantham and S. Leela. *Differential and Integral Inequalities: Theory and Applications*, volume 1 of *Mathematics in Science and Engineering*. Academic Press, New York, 1969. Ordinary Differential Equations.

[28] Ljung Lennart. System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, 28:540, 1999.

[29] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat. State Representation Learning for Control: An Overview. *Neural Networks*, 2018.

[30] Michael Di Loreto and Damien Eberard. Strong left inversion of linear systems and input reconstruction. *IEEE Transactions on Automatic Control*, 68(6):3612–3617, 2023.

[31] Zhixin Lu, Jaideep Pathak, Brian Hunt, Michelle Girvan, Roger Brockett, and Edward Ott. Reservoir observers: Model-free inference of unmeasured variables in chaotic systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(4):041102, 2017.

[32] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer science review*, 3(3):127–149, 2009.

[33] Yasmine Marani, Israel Filho, Tareq Al-Naffouri, and Taous Meriem Laleg Kirati. Unsupervised physics-informed neural network-based nonlinear observer design for autonomous systems using contraction analysis. In *2025 European Control Conference (ECC)*, pages 2327–2332, 2025.

[34] Keyan Miao and Konstantinos Gatsis. Learning robust state observers using neural odes. In *Learning for Dynamics and Control Conference*, pages 208–219. PMLR, 2023.

[35] Boris S. Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107(3):529–530, 2020.

[36] Arthur N. Montanari, Chao Duan, Luis A. Aguirre, and Adilson E. Motter. Functional observability and target state estimation in large-scale networks. *Proceedings of the National Academy of Sciences*, 119(1):e2113750119, 2022.

[37] Muhammad Umar B Niazi, John Cao, Xudong Sun, Amritam Das, and Karl Henrik Johansson. Learning-based design of luenberger observers for autonomous nonlinear systems. In *2023 American Control Conference (ACC)*, pages 3048–3055. IEEE, 2023.

[38] Henk Nijmeijer and Iven M. Y. Mareels. An observer looks at synchronization. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(10):882–890, 1997.

[39] Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting Recurrent Neural Networks for Long Sequences. In *International Conference on Machine Learning (ICML)*, 2023.

[40] Victor Pachy, Vincent Andrieu, Pauline Bernard, Lucas Brivadis, and Laurent Praly. On the existence of kkl observers with nonlinear contracting dynamics. *IFAC-PapersOnLine*, 58(21):262–267, 2024.

[41] J. Peralez and M. Nadri. Deep learning-based luenberger observer design for discrete-time nonlinear systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 1932–1937. IEEE, 2021.

[42] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2020.

[43] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. In *NeurIPS*, 2024.

[44] L. d. C. Ramos, F. Di Meglio, V. Morgenthaler, L. F. F. da Silva, and P. Bernard. Numerical design of luenberger observers for nonlinear systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 5328–5333. IEEE, 2020.

[45] Michael S. Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. Token turing machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[46] Mert Bulent Sariyildiz, Philippe Weinzaepfel, Guillaume Bono, Gianluca Monaci, and Christian Wolf. Kinaema: a recurrent sequence model for memory and pose in motion. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[47] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[48] Mario Spirito, Pauline Bernard, and Lorenzo Marconi. On the existence of robust functional KKL observers. In *2022 American Control Conference (ACC)*, pages 2982–2987, 2022.

[49] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020.

[50] A Stubberud, H Wabgaonkar, and S Stubberud. A neural-network-based system identification technique. In *Conference on Decision and Control*, 1991.

[51] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

[52] Jacek Szarski. *Differential Inequalities*, volume 43 of *Monografie Matematyczne*. Polish Scientific Publishers, Warsaw, 1965.

[53] Wentao Tang. Synthesis of data-driven nonlinear state observers using lipschitz-bounded neural networks. In *2024 American Control Conference (ACC)*, pages 1713–1719. IEEE, 2024.

[54] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press Cambridge, 2005.

[55] Sjoerd van Steenkiste, Daniel Zoran, Yi Yang, Yulia Rubanova, Rishabh Kabra, Carl Doersch, Dilara Gokay, Joseph Heyward, Etienne Pot, Klaus Greff, Drew A. Hudson, Thomas Albert Keck, Joao Carreira, Alexey Dosovitskiy, Mehdi S. M. Sajjadi, and Thomas Kipf. Moving off-the-grid: Scene-grounded video representations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[56] Henk J Van Waarde, Jaap Eising, Harry L Trentelman, and M Kanat Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11):4753–4768, 2020.

[57] Henk J Van Waarde and Rodolphe Sepulchre. Kernel-based models for system analysis. *IEEE Transactions on Automatic Control*, 68(9):5317–5332, 2022.

[58] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017. See also arXiv:1610.01145.

[59] Samuele Zoboli, Steeven Janny, and Mattia Giaccagli. Deep learning-based output tracking via regulation and contraction theory. In *International Federation of Automatic Control (IFAC)*, 2023.

Supplementary Material for: Provable State Estimation with Recurrent Models

A Roadmap of the Proofs

The appendix is long because the paper is a theoretical contribution and the proofs belong in it. This page gives the structure of the argument and the dependencies between its pieces, so that a reader can navigate the proofs or skip the ones they do not need.

Overall structure. The main results form a chain, each one building on the previous:

Lemma B.1 (Appendix B) $\Rightarrow$ Prop. 3.3 (Appendix C) $\Rightarrow$ Thm. 3.4 (Appendix D) $\Rightarrow$ Thm. 3.7 (Appendix E) $\Rightarrow$ Thms. 4.1, 4.2 and Prop. 4.3 (Appendix F) $\Rightarrow$ Thm. 5.3 (Appendix G).

Appendix B: a differential inequality. The main technical tool used throughout is Lemma B.1: a matrix solution of $\dot{\mathcal{M}} = A(t)\mathcal{M} + B(t)$ whose symmetric part $A + A^\top$ is bounded above by $2a_+ I$ satisfies $|\mathcal{M}(t)| \leq e^{a_+(t-t_0)}|\mathcal{M}(t_0)| + \int_{t_0}^{t} e^{a_+(t-s)}|B(s)|ds$. Every contraction, attractivity and robustness estimate of the paper is an instance of this lemma with a different choice of $\mathcal{M}$, A and B. Lemma B.2 complements it with an elementary invariance result for a ball, valid for inputs that are only measurable in time.

Appendix C: contraction and invariance (Proposition 3.3). The monotonicity of Assumption 3.2 is integrated along segments (mean value theorem) to show that (i) the vector field points inward on the boundary of the ball $\mathcal{Z}$, hence $\mathcal{Z}$ is forward invariant (Lemma B.2), and (ii) the difference of two trajectories obeys a linear time-varying equation with symmetric part $\leq -2\kappa I$, hence contracts at rate κ by Lemma B.1. A corollary (Appendix C.3) shows that the invariance survives an additive perturbation of the vector field of magnitude at most $\bar{g}$; it is reused in the injectivity and robustness proofs.

Appendix D: existence and regularity of $\boldsymbol{T}_\kappa$ (Theorem 3.4). The map is defined as $\boldsymbol{T}_\kappa(x_0) = \varpi(0)$, where ϖ is the unique solution of the latent dynamics that is bounded on the whole time axis when driven by the output of the physical trajectory through x_0, extended to negative times. Existence and uniqueness of ϖ follow from contraction (two bounded solutions must coincide as $\tau \to -\infty$). Invariance and attractivity are then direct consequences of Proposition 3.3. Regularity (C^1 with Lipschitz derivative) is the delicate part: the variational equation is analyzed with Lemma B.1, and the condition $\kappa > 2L_f$ ensures that the latent contraction dominates the expansion of the physical flow, so that the family of finite-horizon approximations is equi-Lipschitz and converges by the Ascoli-Arzelà theorem.

Appendix E: Lipschitz injectivity (Theorem 3.7). This is the heart of the paper and the point where the diagonality restriction of [40] is lifted. Working in the observer time $\tau = \kappa t$, in which the observation is slowly varying (its i-th derivative is $\mathcal{O}(\kappa^{-i})$), the bounded solution ϖ is approximated by an explicit sequence $\varphi_0, \ldots, \varphi_\ell$ of functions of the observation and its first derivatives, built recursively by inverting the base dynamics under a constant forcing (implicit function ψ, Lemma E.1; boundedness, Lemma E.2). Three facts are then established: the approximation map is Lipschitz injective with respect to the vector of observation derivatives, with constant of order κ^ℓ, thanks to the rank condition of Assumption 3.6 (Proposition E.3); the approximation error is of order $\kappa^{-(\ell+1)}$ (Proposition E.4); and the difference of two error trajectories obeys a fading-memory bound (Proposition E.6). The synthesis combines these with differential observability (Assumption 3.5) and a large enough κ to conclude.

Appendix F: decoding (Theorems 4.1, 4.2 and Proposition 4.3). The left inverse of $\boldsymbol{T}_\kappa$ is extended to the whole latent space by the McShane-Whitney theorem, preserving its Lipschitz constant up to a factor $\sqrt{n}$. Tunability with the exact inverse is then a one-line consequence of attractivity; the MLP complexity bound follows from the approximation theorem of [58] for Lipschitz functions; and the robustness estimate is obtained by showing that $\mathcal{Z}$ remains invariant under the disturbance and comparing the observer driven by the corrupted output with the nominal one, along the lines of Lemma B.1.

Appendix G: discrete time (Theorem 5.3). The residual RNN is an explicit Euler scheme for the continuous observer. A small-step analysis shows that the discrete update inherits the contraction

for $\delta\kappa$ small enough, and that the global Euler error on the finite horizon $[0, t_o]$ vanishes with δ; combined with Theorem 4.1 this gives discrete tunability.

Appendices H to K: experiments and genericity. Appendix H details the robotics experiments (losses, the Contracting GRU and its contraction certificate, datasets, hyper-parameters, empirical verification of the contraction constant and of the rank condition), Appendix J the simulated system (in particular its differential observability), and Appendix K proves that the rank condition of Assumption 3.6 is generically satisfied by analytic recurrent models.

B A Result on Differential Inequalities

To ease the reading of our proofs, we start by reminding the reader of a standard result on differential inequalities.

In what follows, we denote, for a matrix $\mathcal{M} \in \mathbb{R}^{n\times m}$, the induced Euclidean norm:

$$|\mathcal{M}| = \max_{v\in\mathbb{R}^m : |v|=1} |\mathcal{M}v| \tag{28}$$

Lemma B.1 (See for instance [52] or [27])**.** *Let $\mathcal{M}$ be an absolutely continuous matrix function satisfying the linear ODE*

$$\frac{d\mathcal{M}}{dt} = A(t)\mathcal{M} + B(t) \tag{29}$$

for almost all t, where $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ and $B : \mathbb{R} \to \mathbb{R}^{n\times m}$ are measurable and locally essentially bounded matrix functions satisfying, for almost all t,

$$2a_- I_n \leq A(t) + A(t)^\top \leq 2a_+ I_n \tag{30}$$

where a_+ and a_- are constant real numbers, with a_- possibly $-\infty$, and I_n is the identity matrix of dimension n.

Then we have, for all $t_0 \leq t$,

$$e^{a_-(t-t_0)}|\mathcal{M}(t_0)| - \int_{t_0}^{t} e^{a_-(t-s)}|B(s)|\,ds \leq |\mathcal{M}(t)| \leq e^{a_+(t-t_0)}|\mathcal{M}(t_0)| + \int_{t_0}^{t} e^{a_+(t-s)}|B(s)|\,ds \tag{31}$$

For example, when $a_- = -a_+$ and $B = 0$, we get (using $t_0 = s \leq 0$ and $t = 0$ for the left inequality, and $t_0 = 0$ and $t = s \geq 0$ for the right inequality):

$$e^{-a_+(-s)}|\mathcal{M}(s)| \leq |\mathcal{M}(0)| \quad \forall s \leq 0, \qquad \text{and} \qquad |\mathcal{M}(s)| \leq e^{a_+ s}|\mathcal{M}(0)| \quad \forall s \geq 0, \tag{32}$$

which implies:

$$|\mathcal{M}(s)| \leq e^{a_+|s|}|\mathcal{M}(0)| \qquad \forall s \in \mathbb{R}. \tag{33}$$

Proof. From the ODE, we get readily for any unit vector $v \in \mathbb{R}^m$ and almost all t,

$$\frac{d}{dt}\left\{v^\top \mathcal{M}^\top \mathcal{M} v\right\} = v^\top \mathcal{M}^\top (A(t) + A(t)^\top)\mathcal{M}v + 2v^\top \mathcal{M}^\top B(t)v.$$

Using the bounds on the symmetric part of $A(t)$ and the Cauchy-Schwarz inequality for the second term, we have when $\mathcal{M}v \neq 0$:

$$2a_-|\mathcal{M}v|^2 - 2|\mathcal{M}v||B(t)v| \leq 2|\mathcal{M}v|\frac{d}{dt}\{|\mathcal{M}v|\} \leq 2a_+|\mathcal{M}v|^2 + 2|\mathcal{M}v||B(t)v|.$$

Dividing by $2|\mathcal{M}v|$ yields:

$$a_-|\mathcal{M}v| - |B(t)v| \leq \frac{d}{dt}\{|\mathcal{M}v|\} \leq a_+|\mathcal{M}v| + |B(t)v|.$$

The function $t \mapsto |\mathcal{M}(t)v|$ is absolutely continuous, as the composition of the Lipschitz function $|\cdot|$ with an absolutely continuous function, hence differentiable almost everywhere. At a time where $\mathcal{M}(t)v = 0$ and this function is differentiable, its derivative vanishes, since 0 is its minimum value; the above inequalities therefore hold for almost all t, including when $\mathcal{M}v = 0$. Since $t \mapsto e^{-a_+ t}|\mathcal{M}(t)v|$ and $t \mapsto e^{-a_- t}|\mathcal{M}(t)v|$ are absolutely continuous, it follows that, for almost all t:

$$\frac{d}{dt}\left\{e^{-a_+ t}|\mathcal{M}v|\right\} \leq e^{-a_+ t}|B(t)v| \qquad \text{and} \qquad -e^{-a_- t}|B(t)v| \leq \frac{d}{dt}\left\{e^{-a_- t}|\mathcal{M}v|\right\}.$$

By integration, we obtain for all $t_0 \le t$ and for all v such that $|v| = 1$:

$$e^{-a_+ t}|\mathcal{M}(t)v| - e^{-a_+ t_0}|\mathcal{M}(t_0)v| \le \int_{t_0}^{t} e^{-a_+ s}|B(s)v|\,ds,$$

$$-\int_{t_0}^{t} e^{-a_- s}|B(s)v|\,ds \le e^{-a_- t}|\mathcal{M}(t)v| - e^{-a_- t_0}|\mathcal{M}(t_0)v|.$$

In particular, for $\bar{v}_+(t)$ maximizing $|\mathcal{M}(t)v|$, multiplying the first inequality by $e^{a_+ t}$ gives:

$$|\mathcal{M}(t)| \le e^{a_+(t-t_0)}|\mathcal{M}(t_0)\bar{v}_+(t)| + \int_{t_0}^{t} e^{a_+(t-s)}|B(s)\bar{v}_+(t)|\,ds$$

$$\le e^{a_+(t-t_0)}|\mathcal{M}(t_0)| + \int_{t_0}^{t} e^{a_+(t-s)}|B(s)|\,ds \qquad \forall t_0 \le t.$$

Similarly, for $\bar{v}_-(t_0)$ maximizing $|\mathcal{M}(t_0)v|$ (which implies $|\mathcal{M}(t)\bar{v}_-(t_0)| \le |\mathcal{M}(t)|$), multiplying the second inequality by $e^{a_- t}$ gives:

$$|\mathcal{M}(t)| \ge |\mathcal{M}(t)\bar{v}_-(t_0)| \ge e^{a_-(t-t_0)}|\mathcal{M}(t_0)\bar{v}_-(t_0)| - \int_{t_0}^{t} e^{a_-(t-s)}|B(s)\bar{v}_-(t_0)|\,ds$$

$$\ge e^{a_-(t-t_0)}|\mathcal{M}(t_0)| - \int_{t_0}^{t} e^{a_-(t-s)}|B(s)|\,ds \qquad \forall t_0 \le t.$$

□

We also use the following invariance result for a ball, which is a particular case of the Nagumo-type characterizations of invariant sets, in the setting of an input that is only measurable in time.

Lemma B.2 (Invariance of a ball under measurable inputs)**.** *Let $R > 0$, let $\mathcal{U} \subset \mathbb{R}^p$ be compact, and let $G : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^m$ be continuous, with $\frac{\partial G}{\partial z}$ continuous on $\mathbb{R}^m \times \mathbb{R}^p$. Assume that*

$$z^\top G(z,u) \le 0, \qquad \forall z \in \mathbb{R}^m \text{ with } |z| = R,\ \forall u \in \mathcal{U}. \tag{34}$$

Let $u : [0,+\infty) \to \mathbb{R}^p$ be Lebesgue measurable with $u(t) \in \mathcal{U}$ for almost all $t \ge 0$. Then the solution of $\frac{dz}{dt} = G(z,u(t))$, understood in the sense of Carathéodory, with $|z(0)| \le R$, is defined on $[0,+\infty)$ and satisfies $|z(t)| \le R$ for all $t \ge 0$.

Proof. The right-hand side $(t,z) \mapsto G(z,u(t))$ is measurable in t, continuous in z, and, on any compact set of z, bounded and Lipschitz in z uniformly in t, since $u(t)$ lies in the compact set $\mathcal{U}$ for almost all t and $\frac{\partial G}{\partial z}$ is continuous. By Carathéodory's theorem [19, Chapter I.5], there is a unique solution, defined on a maximal interval $[0, t_{\max})$, and $t_{\max} = +\infty$ if it remains in a compact set.

Step 1: strict inequality. Assume first that (34) holds with a strict inequality. By continuity of G and compactness, $\mu = -\max\{z^\top G(z,u) : |z| = R,\ u \in \mathcal{U}\}$ is positive. Moreover, by continuity of G and compactness again, there exists $\eta > 0$ such that

$$z^\top G(z,u) \le -\frac{\mu}{2}, \qquad \forall z \text{ s.t. } R \le |z| \le R+\eta,\ \forall u \in \mathcal{U}. \tag{35}$$

Assume by contradiction that $|z(t_+)| > R$ for some $t_+ \in [0, t_{\max})$, and let $t^* = \sup\{t \in [0,t_+) : |z(s)| \le R,\ \forall s \in [0,t]\}$. Then $t^* \le t_+ < t_{\max}$ and, by continuity, $|z(t^*)| = R$, and there exists $t_1 \in (t^*, t_{\max})$, maybe equal to t_+, such that $|z(t_1)| > R$ and $|z(s)| \le R+\eta$ for all $s \in [t^*, t_1]$. Let $t_0 = \max\{s \in [t^*, t_1] : |z(s)| \le R\}$, which exists since this set is closed and contains t^*. Then $|z(t_0)| = R$ and $R < |z(s)| \le R+\eta$ for all $s \in (t_0, t_1]$. The function $V(t) = |z(t)|^2$ is absolutely continuous and, by (35), $\frac{dV}{ds}(s) = 2z(s)^\top G(z(s),u(s)) \le -\mu < 0$ for almost all $s \in (t_0,t_1)$. Integrating yields $V(t_1) < V(t_0) = R^2$, which contradicts $|z(t_1)| > R$. Hence $|z(t)| \le R$ for all $t \in [0, t_{\max})$, and therefore $t_{\max} = +\infty$.

Step 2: general case. Fix $T \in (0, t_{\max})$. For $\varepsilon > 0$, let z_ε be the solution, starting from $z(0)$, of $\frac{dz_\varepsilon}{dt} = G(z_\varepsilon, u(t)) - \varepsilon z_\varepsilon$. For $|z| = R$ and $u \in \mathcal{U}$, $z^\top(G(z,u) - \varepsilon z) \le -\varepsilon R^2 < 0$, so that Step 1, applied to $(z,u) \mapsto G(z,u) - \varepsilon z$, shows that z_ε is defined on $[0,+\infty)$ and satisfies $|z_\varepsilon(t)| \le R$ for all $t \ge 0$. Let $\mathcal{K}$ be a closed ball containing the ball of radius R and the compact set $z([0,T])$, and let L be an upper bound of $|\frac{\partial G}{\partial z}|$ on the compact set $\mathcal{K} \times \mathcal{U}$. The difference $t \mapsto \Delta_\varepsilon(t) = z_\varepsilon(t) - z(t)$ is absolutely continuous, satisfies $\Delta_\varepsilon(0) = 0$ and, by the integral mean value theorem, for almost all

$t \in [0, T]$,

$$\frac{d\Delta_\varepsilon}{dt}(t) = A_\varepsilon(t)\Delta_\varepsilon(t) - \varepsilon z_\varepsilon(t), \qquad A_\varepsilon(t) = \int_0^1 \frac{\partial G}{\partial z}\left(z(t) + s\Delta_\varepsilon(t), u(t)\right) ds,$$

where the segment $\{z(t) + s\Delta_\varepsilon(t) : s \in [0,1]\}$ lies in the convex set $\mathcal{K}$ and $u(t) \in \mathcal{U}$, so that A_ε is measurable, $|A_\varepsilon(t)| \le L$ and $A_\varepsilon(t) + A_\varepsilon(t)^\top \le 2LI_m$. Lemma B.1, with $a_+ = L$ and $|\varepsilon z_\varepsilon(t)| \le \varepsilon R$, gives

$$|z_\varepsilon(t) - z(t)| \le \varepsilon R \int_0^t e^{L(t-s)}\, ds \le \varepsilon R T e^{LT}, \qquad \forall t \in [0, T].$$

Letting $\varepsilon \to 0$ and using that the ball is closed, we obtain $|z(t)| \le R$ for all $t \in [0, T]$. Since $T \in (0, t_{\max})$ is arbitrary, z remains in a compact set on $[0, t_{\max})$, which implies $t_{\max} = +\infty$. □

C Proof of Latent Contraction and Invariance (Proposition 3.3)

In this appendix, we prove Proposition 3.3, which states that the compact set $\mathcal{Z}$ is forward-invariant and that any two trajectories driven by the same observation converge exponentially towards each other.

Recall that the latent state z is governed by the dynamics:

$$\frac{dz}{dt} = \kappa g_\Omega(z, o(t))$$

where $o(t)$ takes values in the compact set $\mathcal{O}$. We define the constant $\bar{g}$ as the maximum magnitude of the unforced base vector field on $\mathcal{O}$:

$$\bar{g} = \max_{o \in \mathcal{O}} |g_\Omega(0, o)|. \tag{36}$$

The compact operating region $\mathcal{Z}$ is defined as the closed ball of radius $R_z = 2\bar{g}$, i.e., $\mathcal{Z} = \{z \in \mathbb{R}^m : |z| \le 2\bar{g}\}$.

C.1 Forward Invariance of $\mathcal{Z}$

To prove that $\mathcal{Z}$ is forward-invariant, it is sufficient to show that the vector field points inward on the boundary $\partial\mathcal{Z} = \{z \in \mathbb{R}^m : |z| = 2\bar{g}\}$. The outward normal vector at any point on the boundary is collinear with z. Thus, the vector field points inward if:

$$z^\top \frac{dz}{dt} = \kappa z^\top g_\Omega(z, o) \le 0 \qquad \forall z \in \partial\mathcal{Z}, \quad \forall o \in \mathcal{O}. \tag{37}$$

Using the fundamental theorem of calculus (integral Mean Value Theorem), we can rewrite the base vector field as:

$$\begin{aligned} z^\top g_\Omega(z, o) &= z^\top [g_\Omega(z, o) - g_\Omega(0, o)] + z^\top g_\Omega(0, o) \\ &= z^\top \left[\int_0^1 \frac{\partial g_\Omega}{\partial z}(sz, o)\, ds\right] z + z^\top g_\Omega(0, o). \end{aligned}$$

Since $\mathcal{Z}$ is convex and contains the origin, the segment sz for $s \in [0, 1]$ remains entirely in $\mathcal{Z}$. For any square matrix M, we have $z^\top M z = \frac{1}{2} z^\top (M + M^\top) z$. Applying this to the Jacobian integral and using the monotonicity condition (Assumption 3.2), which guarantees that the symmetric part of the Jacobian is bounded by $-2I_m$, we obtain:

$$\begin{aligned} z^\top \left[\int_0^1 \frac{\partial g_\Omega}{\partial z}(sz, o)\, ds\right] z &= \frac{1}{2} z^\top \left[\int_0^1 \left(\frac{\partial g_\Omega}{\partial z} + \left(\frac{\partial g_\Omega}{\partial z}\right)^\top\right)(sz, o)\, ds\right] z \\ &\le \frac{1}{2} z^\top \left[\int_0^1 -2I_m\, ds\right] z = -|z|^2. \end{aligned}$$

Using the Cauchy-Schwarz inequality ($z^\top g_\Omega(0, o) \le |z||g_\Omega(0, o)| \le |z|\bar{g}$), the inner product is bounded by:

$$z^\top g_\Omega(z, o) \le -|z|^2 + \bar{g}|z| = |z|(\bar{g} - |z|).$$

Evaluating the scaled dynamics on the boundary $\partial\mathcal{Z}$ where $|z| = 2\bar{g}$ and $\kappa > 0$, we get:

$$z^\top \frac{dz}{dt} \leq \kappa\,[2\bar{g}(\bar{g} - 2\bar{g})] = -2\kappa\bar{g}^2 \leq 0.$$

Since the inner product is non-positive on the boundary for all $o \in \mathcal{O}$, Lemma B.2, applied with $G = \kappa g_\Omega$, $\mathcal{U} = \mathcal{O}$ and $u(t) = o(t)$, guarantees that any trajectory starting in $\mathcal{Z}$ remains in it. Therefore, $\mathcal{Z}$ is forward-invariant.

C.2 Exponential Contraction of the Flow

Consider two initial conditions $z_{a0}, z_{b0} \in \mathcal{Z}$ and let $z_a(t)$ and $z_b(t)$ be the corresponding trajectories driven by the exact same observation signal $o(t) \in \mathcal{O}$.

Let $\Delta z(t) = z_a(t) - z_b(t)$. Its time derivative is given by:

$$\frac{d}{dt}\Delta z(t) = \kappa[g_\Omega(z_a(t), o(t)) - g_\Omega(z_b(t), o(t))].$$

By the integral Mean Value Theorem, this can be written as a linear time-varying system:

$$\frac{d}{dt}\Delta z(t) = A(t)\Delta z(t), \quad \text{where} \quad A(t) = \kappa \int_0^1 \frac{\partial g_\Omega}{\partial z}(s z_a(t) + (1-s) z_b(t), o(t))\, ds.$$

Because $\mathcal{Z}$ is forward-invariant and convex, the segment $s z_a(t) + (1-s) z_b(t)$ lies entirely within $\mathcal{Z}$ for all $s \in [0,1]$ and all $t \geq t_0$. Assumption 3.2 then implies that the symmetric part of the Jacobian inside the integral satisfies:

$$A(t) + A(t)^\top = \kappa \int_0^1 \left(\frac{\partial g_\Omega}{\partial z} + \left(\frac{\partial g_\Omega}{\partial z}\right)^\top \right) ds \leq \kappa \int_0^1 -2I_m\, ds = -2\kappa I_m.$$

We can now apply Lemma B.1 with $\mathcal{M} = \Delta z$, $a_+ = -\kappa$, and $B(t) = 0$. This yields directly:

$$|\Delta z(t)| \leq e^{-\kappa(t-t_0)} |\Delta z(t_0)| \qquad \forall t \geq t_0.$$

This concludes the proof of Proposition 3.3.

C.3 Corollary: Invariance under Additive Perturbations

For the subsequent proofs (notably the construction of asymptotic approximations), we highlight a crucial corollary of the invariance analysis. Consider the augmented dynamics $\frac{dz}{dt} = \kappa[g_\Omega(z, o) - v]$, where $v \in \mathbb{R}^m$ is a constant additive perturbation. The boundary condition evaluation becomes:

$$\begin{aligned} z^\top [g_\Omega(z, o) - v] &\leq -|z|^2 + |z|\,(|g_\Omega(0, o)| + |v|) \\ &\leq |z|(\bar{g} + |v| - |z|). \end{aligned}$$

If the perturbation satisfies $|v| \leq \bar{g}$, then on the boundary $|z| = 2\bar{g}$, we obtain:

$$z^\top [g_\Omega(z, o) - v] \leq 2\bar{g}(\bar{g} + \bar{g} - 2\bar{g}) = 0.$$

Multiplying by $\kappa > 0$ preserves the non-positive sign. Hence, for any fixed $o \in \mathcal{O}$ and $v \in \mathbb{R}^m$ with $|v| \leq \bar{g}$, Lemma B.2, applied with $G(z, u) = \kappa[g_\Omega(z, o) - v]$ and the constant input $u = (o, v)$, shows that the compact set $\mathcal{Z}$ remains forward-invariant even when the vector field is shifted by any constant vector v such that $|v| \leq \bar{g}$.

D Existence and Regularity of the State-to-Latent Map (Theorem 3.4)

In this appendix, we prove Theorem 3.4. The proof is structured in four parts: defining the system bounds, establishing the bounds on the physical flow and observations, proving the existence of a unique bounded solution, and finally establishing its invariance, attractivity, and regularity.

It can be observed that, from Assumption 3.1, the physical solutions initialized in the domain of interest coincide in positive time with that of the modified dynamics

$$\dot{x} = \breve{f}(x) = \chi(x) f(x), \tag{38}$$

with χ a smooth function such that

$$\chi(x) = 1 \;\text{ if }\; x \in \mathcal{X}, \qquad \chi(x) = 0 \;\text{ if }\; x \notin \breve{\mathcal{X}}$$

where $\breve{\mathcal{X}}$ is the compact inflation of $\mathcal{X}$ defined below Assumption 3.1. Besides, Assumptions 3.1 (and 3.5) still hold for the modified dynamics (38) with the output map h. In the rest of the proof, we thus assume without loss of generality that $f = \breve{f}$ and we consider that the modified dynamics ensures global (forward and backward) invariance of the extended compact set $\breve{\mathcal{X}}$, as well as completeness and boundedness of all solutions over $\mathbb{R}$.

To analyze the system in the natural time scale of the latent dynamics, we work in the observer time $\tau = \kappa t$. The physical flow initialized at x_0 is thus considered at time $t = \tau/\kappa$, and the corresponding observation fed to the network is:

$$\mathscr{o}_{x_0}(\tau) = h\left(\Phi_f\left(x_0, \frac{\tau}{\kappa}\right)\right). \tag{39}$$

When the explicit dependence on the initial condition is not required for the analysis, we simply denote the observation signal in the latent time scale as $\mathscr{o}(\tau) = o(\tau/\kappa)$. The latent dynamics in this time scale are $\frac{dz}{d\tau} = g_\Omega(z, \mathscr{o}_{x_0}(\tau))$. Let $\Phi_z(z_0, \tau_0, \tau; x_0)$ denote the flow of this latent system with z_0 being the initial condition at time τ_0.

D.1 Preamble: System and Filter Bounds

Because of Assumption 3.1, the physical trajectories remain in a compact set $\breve{\mathcal{X}}$ and the base dynamics are evaluated on the compact set $\mathcal{Z} \times \mathcal{O}$. Consequently, the following finite bounds are well-defined for the physical system:

$$L_h = \sup_{(x,v)\in\breve{\mathcal{X}}\times\mathbb{R}^n} \frac{1}{|v|}\left|\frac{\partial h}{\partial x}(x)v\right|, \qquad L_{hx} = \sup_{(x_a,x_b,v)\in\breve{\mathcal{X}}^2\times\mathbb{R}^n} \frac{\left|\left(\frac{\partial h}{\partial x}(x_a) - \frac{\partial h}{\partial x}(x_b)\right)v\right|}{|x_a - x_b||v|}, \tag{40}$$

$$L_f = \sup_{(x,v)\in\breve{\mathcal{X}}\times\mathbb{R}^n} \frac{1}{|v|}\left|\frac{\partial f}{\partial x}(x)v\right|, \qquad L_{fx} = \sup_{(x_a,x_b,v)\in\breve{\mathcal{X}}^2\times\mathbb{R}^n} \frac{\left|\left(\frac{\partial f}{\partial x}(x_a) - \frac{\partial f}{\partial x}(x_b)\right)v\right|}{|x_a - x_b||v|}. \tag{41}$$

Similarly, we define the bounds for the latent base dynamics g_Ω on $\mathcal{Z} \times \mathcal{O}$:

$$L_{go} = \sup_{(z,o)\in\mathcal{Z}\times\mathcal{O}} \left|\frac{\partial g_\Omega}{\partial o}(z,o)\right|, \qquad L_{goo} = \sup_{(z,o_a,o_b)\in\mathcal{Z}\times\mathcal{O}^2} \frac{\left|\frac{\partial g_\Omega}{\partial o}(z,o_a) - \frac{\partial g_\Omega}{\partial o}(z,o_b)\right|}{|o_a - o_b|}, \tag{42}$$

$$L_{gz} = \sup_{(z,o)\in\mathcal{Z}\times\mathcal{O}} \left|\frac{\partial g_\Omega}{\partial z}(z,o)\right|, \qquad L_{gzz} = \sup_{(z_a,z_b,o)\in\mathcal{Z}^2\times\mathcal{O}} \frac{\left|\frac{\partial g_\Omega}{\partial z}(z_a,o) - \frac{\partial g_\Omega}{\partial z}(z_b,o)\right|}{|z_a - z_b|}, \tag{43}$$

along with the cross-derivative bounds L_{goz} and L_{gzo} defined symmetrically.

D.2 Bounds on the Physical Flow and Observations

We establish the exponential growth bounds for the variations of the observation $\mathscr{o}_{x_0}(\tau)$ with respect to the initial condition x_0. The variational equation for the physical flow is:

$$\frac{\partial^2 \Phi_f}{\partial x \partial t}(x_0,t) = \frac{\partial f}{\partial x}(\Phi_f(x_0,t))\frac{\partial \Phi_f}{\partial x}(x_0,t), \qquad \frac{\partial \Phi_f}{\partial x}(x_0,0) = I_n.$$

With (40), the symmetric part of the Jacobian is bounded by $2L_f I_n$. Invoking Lemma B.1 yields:

$$\left|\frac{\partial \Phi_f}{\partial x}(x_0,t)\right| \le e^{L_f|t|} \qquad \forall (t,x_0) \in \mathbb{R}\times\breve{\mathcal{X}}. \tag{44}$$

To bound the second-order variations, let $x_a, x_b \in \breve{\mathcal{X}}$. We have:

$$\begin{aligned}\frac{\partial}{\partial t}\left[\frac{\partial \Phi_f}{\partial x}(x_a,t) - \frac{\partial \Phi_f}{\partial x}(x_b,t)\right] &= \left[\frac{\partial f}{\partial x}(\Phi_f(x_a,t)) - \frac{\partial f}{\partial x}(\Phi_f(x_b,t))\right]\frac{\partial \Phi_f}{\partial x}(x_a,t) \\ &\quad + \frac{\partial f}{\partial x}(\Phi_f(x_b,t))\left[\frac{\partial \Phi_f}{\partial x}(x_a,t) - \frac{\partial \Phi_f}{\partial x}(x_b,t)\right].\end{aligned}$$

The norm of the first term is bounded by $L_{fx}|\Phi_f(x_a,t) - \Phi_f(x_b,t)|e^{L_f|t|}$. Using the Mean Value Theorem, $|\Phi_f(x_a,t) - \Phi_f(x_b,t)| \le e^{L_f|t|}|x_a - x_b|$. Thus, the first term is bounded by

$L_{fx}e^{2L_f|t|}|x_a - x_b|$. Applying Lemma B.1 again (with initial difference 0), we obtain:

$$\left|\frac{\partial \Phi_f}{\partial x}(x_a,t) - \frac{\partial \Phi_f}{\partial x}(x_b,t)\right| \leq \int_0^{|t|} e^{L_f(|t|-|s|)} L_{fx} e^{2L_f|s|}\, ds\, |x_a - x_b| \leq \frac{L_{fx}}{L_f} e^{2L_f|t|}|x_a - x_b|. \tag{45}$$

By applying the chain rule to $\mathscr{o}_{x_0}(\tau) = h(\Phi_f(x_0, \tau/\kappa))$, we obtain the bounds on the observation derivatives. Let $\lambda = L_f/\kappa$. We get:

$$\left|\frac{\partial \mathscr{o}_{x_0}}{\partial x_0}(\tau)\right| \leq L_h e^{\lambda|\tau|} \tag{46}$$

$$\left|\frac{\partial \mathscr{o}_{x_a}}{\partial x_0}(\tau) - \frac{\partial \mathscr{o}_{x_b}}{\partial x_0}(\tau)\right| \leq L_{ox} e^{2\lambda|\tau|}|x_a - x_b| \tag{47}$$

where $L_{ox} = \left(L_{hx} + L_h \frac{L_{fx}}{L_f}\right)$. If $\kappa > 2L_f$, it implies $\lambda < \frac{1}{2}$.

D.3 Well-Definedness of the State-to-Latent Map

Let $\mathcal{F} = C_b^0(\mathbb{R}, \mathcal{Z})$ be the complete metric space of continuous bounded functions $\omega : \mathbb{R} \to \mathcal{Z}$, equipped with the uniform norm $\|\omega\|_\infty = \sup_\tau |\omega(\tau)|$. Fix an arbitrary time shift $\theta > 0$. We define the operator $\mathbb{O}_\theta : \mathcal{F} \times \breve{\mathcal{X}} \to \mathcal{F}$ by:

$$\mathbb{O}_\theta(\omega, x_0)(\tau) = \Phi_z(\omega(\tau - \theta), \tau - \theta, \tau; x_0). \tag{48}$$

By the forward invariance of $\mathcal{Z}$ (Proposition 3.3), $\mathbb{O}_\theta$ maps $\mathcal{F}$ into itself. Furthermore, the contraction property of the flow guarantees that for any $\omega_a, \omega_b \in \mathcal{F}$:

$$|\mathbb{O}_\theta(\omega_a, x_0)(\tau) - \mathbb{O}_\theta(\omega_b, x_0)(\tau)| \leq e^{-\theta}|\omega_a(\tau - \theta) - \omega_b(\tau - \theta)| \leq e^{-\theta}\|\omega_a - \omega_b\|_\infty.$$

Since $\theta > 0$, $e^{-\theta} < 1$, making $\mathbb{O}_\theta$ a strict contraction on $\mathcal{F}$ which is uniform in x_0 in $\breve{\mathcal{X}}$. By the Banach Fixed-Point Theorem, there exists a unique fixed point $\varpi_{x_0} \in \mathcal{F}$.

A priori, the abstract fixed point $\varpi_{\theta,x_0} \in C_b^0(\mathbb{R}, \mathcal{Z})$ provided by Banach's theorem depends on the chosen time shift θ, and it is not trivially a continuous-time solution of the differential equation everywhere. We now demonstrate that it is indeed the unique bounded exact solution.

Note that by iteration we get for any integer $m \in \mathbb{N}$:

$$\begin{aligned}
\varpi_{\theta,x_0}(\tau) &= \Phi_z(\varpi_{\theta,x_0}(\tau - \theta), \tau - \theta, \tau; x_0) \\
&= \Phi_z(\Phi_z(\varpi_{\theta,x_0}(\tau - 2\theta), \tau - 2\theta, \tau - \theta; x_0), \tau - \theta, \tau; x_0) \\
&= \Phi_z(\varpi_{\theta,x_0}(\tau - m\theta), \tau - m\theta, \tau; x_0).
\end{aligned}$$

Namely, ϖ_{θ,x_0} is also the fixed point of $\mathbb{O}_{m\theta}$ for all integers m.

Now, consider the fractioned sequence $\theta_k = \frac{\theta}{k}$ for $k \in \mathbb{N}^*$. By the exact same Banach theorem, there exists a unique fixed point $\varpi_{\theta_k,x_0} \in C_b^0(\mathbb{R}, \mathcal{Z})$ satisfying:

$$\varpi_{\theta_k,x_0}(\tau) = \Phi_z(\varpi_{\theta_k,x_0}(\tau - \tfrac{m}{k}\theta), \tau - \tfrac{m}{k}\theta, \tau; x_0) \qquad \forall \tau \in \mathbb{R},\ \forall m \in \mathbb{N}.$$

In particular, for $m = k$, we obtain that ϖ_{θ_k,x_0} is also a fixed point of $\mathbb{O}_\theta$. Uniqueness of the fixed point implies that $\varpi_{\theta_k,x_0} = \varpi_{\theta,x_0}$ for all $k \in \mathbb{N}^*$, meaning:

$$\varpi_{\theta,x_0}(\tau) = \Phi_z(\varpi_{\theta,x_0}(\tau - \tfrac{m}{k}\theta), \tau - \tfrac{m}{k}\theta, \tau; x_0) \qquad \forall \tau,\ \forall m \in \mathbb{N},\ \forall k \in \mathbb{N}^*. \tag{49}$$

To show that ϖ_{θ,x_0} is a true differential solution, let $\tau_0 < \tau$ be arbitrary and denote $m(k)$ the integer satisfying:

$$\tfrac{m(k)}{k}\theta \leq \tau - \tau_0 < \tfrac{m(k)+1}{k}\theta.$$

We have $\lim_{k\to\infty} \tau - \frac{m(k)}{k}\theta = \tau_0$. Because the functions $\tau \mapsto \varpi_{\theta,x_0}(\tau)$ and $(\tau_0, z_0) \mapsto \Phi_z(z_0, \tau_0, \tau; x_0)$ are continuous, passing to the limit $k \to \infty$ in (49) yields:

$$\varpi_{\theta,x_0}(\tau) = \Phi_z(\varpi_{\theta,x_0}(\tau_0), \tau_0, \tau; x_0).$$

For $\tau \leq \tau_0$, with

$$\tfrac{m(k)}{k}\theta \leq \tau_0 - \tau < \tfrac{m(k)+1}{k}\theta,$$

we have

$$\lim_{k\to\infty} \tau + \tfrac{m(k)}{k}\theta = \tau_0.$$

Also, since we have uniqueness of solutions, we can invert the flow in (49) to get:

$$\Phi_z(\varpi_{\theta,x_0}(s), s, s - \tfrac{m(k)}{k}\theta; x_0) = \varpi_{\theta,x_0}(s - \tfrac{m(k)}{k}\theta) \qquad \forall s \in \mathbb{R},\ \forall k \in \mathbb{N}^*,$$

and consequently, by setting $s = \tau + \frac{m(k)}{k}\theta$:

$$\Phi_z(\varpi_{\theta,x_0}(\tau + \tfrac{m(k)}{k}\theta), \tau + \tfrac{m(k)}{k}\theta, \tau; x_0) = \varpi_{\theta,x_0}(\tau).$$

So as above, by invoking continuity, the left hand side tends to $\Phi_z(\varpi_{\theta,x_0}(\tau_0), \tau_0, \tau; x_0)$ when $k \to \infty$, and we can conclude that:

$$\varpi_{\theta,x_0}(\tau) = \Phi_z(\varpi_{\theta,x_0}(\tau_0), \tau_0, \tau; x_0).$$

Consequently, ϖ_{θ,x_0} is an exact continuous-time trajectory of the latent dynamics everywhere. Because it is globally bounded, and we previously showed that at most one bounded solution can exist, it is strictly independent of the choice of $\theta > 0$. We denote it simply as ϖ_{x_0}.

We can thus define the state-to-latent map $\boldsymbol{T}_\kappa : \breve{\mathcal{X}} \to \mathcal{Z}$ as:

$$\boldsymbol{T}_\kappa(x_0) = \varpi_{x_0}(0). \tag{50}$$

D.4 Invariance and Attractivity

By the semi-group property of the physical system, $\Phi_f(x_0, \frac{\tau+s}{\kappa}) = \Phi_f(\Phi_f(x_0, \frac{s}{\kappa}), \frac{\tau}{\kappa})$, which implies that the input observation satisfies $o_{\Phi_f(x_0,s/\kappa)}(\tau) = o_{x_0}(\tau + s)$. Consequently, $\tau \mapsto \varpi_{x_0}(\tau + s)$ and $\tau \mapsto \varpi_{\Phi_f(x_0,s/\kappa)}(\tau)$ are both bounded solutions on $\mathbb{R}$ driven by the exact same observation sequence. By uniqueness of the bounded solution, they must be equal:

$$\varpi_{x_0}(\tau + s) = \varpi_{\Phi_f(x_0,s/\kappa)}(\tau) \qquad \forall(\tau, s).$$

Evaluating at $\tau = 0$ and setting $s = \kappa t$, we get $\boldsymbol{T}_\kappa(\Phi_f(x_0, t)) = \varpi_{x_0}(\kappa t)$. Since ϖ_{x_0} is a valid flow trajectory passing through $\varpi_{x_0}(0) = \boldsymbol{T}_\kappa(x_0)$, this proves the invariance property.

The attractivity follows directly from the contraction property of Proposition 3.3:

$$|\Phi_z(z_0, 0, \kappa t; x_0) - \boldsymbol{T}_\kappa(\Phi_f(x_0, t))| = |\Phi_z(z_0, 0, \kappa t; x_0) - \varpi_{x_0}(\kappa t)| \leq e^{-\kappa t}|z_0 - \boldsymbol{T}_\kappa(x_0)|.$$

D.5 Regularity via Ascoli-Arzelà Theorem

In this section, we restrict our attention to negative times $\tau \leq 0$. We prove that the unique bounded map ϖ is C^1 with a Lipschitz derivative on the compact set $\breve{\mathcal{X}}$. The solution ϖ_{x_0} being the limit of a sequence of bounded functions, it is sufficient to show that the sequence made of their partial derivatives is uniformly bounded and equicontinuous, allowing us to invoke the Ascoli-Arzelà theorem.

Step 1: The Iterative Sequence and its Formal Derivatives

Let $\theta > 0$ be a fixed time shift. We define the iterative sequence of functions $\omega_k : \mathbb{R} \times \breve{\mathcal{X}} \to \mathcal{Z}$ by:

$$\omega_0(\tau, x_0) \equiv 0, \qquad \omega_{k+1}(\tau, x_0) = \mathbb{O}_\theta(\omega_k, x_0)(\tau) = \Phi_z(\omega_k(\tau - \theta, x_0), \tau - \theta, \tau; x_0). \tag{51}$$

We have shown in Section D.3 that ω_k converges to ϖ uniformly in (τ, x_0) in $]-\infty, 0] \times \breve{\mathcal{X}}$.

Since the vector fields g_Ω and f, as well as the map h, are C^2 by Assumption 3.1, the flow $\Phi_z(z_0, \tau_0, \tau; x_0)$ is C^1 with respect to both z_0 and x_0. By induction, each ω_k is C^1 with respect to x_0. Differentiating the recurrence yields:

$$\frac{\partial \omega_{k+1}}{\partial x_0}(\tau, x_0) = J_z(\tau)\frac{\partial \omega_k}{\partial x_0}(\tau - \theta, x_0) + J_x(\tau), \tag{52}$$

where $J_z(\tau) = \frac{\partial \Phi_z}{\partial z_0}$ and $J_x(\tau) = \frac{\partial \Phi_z}{\partial x_0}$ are the Jacobians of the flow evaluated along the trajectory initialized at $z_0 = \omega_k(\tau - \theta, x_0)$ at time $\tau_0 = \tau - \theta$.

Step 2: First-Order Variational Equations and Bounds
To bound the sequence (52), we first bound the Jacobians. They satisfy the variational equations:

$$\frac{d}{d\tau}J_z = \frac{\partial g_\Omega}{\partial z}(\Phi_z, \mathscr{o}_{x_0})J_z, \qquad J_z(\tau_0) = I_m, \tag{53}$$

$$\frac{d}{d\tau}J_x = \frac{\partial g_\Omega}{\partial z}(\Phi_z, \mathscr{o}_{x_0})J_x + \frac{\partial g_\Omega}{\partial o}(\Phi_z, \mathscr{o}_{x_0})\frac{\partial \mathscr{o}_{x_0}}{\partial x_0}(\tau), \qquad J_x(\tau_0) = 0. \tag{54}$$

By Assumption 3.2, the symmetric part of $\frac{\partial g_\Omega}{\partial z}$ is bounded by $-2I_m$. Applying Lemma B.1 directly to (53) gives:

$$|J_z(\tau)| \leq e^{-(\tau-\tau_0)} \qquad \forall \tau \geq \tau_0. \tag{55}$$

For J_x, we use the bound on the observation variations $|\frac{\partial \mathscr{o}_{x_0}}{\partial x_0}(\tau)| \leq L_h e^{\lambda|\tau|}$ from (46), where $\lambda = L_f/\kappa$. Applying Lemma B.1 to (54) yields:

$$|J_x(\tau)| \leq \int_{\tau_0}^{\tau} e^{-(\tau-s)} L_{go} L_h e^{-\lambda s} ds \leq \frac{L_{go}L_h}{1-\lambda} e^{\lambda|\tau|} \qquad \forall \tau_0 \leq \tau \leq 0. \tag{56}$$

Step 3: Uniform Boundedness of the Sequence of Derivatives
Substituting the bounds (55) and (56) into the norm of the recursion (52) (with $\tau_0 = \tau - \theta$) and multiplying by $e^{-\lambda|\tau|}$ gives:

$$\left|\frac{\partial \omega_{k+1}}{\partial x_0}(\tau, x_0)\right| e^{-\lambda|\tau|} \leq e^{-\theta}\left|\frac{\partial \omega_k}{\partial x_0}(\tau-\theta, x_0)\right| e^{-\lambda|\tau|} + \frac{L_{go}L_h}{1-\lambda}.$$

and therefore, with $|\tau| \geq |\tau-\theta| - \theta$.

$$\sup_{\tau\leq 0, x_0 \in \breve{\mathfrak{X}}}\left|\frac{\partial \omega_k}{\partial x_0}(\tau, x_0)\right| e^{-\lambda|\tau|} \leq e^{-(1-\lambda)\theta} \sup_{\tau\leq 0, x_0 \in \breve{\mathfrak{X}}}\left|\frac{\partial \omega_k}{\partial x_0}(\tau-\theta, x_0)\right| e^{-\lambda|\tau-\theta|} + \frac{L_{go}L_h}{1-\lambda}.$$

Because $\omega_0 \equiv 0$, we get by induction that the sequence is uniformly bounded for all $k \geq 0$:

$$\sup_{\tau\leq 0, x_0 \in \breve{\mathfrak{X}}}\left|\frac{\partial \omega_k}{\partial x_0}(\tau, x_0)\right| e^{-\lambda|\tau|} \leq \frac{L_{go}L_h}{(1-\lambda)(1-e^{-(1-\lambda)\theta})} := C_{\omega'} \tag{57}$$

where $\lambda = L_f/\kappa < \frac{1}{2}$. This establishes the sequence $\left\{\frac{\partial \omega_k}{\partial x_0}\right\}$ is uniformly bounded in $[-q, 0] \times \breve{\mathfrak{X}}$, for any $q > 0$.

Step 4: Second-Order Variations and Lipschitz Bounds
To apply Ascoli-Arzelà, we need to guarantee the equicontinuity of the derivatives, which can be done by bounding the second-order variations. Consider two parameters $x_{a0}, x_{b0} \in \breve{\mathfrak{X}}$ and define $\Delta J_z(\tau) = J_{za}(\tau) - J_{zb}(\tau)$. Subtracting the variational equations (53) evaluated along the two trajectories yields:

$$\frac{d}{d\tau}\Delta J_z = \left.\frac{\partial g_\Omega}{\partial z}\right|_a \Delta J_z + \underbrace{\left(\left.\frac{\partial g_\Omega}{\partial z}\right|_a - \left.\frac{\partial g_\Omega}{\partial z}\right|_b\right)}_{:=F_z(\tau)} J_{zb}\,.$$

We have $|F_z(\tau)| \leq \big(L_{gzz}|\Delta z(\tau)| + L_{gzo}|\Delta \mathscr{o}(\tau)|\big)|J_{zb}(\tau)|$. And with (55) and $|\Delta \mathscr{o}(\tau)| \leq L_h e^{\lambda|\tau|}|\Delta x_0|$, alongside the explicit latent flow variation bound $|\Delta z(\tau)| \leq e^{-(\tau-\tau_0)}|\Delta z_0| + C_{\Phi x}e^{\lambda|\tau|}|\Delta x_0|$, we obtain:

$$\begin{aligned}|F_z(\tau)| &\leq \left(L_{gzz}\left[e^{-(\tau-\tau_0)}|\Delta z_0| + C_{\Phi x}e^{\lambda|\tau|}|\Delta x_0|\right] + L_{gzo}L_h e^{\lambda|\tau|}|\Delta x_0|\right) e^{-(\tau-\tau_0)} \\ &= L_{gzz}e^{-2(\tau-\tau_0)}|\Delta z_0| + \lambda \wp_{zx} e^{\lambda|\tau|}e^{-(\tau-\tau_0)}|\Delta x_0|,\end{aligned}$$

where we define the structural constant $\wp_{zx} = \frac{1}{\lambda}(L_{gzz}C_{\Phi x} + L_{gzo}L_h)$. Applying Lemma B.1 to the differential equation governing ΔJ_z with the initial condition $\Delta J_z(\tau_0) = 0$ requires evaluating the convolution integral $\int_{\tau_0}^{\tau} e^{-(\tau-s)}|F_z(s)|\,ds$. Because the integration domain is strictly non-positive

($\tau_0 \le s \le \tau \le 0$), we substitute $|s| = -s$. The exact integrations therefore yield:

$$\int_{\tau_0}^{\tau} e^{-(\tau-s)} e^{-2(s-\tau_0)} ds \le e^{-(\tau-\tau_0)},$$

$$\int_{\tau_0}^{\tau} e^{-(\tau-s)} e^{-\lambda s} e^{-(s-\tau_0)} ds = \frac{1}{\lambda} e^{-(\tau-\tau_0)} \left(e^{\lambda|\tau_0|} - e^{\lambda|\tau|}\right).$$

This directly provides the structural Lipschitz bound for the state Jacobian:

$$|\Delta J_z(\tau)| \le L_{gzz} e^{-(\tau-\tau_0)} |z_{a0} - z_{b0}| + \wp_{zx} e^{-(\tau-\tau_0)} \left(e^{\lambda|\tau_0|} - e^{\lambda|\tau|}\right) |x_{a0} - x_{b0}|. \tag{58}$$

A similar procedure is applied to the parameter Jacobian difference $\Delta J_x(\tau)$. Subtracting the equations (54) reveals a forcing term $F_x(\tau)$ containing four distinct differences. Bounding them using the known structural constants gives:

$$\begin{aligned} |F_x(\tau)| &\le \left(L_{gzz}|\Delta z(\tau)| + L_{gzo}|\Delta o(\tau)|\right) |J_{xb}(\tau)| + L_{go} \left|\Delta \frac{\partial o}{\partial x_0}(\tau)\right| \\ &\quad + \left(L_{gzo}|\Delta z(\tau)| + L_{goo}|\Delta o(\tau)|\right) \left|\frac{\partial o}{\partial x_0}\right|_b \\ &\le \lambda \wp_{zx} e^{\lambda|\tau|} e^{-(\tau-\tau_0)} |\Delta z_0| + \wp_{xx} e^{2\lambda|\tau|} |\Delta x_0|, \end{aligned}$$

where $\wp_{xx} = L_{gzz} C_{\Phi x}^2 + 2 L_{gzo} C_{\Phi x} L_h + L_{go} L_{ox} + L_{goo} L_h^2$. Integrating this via Lemma B.1 requires bounding the cross-term integral:

$$\int_{\tau_0}^{\tau} e^{-(\tau-s)} e^{-\lambda s} e^{-(s-\tau_0)} ds = \frac{1}{\lambda} e^{-(\tau-\tau_0)} e^{\lambda|\tau_0|} - \frac{1}{\lambda} e^{-(\tau-\tau_0)} e^{\lambda|\tau|} \le \frac{1}{\lambda} e^{-(1-\lambda)(\tau-\tau_0)} e^{\lambda|\tau|},$$

where we used the identity $\lambda|\tau_0| = \lambda|\tau| + \lambda(\tau - \tau_0)$ derived from the fact that $s \le \tau \le 0$. The second term integrates as $\int_{\tau_0}^{\tau} e^{-(\tau-s)} e^{-2\lambda s} ds \le \frac{1}{1-2\lambda} e^{2\lambda|\tau|}$. Combining these bounds strictly yields:

$$|\Delta J_x(\tau)| \le \wp_{zx} e^{-(1-\lambda)(\tau-\tau_0)} e^{\lambda|\tau|} |z_{a0} - z_{b0}| + \frac{\wp_{xx}}{1-2\lambda} e^{2\lambda|\tau|} |x_{a0} - x_{b0}|. \tag{59}$$

Note that the integration bounding the parameter sensitivity requires the exponent $1 - 2\lambda$ to be strictly positive, which natively enforces the required high-gain condition $\kappa > 2L_f$.

Step 5: Equicontinuity

By injecting the Lipschitz bounds (58) and (59) back into the difference of the recursive equation (52), multiplying by $e^{-2\lambda|\tau|}$ and utilizing the uniform bound $C_{\omega'}$ obtained in (57), algebraic recombination leads to:

$$\left|\Delta \frac{\partial \omega_{k+1}}{\partial x_0}(\tau)\right| e^{-2\lambda|\tau|} \le e^{-\theta} \left|\Delta \frac{\partial \omega_k}{\partial x_0}(\tau - \theta)\right| e^{-2\lambda|\tau|} + \wp_p |x_{a0} - x_{b0}|,$$

where $\wp_p$ is a positive constant grouping the previous structural terms. With the same arguments as those used in Step 3, we conclude, for all $k \ge 0$,

$$\sup_{\tau \le 0, (x_{a0}, x_{b0}) \in \breve{\mathfrak{X}}} \left|\Delta \frac{\partial \omega_{k+1}}{\partial x_0}\right| e^{-2\lambda|\tau|} \le \frac{\wp_p}{1 - e^{-(1-2\lambda)\theta}} |x_{a0} - x_{b0}| \tag{60}$$

This establishes the sequence $\left\{\frac{\partial \omega_k}{\partial x_0}\right\}$ is Lipschitz in $[-q, 0] \times \breve{\mathfrak{X}}$, uniformly in k, for any $q > 0$. This implies this sequence is equicontinuous on $[-q, 0] \times \breve{\mathfrak{X}}$, for any $q > 0$.

Step 6: Convergence

For $q = 1$, Arzela-Ascoli theorem guarantees the existence of a subsequence $\{k_{1,j}\}$ such that $\left\{\frac{\partial \omega_{k_{1,j}}}{\partial x_0}\right\}$ converges uniformly on $[-1, 0] \times \breve{\mathfrak{X}}$ to a Lipschitz limit, say $D\varpi_1$. For $q = 2$ we obtain a subsequence $\{k_{2,j}\}$ of the previous subsequence such that $\left\{\frac{\partial \omega_{k_{2,j}}}{\partial x_0}\right\}$ converges uniformly on $[-2, 0] \times \breve{\mathfrak{X}}$ to a Lipschitz limit, say $D\varpi_2$ which coincides with $D\varpi_1$ on $[-1, 0]$. We go on recursively for all integer q

We remark that, since $\breve{\mathfrak{X}}$ is convex, the Mean Value Theorem gives, for any x_0 in $\breve{\mathfrak{X}}$, any unit vector v and any non zero δ such that $x_0 + \delta v$ is also in $\breve{\mathfrak{X}}$,

$$\frac{\left|\omega_k(\tau, x_0 + \delta v) - \omega_k(\tau, x_0) - \delta \frac{\partial \omega_k}{\partial x_0}(\tau, x_0) v\right|}{\delta} = \left| \int_0^1 \left[\frac{\partial \omega_k}{\partial x_0}(\tau, x_0 + s\delta v) - \frac{\partial \omega_k}{\partial x_0}(\tau, x_0) \right] v ds \right| \leq \frac{\wp_p}{1 - e^{-(1-2\lambda)\theta}} e^{2\lambda|\tau|} \frac{\delta}{2}. \tag{61}$$

The convergence of subsequences $\{\omega_{k_{q,j}}\}$ and $\left\{\frac{\partial \omega_{k_{q,j}}}{\partial x_0}\right\}$ to ϖ and $D\varpi_q$, respectively, being uniform on $[-q, 0] \times \breve{\mathfrak{X}}$, by taking the limit in (61) when first $k_{q,j}$ goes to infinity and then when δ goes to 0, we get that the limit of the derivatives $D\varpi_q$ is the derivative $\frac{\partial \varpi}{\partial x_0}$ of the limit ϖ.

This establishes that, for each $\tau \leq 0$, the function $x_0 \mapsto \varpi(\tau, x_0)$ is C^1 with Lipschitz derivative in $\breve{\mathfrak{X}}$.

E Injectivity of the State-to-Latent Map (Theorem 3.7)

This appendix is dedicated to the proof of Theorem 3.7. The proof relies on an asymptotic analysis of the contracting latent dynamics when driven by slow-varying inputs. We construct a sequence of functions that approximates the exact unique bounded solution of the filter, and we show that this approximation is injective with respect to the input and its time derivatives.

E.1 Preamble and Definitions

Restrictions and Bounds. We assume the base dynamics $g_\Omega : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^m$ satisfy the regularity of Assumption 3.1 and the monotonicity property of Assumption 3.2. Consequently, all the bounds defined in (42) are available on the compact set $\mathcal{Z} \times \mathcal{O}$. We denote $\bar{o} = \max_{o \in \mathcal{O}} |o|$.

Input Derivatives and Dummy Variables. We are interested in the first ℓ derivatives with respect to the observer time τ of the input $\mathscr{o}(\tau)$, with ℓ fixed by the observability index (Assumption 3.5). We introduce functions whose arguments will ultimately be the time sequence:

$$\tau \mapsto \left(\mathscr{o}(\tau), \frac{d\mathscr{o}}{d\tau}(\tau), \ldots, \frac{d^i \mathscr{o}}{d\tau^i}(\tau) \right).$$

To establish properties independent of specific time trajectories, we introduce dummy variables $\eta_i \in \mathbb{R}^p$ acting as generic arguments. We use the notation compact block notations:

$$\boldsymbol{\mathscr{o}}_i(\tau) = \left(\frac{d\mathscr{o}}{d\tau}(\tau), \ldots, \frac{d^i \mathscr{o}}{d\tau^i}(\tau) \right)^\top, \qquad \boldsymbol{\eta}_i = (\eta_1, \ldots, \eta_i)^\top. \tag{62}$$

We define a weighted magnitude function ρ to scale the successive time derivatives appropriately:

$$\rho(\boldsymbol{\eta}_{\ell+1}) = \max \left\{ |\eta_1|, |\eta_2|^{1/2}, \ldots, |\eta_{\ell+1}|^{1/(\ell+1)} \right\}. \tag{63}$$

For any $\varepsilon > 0$, we define the compact domain $\mathfrak{Y}(\varepsilon)$ as:

$$\mathfrak{Y}(\varepsilon) = \left\{ (\eta_0, \boldsymbol{\eta}_{\ell+1}) : |\eta_0| \leq \bar{o}, \rho(\boldsymbol{\eta}_{\ell+1}) \leq \varepsilon \right\}. \tag{64}$$

An input function $\tau \mapsto \mathscr{o}(\tau)$ is said to be *admissible* if there exists a sufficiently small $\varepsilon_{ad} > 0$ such that:

$$(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_{\ell+1}(\tau)) \in \mathfrak{Y}(\varepsilon_{ad}) \qquad \forall \tau \in \mathbb{R}. \tag{65}$$

The Operator $\mathring{-}$. To manipulate time derivatives algebraically, we introduce the operator $\mathring{-}$. It maps a continuously differentiable function $(\eta_0, \boldsymbol{\eta}_i) \mapsto p(\eta_0, \boldsymbol{\eta}_i)$ to a function $(\eta_0, \boldsymbol{\eta}_{i+1}) \mapsto \mathring{p}(\eta_0, \boldsymbol{\eta}_{i+1})$ defined as:

$$\mathring{p}(\eta_0, \boldsymbol{\eta}_{i+1}) = \sum_{k=0}^{i} \frac{\partial p}{\partial \eta_k}(\eta_0, \boldsymbol{\eta}_i) \eta_{k+1}. \tag{66}$$

By construction, $\mathring{\eta}_i = \eta_{i+1}$. Evaluated along a valid time trajectory where $\eta_i = \boldsymbol{o}_i(\tau) = \frac{d^i \boldsymbol{o}}{d\tau^i}(\tau)$, this operator provides the exact time derivative:

$$\mathring{p}(\boldsymbol{o}(\tau), \boldsymbol{o}_{i+1}(\tau)) = \frac{d}{d\tau}\left\{p(\boldsymbol{o}(\tau), \boldsymbol{o}_i(\tau))\right\}. \tag{67}$$

The Implicit Function ψ. We need to invert the base dynamics with respect to the latent state when subject to a constant additive forcing v. We define the function ψ implicitly as the solution to:

$$g_\Omega(\psi(v, \eta_0), \eta_0) = v. \tag{68}$$

Lemma E.1. *Under Assumptions 3.1 and 3.2, there exists a unique function $\psi : \{(v, \eta_0) : |v| \leq \bar{g}, |\eta_0| \leq \bar{o}\} \to \mathcal{Z}$ which solves (68). Furthermore, ψ inherits the smoothness of g_Ω.*

Proof. Consider the auxiliary dynamics $\frac{d\xi}{d\tau} = g_\Omega(\xi, \eta_0) - v$. As demonstrated in Appendix C.3, for any fixed $v \in \mathbb{R}^m$ satisfying $|v| \leq \bar{g}$, the compact convex set $\mathcal{Z}$ remains forward-invariant. Therefore, there exists an equilibrium point $\xi_{eq} \in \mathcal{Z}$ (see e.g., [19, Theorem I.8.2]). This equilibrium satisfies $g_\Omega(\xi_{eq}, \eta_0) = v$. Due to the contraction property inside $\mathcal{Z}$ (Assumption 3.2), this equilibrium is unique. We thus well-define the map $(v, \eta_0) \mapsto \psi(v, \eta_0) := \xi_{eq}$. Its smoothness follows directly from the Implicit Function Theorem, since g_Ω is smooth and its Jacobian $\frac{\partial g_\Omega}{\partial z}$ is negative definite (hence invertible) everywhere in $\mathcal{Z} \times \mathcal{O}$. □

E.2 Definition of the Approximation Sequence

As established previously, the state-to-latent map $\boldsymbol{T}_\kappa$ defined in (50) relies on the unique bounded latent trajectory. To lighten the notation in the following approximation analysis, we drop the explicit dependence on the initial physical state x_0 and simply denote this trajectory as $\varpi(\tau)$ when driven by a given observation signal. However, because $\varpi(\tau)$ is the solution to a differential equation, its injectivity with respect to the physical state cannot be analyzed directly. To overcome this, we construct an algebraic approximation of $\varpi(\tau)$. Because the contracting latent dynamics heavily filter the input, the latent state relies primarily on the instantaneous observation $\boldsymbol{o}(\tau)$ and its recent history. We thus approximate the exact state $\varpi(\tau)$ using a sequence of static maps φ_i that depend explicitly on the observation and its first i time derivatives.

We define this sequence of approximation functions φ_i via the following recursion:

$$\begin{aligned} \varphi_{-1} &= 0, \\ \varphi_0(\eta_0) &= \psi(0, \eta_0), \\ \varphi_i(\eta_0, \boldsymbol{\eta}_i) &= \psi(\mathring{\varphi}_{i-1}(\eta_0, \boldsymbol{\eta}_i), \eta_0) \qquad \forall i \in \{1, \dots, \ell+1\}. \end{aligned} \tag{69}$$

Two direct consequences follow from this definition:

1. Since the definition (66) of the operator $\mathring{}$ gives readily $\mathring{\varphi}_{i-1}(\eta_0, 0, \dots, 0) = 0$, we have:
$$\varphi_i(\eta_0, 0, \dots, 0) = \psi(0, \eta_0) = \varphi_0(\eta_0). \tag{70}$$
2. The implicit relation (68) of ψ directly yields the fundamental identity:
$$\mathring{\varphi}_{i-1}(\eta_0, \boldsymbol{\eta}_i) = g_\Omega(\varphi_i(\eta_0, \boldsymbol{\eta}_i), \eta_0). \tag{71}$$

Lemma E.2 (Well-Definedness and Bounds on the Approximations)**.** *Suppose Assumptions 3.1 and 3.2 are satisfied. There exists an admissibility threshold $\varepsilon_{ad} > 0$ such that the functions φ_i are well-defined on the compact domain $\mathfrak{Y}(\varepsilon_{ad})$ and take values in $\mathcal{Z}$. Furthermore, if g_Ω is C^q (with $q \geq \ell + 1$), then φ_i is C^{q-i}, and there exist constants $\bar{\phi}$ and $\overline{\partial\phi}$ such that for all $i \in \{0, \dots, \ell\}$, and all $(\eta_0, \boldsymbol{\eta}_{\ell+1}) \in \mathfrak{Y}(\varepsilon_{ad})$:*

$$|\varphi_i(\eta_0, \boldsymbol{\eta}_i)| \leq 2\bar{g} \tag{72}$$

$$|\varphi_{i+1}(\eta_0, \boldsymbol{\eta}_{i+1}) - \varphi_i(\eta_0, \boldsymbol{\eta}_i)| \leq \rho(\boldsymbol{\eta}_{\ell+1})^{i+1}\bar{\phi} \tag{73}$$

Moreover, for any two points $(\eta_{a0}, \boldsymbol{\eta}_{a(\ell+1)})$ and $(\eta_{b0}, \boldsymbol{\eta}_{b(\ell+1)})$ in $\mathfrak{Y}(\varepsilon_{ad})$:

$$\begin{aligned} &\Big|[\varphi_{i+1}(\eta_{a0}, \boldsymbol{\eta}_{a(i+1)}) - \varphi_i(\eta_{a0}, \boldsymbol{\eta}_{ai})] - [\varphi_{i+1}(\eta_{b0}, \boldsymbol{\eta}_{b(i+1)}) - \varphi_i(\eta_{b0}, \boldsymbol{\eta}_{bi})]\Big| \\ &\leq \overline{\partial\phi} \sum_{k=0}^{i+1} \max\left\{\rho(\boldsymbol{\eta}_{a(\ell+1)}), \rho(\boldsymbol{\eta}_{b(\ell+1)})\right\}^{i+1-k} |\eta_{ak} - \eta_{bk}|. \end{aligned} \tag{74}$$

Proof. We proceed in three steps: well-definedness by induction, bounding the approximation differences via homogeneity scaling, and finally establishing the local Lipschitz bound.

1. Well-definedness (Induction on i). *Base case ($i = 0$):* By Lemma E.1, the function $\eta_0 \mapsto \varphi_0(\eta_0) = \psi(0, \eta_0)$ solving $g_\Omega(\varphi_0(\eta_0), \eta_0) = 0$ is well-defined, unique, takes values in $\mathcal{Z}$, and is C^q. We set $\varepsilon_0 = 1$. The bound (72) is trivially satisfied since $\mathcal{Z}$ is a ball of radius $2\bar{g}$.

Inductive step: Assume that, for some $i \leq \ell$, there exists $\varepsilon_i \in (0, 1]$ such that for all $\varepsilon \leq \varepsilon_i$, the functions $\varphi_0, \dots, \varphi_i$ are well-defined on the compact set $\mathfrak{Y}(\varepsilon_i)$, take values in $\mathcal{Z}$, and are C^{q-j} respectively. By this inductive hypothesis, the partial derivatives of φ_i are continuous and bounded on $\mathfrak{Y}(\varepsilon_i)$. Let us define the constant:

$$L_{\varphi_i} := (i+1) \max_{k \in \{0,\dots,i\}} \left(\sup_{(\eta_0, \boldsymbol{\eta}_i) \in \mathfrak{Y}(\varepsilon_i)} \left| \frac{\partial \varphi_i}{\partial \eta_k}(\eta_0, \boldsymbol{\eta}_i) \right| \right).$$

Recalling the definition (66) of the operator $\mathring{-}$, we have:

$$|\mathring{\varphi}_i(\eta_0, \boldsymbol{\eta}_{i+1})| = \left| \sum_{k=0}^{i} \frac{\partial \varphi_i}{\partial \eta_k}(\eta_0, \boldsymbol{\eta}_i) \eta_{k+1} \right| \leq \frac{1}{i+1} L_{\varphi_i} \sum_{k=0}^{i} |\eta_{k+1}|.$$

For any $(\eta_0, \boldsymbol{\eta}_{i+1}) \in \mathfrak{Y}(\varepsilon_i)$ with $\varepsilon_i \leq 1$, the definition of ρ ensures $|\eta_{k+1}| \leq \rho(\boldsymbol{\eta}_{i+1})^{k+1} \leq \rho(\boldsymbol{\eta}_{i+1})$. This yields:

$$|\mathring{\varphi}_i(\eta_0, \boldsymbol{\eta}_{i+1})| \leq L_{\varphi_i} \rho(\boldsymbol{\eta}_{i+1}). \tag{75}$$

The subsequent function is defined as $\varphi_{i+1}(\eta_0, \boldsymbol{\eta}_{i+1}) = \psi(\mathring{\varphi}_i(\eta_0, \boldsymbol{\eta}_{i+1}), \eta_0)$. Lemma E.1 guarantees this is well-defined and evaluates inside $\mathcal{Z}$ provided that the forcing magnitude $|\mathring{\varphi}_i| \leq \bar{g}$. By choosing the refined admissibility threshold:

$$\varepsilon_{i+1} = \min\left(\varepsilon_i, \frac{\bar{g}}{L_{\varphi_i}} \right),$$

we ensure that $\rho(\boldsymbol{\eta}_{i+1}) \leq \varepsilon_{i+1} \implies |\mathring{\varphi}_i| \leq \bar{g}$. Thus φ_{i+1} is well-defined and $C^{q-(i+1)}$. Setting $\varepsilon_{ad} = \varepsilon_{\ell+1}$ concludes the induction, proving (72) up to $\ell + 1$.

2. Scaling and Difference Bound. To prove (73), we apply the Hadamard Lemma to ψ between the two forcings:

$$\begin{aligned} \varphi_{i+1}(\eta_0, \boldsymbol{\eta}_{i+1}) - \varphi_i(\eta_0, \boldsymbol{\eta}_i) &= \psi(\mathring{\varphi}_i(\eta_0, \boldsymbol{\eta}_{i+1}), \eta_0) - \psi(\mathring{\varphi}_{i-1}(\eta_0, \boldsymbol{\eta}_i), \eta_0) \\ &= K_i(\eta_0, \boldsymbol{\eta}_{i+1}) \left[\mathring{\varphi}_i(\eta_0, \boldsymbol{\eta}_{i+1}) - \mathring{\varphi}_{i-1}(\eta_0, \boldsymbol{\eta}_i) \right], \end{aligned}$$

where $K_i(\eta_0, \boldsymbol{\eta}_{i+1}) = \int_0^1 \frac{\partial \psi}{\partial v}(s\mathring{\varphi}_i + (1-s)\mathring{\varphi}_{i-1}, \eta_0)\, ds$ is C^{q-i-1} and bounded on $\mathfrak{Y}(\varepsilon_{ad})$.

Let $r = \rho(\boldsymbol{\eta}_{\ell+1}) \leq \varepsilon_{ad}$. We define the scaled variables $\xi_k = \frac{\eta_k}{r^k}$, denoting $\boldsymbol{\xi}_i = (\xi_1, \dots, \xi_i)^\top$, and $\xi_0 = \eta_0$. The original variables satisfy $\eta_k = r^k \xi_k$. For $i = 0$, since $\mathring{\varphi}_{-1} = 0$, we have $\mathring{\varphi}_0(\xi_0, r\xi_1) = \frac{\partial \psi}{\partial o}(0, \xi_0) r\xi_1 := r\check{p}_0(\xi_0, \xi_1)$ where $\check{p}_0$ is C^{q-1}. Thus:

$$\varphi_1(\xi_0, r\xi_1) - \varphi_0(\xi_0) = r K_0(\xi_0, r\xi_1) \check{p}_0(\xi_0, \xi_1) := r\tilde{p}_1(r, \xi_0, \xi_1).$$

By induction, if a function p can be factorized as $p(\xi_0, r\xi_1, \dots, r^i \xi_i) = r^i \tilde{p}(r, \xi_0, \dots, \xi_i)$, then applying the $\mathring{-}$ operator yields:

$$\begin{aligned} \mathring{p}(\xi_0, r\xi_1, \dots, r^{i+1}\xi_{i+1}) &= \sum_{k=0}^{i} \frac{\partial p}{\partial \eta_k}(\xi_0, r\xi_1, \dots, r^i \xi_i) r^{k+1} \xi_{k+1} \\ &= r^i \sum_{k=0}^{i} \frac{\partial \tilde{p}}{\partial \xi_k}(r, \xi_0, \dots, \xi_i) \frac{\partial \xi_k}{\partial \eta_k} r^{k+1} \xi_{k+1} \\ &= r^{i+1} \sum_{k=0}^{i} \frac{\partial \tilde{p}}{\partial \xi_k}(r, \xi_0, \dots, \xi_i) \xi_{k+1} := r^{i+1} \check{p}(r, \xi_0, \boldsymbol{\xi}_{i+1}). \end{aligned} \tag{76}$$

Following this recursive factorization, there exist C^{q-i-1} functions $\tilde{p}_{i+1}$ such that:

$$\varphi_{i+1}(\xi_0, r\xi_1, \dots, r^{i+1}\xi_{i+1}) - \varphi_i(\xi_0, r\xi_1, \dots, r^i \xi_i) = r^{i+1} \tilde{p}_{i+1}(r, \xi_0, \boldsymbol{\xi}_{i+1}).$$

Whenever $(\eta_0, \boldsymbol{\eta}_{\ell+1}) \in \mathfrak{Y}(\varepsilon_{ad})$, the scaled vector $(\xi_0, \boldsymbol{\xi}_{i+1})$ belongs to the compact set $\{(\xi_0, \boldsymbol{\xi}_{i+1}) : |\xi_0| \leq \bar{o}, \rho(\boldsymbol{\xi}_{\ell+1}) = 1\}$. By continuity, $\tilde{p}_{i+1}$ is bounded on this compact domain by a constant $\bar{\phi}$. Replacing r by $\rho(\boldsymbol{\eta}_{\ell+1})$ yields (73).

3. Lipschitz of the Approximation Differences. To establish (74), we apply the Hadamard Lemma to the factorized function $\tilde{p}_{i+1}$ with respect to the scaled variables ξ_k:

$$\tilde{p}_{i+1}(r, \xi_{a0}, \boldsymbol{\xi}_{a(i+1)}) - \tilde{p}_{i+1}(r, \xi_{b0}, \boldsymbol{\xi}_{b(i+1)}) = \sum_{k=0}^{i+1} \pi_k(r, \xi_{a0}, \boldsymbol{\xi}_{a(i+1)}, \xi_{b0}, \boldsymbol{\xi}_{b(i+1)})(\xi_{ak} - \xi_{bk}),$$

where π_k are continuous functions. Multiplying both sides by r^{i+1} gives the difference in the original unscaled functions $p = \varphi_{i+1} - \varphi_i$:

$$\begin{aligned} p(\eta_{a0}, \boldsymbol{\eta}_{a(i+1)}) - p(\eta_{b0}, \boldsymbol{\eta}_{b(i+1)}) &= r^{i+1} \sum_{k=0}^{i+1} \pi_k(\dots)(\xi_{ak} - \xi_{bk}) \\ &= \sum_{k=0}^{i+1} r^{i+1-k} \pi_k(\dots)(r^k \xi_{ak} - r^k \xi_{bk}) \\ &= \sum_{k=0}^{i+1} \pi_k(\dots) r^{i+1-k} (\eta_{ak} - \eta_{bk}). \end{aligned}$$

Let $r = \max\{\rho(\boldsymbol{\eta}_{a(\ell+1)}), \rho(\boldsymbol{\eta}_{b(\ell+1)})\}$. Since $r \leq \varepsilon_{ad}$, the scaled arguments lie in a compact set, meaning there exists a uniform upper bound $\overline{\partial\phi}$ for all functions $|\pi_k(\dots)|$. Factoring out this bound immediately yields (74). □

E.3 Lipschitz Injectivity of the Approximation Map

The next crucial step is to prove that the approximation map φ_ℓ is injective on $\mathfrak{Y}(\varepsilon)$ when ε is sufficiently small. This ensures that the time-derivatives of the input observation can be uniquely recovered from the latent state approximation.

Proposition E.3 (Lipschitz Injectivity of the Approximation)**.** *Suppose Assumptions 3.1, 3.2, and 3.6 are satisfied. Then, there exist real numbers $\varepsilon_{inj} \in (0, \varepsilon_{ad}]$ and $L_{inj} > 0$ such that, for any two pairs $(\eta_{a0}, \boldsymbol{\eta}_{a\ell})$ and $(\eta_{b0}, \boldsymbol{\eta}_{b\ell})$ in the compact set $\mathfrak{Y}(\varepsilon_{inj})$, we have:*

$$|\eta_{a0} - \eta_{b0}| + |\boldsymbol{\eta}_{a\ell} - \boldsymbol{\eta}_{b\ell}| \leq L_{inj} |\varphi_\ell(\eta_{a0}, \boldsymbol{\eta}_{a\ell}) - \varphi_\ell(\eta_{b0}, \boldsymbol{\eta}_{b\ell})|. \tag{77}$$

Proof. The proof is organized into four main steps. First, we study the injectivity of the base map φ_0. Second, we derive a recursive formula for the Jacobian of φ_i, leading to an explicit expression for $J_\ell(\eta_0, 0)$ and establishing its left-invertibility. Third, we prove the injectivity of φ_ℓ on $\mathfrak{Y}(\varepsilon_{inj})$ via a contradiction argument. Finally, we deduce the uniform Lipschitz injectivity bound.

Let ε_{ad} be the admissibility threshold given by Lemma E.2, guaranteeing that the sequence of approximations φ_i is well-defined, smooth, and takes values in $\mathcal{Z}$. Let J_i denote the Jacobian matrix of φ_i with respect to all its arguments:

$$J_i(\eta_0, \boldsymbol{\eta}_i) = \begin{bmatrix} \frac{\partial \varphi_i}{\partial \eta_0} & \frac{\partial \varphi_i}{\partial \eta_1} & \dots & \frac{\partial \varphi_i}{\partial \eta_i} \end{bmatrix}.$$

We are particularly interested in evaluating J_ℓ on the zero-derivative manifold, i.e., $J_\ell(\eta_0, 0)$. Let $A(\eta_0) = \frac{\partial g_\Omega}{\partial z}(\varphi_0(\eta_0), \eta_0)$ and $B(\eta_0) = \frac{\partial g_\Omega}{\partial o}(\varphi_0(\eta_0), \eta_0)$. By the monotonicity property (Assumption 3.2), $A(\eta_0) + A(\eta_0)^\top$ is strictly negative definite and therefore $A(\eta_0)$ is invertible.

1. Study of the base map φ_0. The map $\varphi_0(\eta_0)$ is defined by the implicit relation $g_\Omega(\varphi_0(\eta_0), \eta_0) = 0$. Assume there exist $\eta_{a0}, \eta_{b0} \in \mathcal{O}$ such that $\varphi_0(\eta_{a0}) = \varphi_0(\eta_{b0}) := z^*$. This implies $g_\Omega(z^*, \eta_{a0}) = g_\Omega(z^*, \eta_{b0})$. By the Observation Injectivity property assumed in the Filter Rank Condition (Assumption 3.6), this requires $\eta_{a0} = \eta_{b0}$. Hence, φ_0 is injective. Moreover, differentiating the implicit relation using the chain rule yields:

$$\frac{\partial g_\Omega}{\partial z}(\varphi_0(\eta_0), \eta_0) \frac{\partial \varphi_0}{\partial \eta_0}(\eta_0) + \frac{\partial g_\Omega}{\partial o}(\varphi_0(\eta_0), \eta_0) = 0 \implies A(\eta_0) J_0(\eta_0) + B(\eta_0) = 0.$$

This explicitly gives $J_0(\eta_0) = -A(\eta_0)^{-1} B(\eta_0)$.

2. Jacobian Recursion and Left-Invertibility. We now derive a recursive formula for $J_{i+1}(\eta_0, 0)$. The map φ_{i+1} satisfies the fundamental identity (71): $g_\Omega(\varphi_{i+1}(\eta_0, \boldsymbol{\eta}_{i+1}), \eta_0) = \mathring{\varphi}_i(\eta_0, \boldsymbol{\eta}_{i+1})$. We differentiate this identity with respect to each block component η_j of the augmented vector $(\eta_0, \dots, \eta_{i+1})$ and then evaluate the result at $\boldsymbol{\eta}_{i+1} = 0$.

- **Derivative w.r.t. η_0:**

$$\frac{\partial g_\Omega}{\partial z}\frac{\partial \varphi_{i+1}}{\partial \eta_0}(\eta_0, \boldsymbol{\eta}_{i+1}) + \frac{\partial g_\Omega}{\partial o} = \frac{\partial}{\partial \eta_0}\left(\sum_{k=0}^{i}\frac{\partial \varphi_i}{\partial \eta_k}(\eta_0, \boldsymbol{\eta}_i)\eta_{k+1}\right).$$

  Evaluating at $\boldsymbol{\eta}_{i+1} = 0$, the right-hand side vanishes because every term in the sum contains a factor η_{k+1} evaluated at 0. With $\varphi_{i+1}(\eta_0, 0) = \varphi_0(\eta_0)$, this yields:

$$A(\eta_0)\frac{\partial \varphi_{i+1}}{\partial \eta_0}(\eta_0, 0) + B(\eta_0) = 0 \implies \frac{\partial \varphi_{i+1}}{\partial \eta_0}(\eta_0, 0) = -A(\eta_0)^{-1}B(\eta_0).$$

- **Derivative w.r.t. η_j (for $j \in \{1, \dots, i+1\}$):**

$$\frac{\partial g_\Omega}{\partial z}\frac{\partial \varphi_{i+1}}{\partial \eta_j}(\eta_0, \boldsymbol{\eta}_{i+1}) + 0 = \frac{\partial}{\partial \eta_j}\left(\sum_{k=0}^{i}\frac{\partial \varphi_i}{\partial \eta_k}(\eta_0, \boldsymbol{\eta}_i)\eta_{k+1}\right).$$

  Evaluating at $\boldsymbol{\eta}_{i+1} = 0$, the sum on the right-hand side drops all terms except the one where the derivative explicitly hits η_j (i.e., when $k+1 = j$). This gives:

$$A(\eta_0)\frac{\partial \varphi_{i+1}}{\partial \eta_j}(\eta_0, 0) = \frac{\partial \varphi_i}{\partial \eta_{j-1}}(\eta_0, 0) \implies \frac{\partial \varphi_{i+1}}{\partial \eta_j}(\eta_0, 0) = A(\eta_0)^{-1}\frac{\partial \varphi_i}{\partial \eta_{j-1}}(\eta_0, 0).$$

Combining these blocks, we obtain the recursive block-matrix relation:

$$\begin{aligned} J_{i+1}(\eta_0, 0) &= \left[\tfrac{\partial \varphi_{i+1}}{\partial \eta_0}(\eta_0, 0) \quad \tfrac{\partial \varphi_{i+1}}{\partial \eta_1}(\eta_0, 0) \quad \dots \quad \tfrac{\partial \varphi_{i+1}}{\partial \eta_{i+1}}(\eta_0, 0)\right] \\ &= A(\eta_0)^{-1}\left[-B(\eta_0) \quad \tfrac{\partial \varphi_i}{\partial \eta_0}(\eta_0, 0) \quad \dots \quad \tfrac{\partial \varphi_i}{\partial \eta_i}(\eta_0, 0)\right] \\ &= A(\eta_0)^{-1}\left[-B(\eta_0) \quad J_i(\eta_0, 0)\right]. \end{aligned}$$

Unrolling this recursion from $i = 0$ up to ℓ yields the explicit form:

$$\begin{aligned} J_\ell(\eta_0, 0) &= -A(\eta_0)^{-(\ell+1)}\left[A(\eta_0)^\ell B(\eta_0) \quad A(\eta_0)^{\ell-1}B(\eta_0) \quad \dots \quad B(\eta_0)\right], \\ &= -A(\eta_0)^{-(\ell+1)}C_\ell(\eta_0)P, \end{aligned}$$

where $C_\ell(\eta_0)$ is the matrix defined in the Filter Rank Condition (Assumption 3.6) and $P \in \mathbb{R}^{p(\ell+1)\times p(\ell+1)}$ is an invertible permutation matrix . Since $C_\ell(\eta_0)$ is left-invertible by assumption, $J_\ell(\eta_0, 0)$ is also left-invertible for all $\eta_0 \in \mathcal{O}$. By the Local Inversion Theorem, φ_ℓ is locally injective in an open neighborhood around every point $(\eta_0, 0)$.

3. Existence of ε_{inj} via Contradiction. We prove that there exists $\varepsilon_{inj} \in (0, \varepsilon_{ad}]$ such that φ_ℓ is injective on the entire set $\mathfrak{Y}(\varepsilon_{inj})$. We proceed by contradiction. Assume no such ε_{inj} exists. Then, for every $k \in \mathbb{N}^*$, we can pick $\epsilon_k = 1/k$. There must exist two distinct points $(\eta_{ak0}, \boldsymbol{\eta}_{ak\ell}) \neq (\eta_{bk0}, \boldsymbol{\eta}_{bk\ell})$ satisfying:

1. $(\eta_{ak0}, \boldsymbol{\eta}_{ak\ell}) \in \mathfrak{Y}(\epsilon_k)$ and $(\eta_{bk0}, \boldsymbol{\eta}_{bk\ell}) \in \mathfrak{Y}(\epsilon_k)$,
2. $\varphi_\ell(\eta_{ak0}, \boldsymbol{\eta}_{ak\ell}) = \varphi_\ell(\eta_{bk0}, \boldsymbol{\eta}_{bk\ell})$.

Since all points lie in the compact set $\mathfrak{Y}(\varepsilon_{ad})$, we can extract convergent subsequences. Let the limits be $(\eta_{ak0}, \boldsymbol{\eta}_{ak\ell}) \to (\eta_{a*0}, \boldsymbol{\eta}_{a*\ell})$ and $(\eta_{bk0}, \boldsymbol{\eta}_{bk\ell}) \to (\eta_{b*0}, \boldsymbol{\eta}_{b*\ell})$. Because $\rho(\boldsymbol{\eta}_{ak\ell}) \leq 1/k \to 0$ and $\rho(\boldsymbol{\eta}_{bk\ell}) \leq 1/k \to 0$, the limits of the derivative vectors must be strictly zero: $\boldsymbol{\eta}_{a*\ell} = 0$ and $\boldsymbol{\eta}_{b*\ell} = 0$. By the continuity of φ_ℓ, taking the limit in the equality yields:

$$\varphi_\ell(\eta_{a*0}, 0) = \lim_{k\to\infty}\varphi_\ell(\eta_{ak0}, \boldsymbol{\eta}_{ak\ell}) = \lim_{k\to\infty}\varphi_\ell(\eta_{bk0}, \boldsymbol{\eta}_{bk\ell}) = \varphi_\ell(\eta_{b*0}, 0).$$

Using (70), we know that $\varphi_\ell(\eta_0, 0) = \varphi_0(\eta_0)$. Thus, $\varphi_0(\eta_{a*0}) = \varphi_0(\eta_{b*0})$. By the global injectivity of φ_0 (proved in Step 1), this forces $\eta_{a*0} = \eta_{b*0} := \eta_{*0}$. Both sequences thus converge to the exact same limit point $(\eta_{*0}, 0)$.

However, we established in Step 2 that $J_\ell(\eta_{*0}, 0)$ is left-invertible, meaning φ_ℓ is locally injective in some neighborhood $\mathcal{N}(\eta_{*0}, 0)$. For k sufficiently large, both distinct points $(\eta_{ak0}, \boldsymbol{\eta}_{ak\ell})$ and $(\eta_{bk0}, \boldsymbol{\eta}_{bk\ell})$ will simultaneously enter this neighborhood while still being mapped to the identical value $\varphi_\ell(\dots)$. This strictly contradicts the local injectivity. Thus, a strictly positive ε_{inj} ensuring global injectivity on $\mathfrak{Y}(\varepsilon_{inj})$ must exist.

4. Lipschitz Injectivity Bound. The function φ_ℓ is injective and continuously differentiable on the compact convex set $\mathfrak{Y}(\varepsilon_{inj})$, and its Jacobian is full column rank everywhere on this set. Standard analysis results guarantee that the inverse map is globally Lipschitz on the image set $\varphi_\ell(\mathfrak{Y}(\varepsilon_{inj}))$. This guarantees the existence of the constant $L_{inj} > 0$ such that (77) holds. □

E.4 Approximation Error Dynamics and Fading Memory

We now study the properties of the true bounded solution $\varpi(\tau)$ relative to its approximation φ_ℓ. We define the approximation error as $e_\ell(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_\ell(\tau)) = \varpi(\tau) - \varphi_\ell(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_\ell(\tau))$.

Proposition E.4 (Approximation Error Magnitude)**.** *Suppose Assumptions 3.1 and 3.2 are satisfied. There exists a constant $\bar{\phi}$ such that, for any admissible input function $\tau \mapsto \mathscr{o}(\tau)$, we have:*

$$\sup_{\tau\in\mathbb{R}} |\varpi(\tau) - \varphi_\ell(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_\ell(\tau))| \le L_{gz}\bar{\phi}\left[\sup_{\tau\in\mathbb{R}} \rho(\boldsymbol{\mathscr{o}}_{\ell+1}(\tau))\right]^{\ell+1}. \tag{78}$$

Proof. As established in (67) and (71), when the argument of φ_ℓ is the valid time trajectory, its derivative is exactly:

$$\frac{d}{d\tau}\{\varphi_\ell(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_\ell(\tau))\} = \mathring{\varphi}_\ell(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_{\ell+1}(\tau)) = g_\Omega(\varphi_{\ell+1}(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_{\ell+1}(\tau)), \mathscr{o}(\tau)).$$

Using the shorthand notation $\varphi_i(\tau) = \varphi_i(\mathscr{o}(\tau), \boldsymbol{\mathscr{o}}_i(\tau))$, the dynamics of the approximation error are:

$$\frac{de_\ell}{d\tau}(\tau) = \frac{d\varpi}{d\tau}(\tau) - \frac{d\varphi_\ell}{d\tau}(\tau) = g_\Omega(\varpi(\tau), \mathscr{o}(\tau)) - g_\Omega(\varphi_{\ell+1}(\tau), \mathscr{o}(\tau)). \tag{79}$$

Applying the Mean Value Theorem to the first term yields:

$$\begin{aligned} g_\Omega(\varpi(\tau), \mathscr{o}(\tau)) &= g_\Omega(\varphi_\ell(\tau) + e_\ell(\tau), \mathscr{o}(\tau)) \\ &= g_\Omega(\varphi_\ell(\tau), \mathscr{o}(\tau)) + \underbrace{\left(\int_0^1 \frac{\partial g_\Omega}{\partial z}(\varphi_\ell(\tau) + se_\ell(\tau), \mathscr{o}(\tau))\, ds\right)}_{:=A_\ell(\tau)} e_\ell(\tau). \end{aligned}$$

Similarly, with the notation $\phi_{\ell+1}(\tau) = \varphi_{\ell+1}(\tau) - \varphi_\ell(\tau)$:

$$\begin{aligned} g_\Omega(\varphi_{\ell+1}(\tau), \mathscr{o}(\tau)) &= g_\Omega(\varphi_\ell(\tau) + \phi_{\ell+1}(\tau), \mathscr{o}(\tau)) \\ &= g_\Omega(\varphi_\ell(\tau), \mathscr{o}(\tau)) + \underbrace{\left(\int_0^1 \frac{\partial g_\Omega}{\partial z}(\varphi_\ell(\tau) + s\phi_{\ell+1}(\tau), \mathscr{o}(\tau))\, ds\right)}_{:=B_\ell(\tau)} \phi_{\ell+1}(\tau). \end{aligned}$$

Note that by Lemma E.2, $|\phi_{\ell+1}(\tau)| \le \bar{\phi}\rho(\boldsymbol{\mathscr{o}}_{\ell+1}(\tau))^{\ell+1}$ as long as the input is admissible. Substituting these expansions back into (79), the common term $g_\Omega(\varphi_\ell, \mathscr{o})$ cancels out:

$$\frac{de_\ell}{d\tau}(\tau) = A_\ell(\tau)e_\ell(\tau) - B_\ell(\tau)\phi_{\ell+1}(\tau). \tag{80}$$

Because the sets $\mathcal{Z}$ and $\mathcal{O}$ are convex and invariant, the paths of integration remain inside $\mathcal{Z}$. By the contraction assumption, $A_\ell(\tau) + A_\ell(\tau)^\top \le -2I_m$. Moreover, $|B_\ell(\tau)| \le L_{gz}$. Applying Lemma B.1 gives:

$$|e_\ell(\tau)| \le e^{-(\tau-\tau_0)}|e_\ell(\tau_0)| + \int_{\tau_0}^{\tau} e^{-(\tau-s)} L_{gz}\bar{\phi}\rho(\boldsymbol{\mathscr{o}}_{\ell+1}(s))^{\ell+1}\, ds.$$

Since both $\varpi(\tau)$ and $\varphi_\ell(\tau)$ are bounded on $\mathbb{R}$, $e_\ell(\tau)$ is bounded. Letting $\tau_0 \to -\infty$, the initial condition term vanishes, and bounding the supremum of ρ directly yields (78). □

We now examine how this approximation error responds to input variations. We consider two admissible input histories, $\tau \mapsto \mathscr{o}_a(\tau)$ and $\tau \mapsto \mathscr{o}_b(\tau)$, generating solutions $\tau \mapsto \varpi_a(\tau)$ and $\tau \mapsto \varpi_b(\tau)$. Let $\Delta e_\ell = e_{\ell,a} - e_{\ell,b}$, $\Delta\boldsymbol{\mathscr{o}}_k = \boldsymbol{\mathscr{o}}_{ak} - \boldsymbol{\mathscr{o}}_{bk}$ and $\Delta\mathscr{o}_k = \mathscr{o}_{ak} - \mathscr{o}_{bk}$ where $\mathscr{o}_{ak} = \frac{d^k\mathscr{o}_a}{d\tau^k}$.

Corollary E.5 (Lipschitz Property of the Error w.r.t. Input History)**.** *Under the assumptions of Proposition E.3, there exist constants $c_e, c_u, c_{\ell+1} > 0$ such that for any decay rate $\lambda \in [0, 1/2)$ and bound $K > 0$, there exist global constants $L_{e\boldsymbol{o}}$ and $L_{e\mathscr{o}_{\ell+1}}$ such that, if the inputs take values in $\mathfrak{Y}(\rho_{fm})$ with*

$$\rho_{fm} \in \left[0, \min\left\{1, \left(\frac{1/2-\lambda}{c_e}\right)^{\frac{1}{\ell+1}}\right\}\right),$$

and satisfy the fading memory property for all $s \leq 0$:

$$e^{\lambda s}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\mathscr{o}_k(s)| \leq K\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\mathscr{o}_k(0)|, \tag{81}$$

then the approximation error difference at $\tau = 0$ is bounded by:

$$|\Delta e_\ell(0)| \leq L_{e\boldsymbol{o}}\rho_{fm}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\mathscr{o}_k(0)| + L_{e\mathscr{o}_{\ell+1}}\sup_{s\leq 0}e^{\lambda s}|\Delta\mathscr{o}_{\ell+1}(s)|. \tag{82}$$

Proof. Subtracting the dynamics (80) evaluated for the two trajectories yields:

$$\begin{aligned}\frac{d\Delta e_\ell}{d\tau}(\tau) &= [A_{a\ell}(\tau)e_{\ell,a}(\tau) - B_{a\ell}(\tau)\phi_{a(\ell+1)}(\tau)] - [A_{b\ell}(\tau)e_{\ell,b}(\tau) - B_{b\ell}(\tau)\phi_{b(\ell+1)}(\tau)]\\ &= A_{a\ell}(\tau)\Delta e_\ell(\tau) + F(\tau),\end{aligned}$$

where $F(\tau) = \Delta A_\ell(\tau)e_{\ell,b}(\tau) - B_{a\ell}(\tau)\Delta\phi_{\ell+1}(\tau) - \Delta B_\ell(\tau)\phi_{b(\ell+1)}(\tau)$, with $\Delta A_\ell = A_{a\ell} - A_{b\ell}$ and $\Delta B_\ell = B_{a\ell} - B_{b\ell}$.

Because the sets $\mathcal{Z}$ and $\mathcal{O}$ are convex and invariant, the paths of integration remain inside $\mathcal{Z}$. The monotonicity property guarantees $A_{a\ell}(\tau) + A_{a\ell}(\tau)^\top \leq -2I_m$. Applying Lemma B.1 gives:

$$|\Delta e_\ell(\tau)| \leq e^{-(\tau-\tau_0)}|\Delta e_\ell(\tau_0)| + \int_{\tau_0}^{\tau} e^{-(\tau-s)}|F(s)|\,ds. \tag{83}$$

Using the bounds from (42) and Lemma E.2, since all intermediate variables lie in the invariant set $\mathcal{Z}$ and observations in $\mathcal{O}$, we can explicitly bound each component:

$$\begin{aligned}
|\Delta A_\ell| &\leq L_{gzz}[|\Delta\varphi_\ell| + |\Delta e_\ell|] + L_{gzo}|\Delta\mathscr{o}|\\
|e_{\ell,b}(\tau)| &\leq L_{gz}\bar{\phi}\rho(\boldsymbol{o}_{b(\ell+1)})^{\ell+1} \quad \text{(Proposition E.4)}\\
|B_{a\ell}| &\leq L_{gz}\\
|\Delta\phi_{\ell+1}| &\leq \overline{\partial\phi}\sum_{k=0}^{\ell+1}\max\left\{\rho(\boldsymbol{o}_{a(\ell+1)}), \rho(\boldsymbol{o}_{b(\ell+1)})\right\}^{\ell+1-k}|\Delta\mathscr{o}_k| \quad \text{(Lemma E.2)}\\
|\Delta B_\ell| &\leq L_{gzz}[|\Delta\varphi_\ell| + |\Delta\phi_{\ell+1}|] + L_{gzo}|\Delta\mathscr{o}|\\
|\phi_{b(\ell+1)}| &\leq \bar{\phi}\rho(\boldsymbol{o}_{b(\ell+1)})^{\ell+1} \quad \text{(Lemma E.2)}\\
|\Delta\varphi_\ell| &\leq L_\varphi\sum_{k=0}^{\ell}|\Delta\mathscr{o}_k|\\
|\Delta\mathscr{o}| = |\Delta\mathscr{o}_0| &\leq \sum_{k=0}^{\ell}|\Delta\mathscr{o}_k| \leq \sum_{k=0}^{\ell}\rho_{fm}^{-k}|\Delta\mathscr{o}_k|.
\end{aligned}$$

Because we assume $\sup_\tau \max\left\{\rho(\boldsymbol{\sigma}_{a(\ell+1)}(\tau)), \rho(\boldsymbol{\sigma}_{b(\ell+1)}(\tau))\right\} \leq \rho_{fm}$, we can substitute these bounds into the expression of $F(\tau)$ and distribute the terms:

$$\begin{aligned}|F(\tau)| \leq{}& \left(L_{gzz}[|\Delta\varphi_\ell| + |\Delta e_\ell|] + L_{gzo}|\Delta\sigma|\right) L_{gz}\bar{\phi}\rho_{fm}^{\ell+1} \\ &+ L_{gz}\overline{\partial\phi}\sum_{k=0}^{\ell+1}\rho_{fm}^{\ell+1-k}|\Delta\sigma_k(\tau)| \\ &+ \left(L_{gzz}\left[|\Delta\varphi_\ell| + \overline{\partial\phi}\sum_{k=0}^{\ell+1}\rho_{fm}^{\ell+1-k}|\Delta\sigma_k(\tau)|\right] + L_{gzo}|\Delta\sigma|\right)\bar{\phi}\rho_{fm}^{\ell+1}\end{aligned}$$

This gives,

$$\begin{aligned}|F(\tau)| \leq{}& L_{gzz}L_{gz}\bar{\phi}\rho_{fm}^{\ell+1}|\Delta\varphi_\ell| + L_{gzz}\bar{\phi}\rho_{fm}^{\ell+1}|\Delta\varphi_\ell| + L_{gzz}L_{gz}\bar{\phi}\rho_{fm}^{\ell+1}|\Delta e_\ell(\tau)| \\ &+ L_{gzo}L_{gz}\bar{\phi}\rho_{fm}^{\ell+1}|\Delta\sigma| + L_{gzo}\bar{\phi}\rho_{fm}^{\ell+1}|\Delta\sigma| \\ &+ L_{gz}\overline{\partial\phi}\sum_{k=0}^{\ell+1}\rho_{fm}^{\ell+1-k}|\Delta\sigma_k(\tau)| + L_{gzz}\bar{\phi}\rho_{fm}^{\ell+1}\overline{\partial\phi}\sum_{k=0}^{\ell+1}\rho_{fm}^{\ell+1-k}|\Delta\sigma_k(\tau)|\end{aligned}$$

and,

$$\begin{aligned}|F(\tau)| \leq{}& L_{gzz}[L_{gz}+1]\bar{\phi}\rho_{fm}^{\ell+1}L_\varphi\sum_{k=0}^{\ell}|\Delta\sigma_k(\tau)| + L_{gzz}L_{gz}\bar{\phi}\rho_{fm}^{\ell+1}|\Delta e_\ell(\tau)| \\ &+ L_{gzo}[L_{gz}+1]\bar{\phi}\rho_{fm}^{\ell+1}|\Delta\sigma| \\ &+ [L_{gz}+L_{gzz}\bar{\phi}\rho_{fm}^{\ell+1}]\overline{\partial\phi}\sum_{k=0}^{\ell}\rho_{fm}^{\ell+1-k}|\Delta\sigma_k(\tau)| + [L_{gz}+L_{gzz}\bar{\phi}\rho_{fm}^{\ell+1}]\overline{\partial\phi}|\Delta\sigma_{\ell+1}(\tau)| \\ \leq{}& L_{gzz}L_{gz}\bar{\phi}\rho_{fm}^{\ell+1}|\Delta e_\ell(\tau)| \\ &+ \rho_{fm}\left([L_{gzz}L_\varphi + L_{gzo}][L_{gz}+1]\bar{\phi} + [L_{gz}+L_{gzz}\bar{\phi}\rho_{fm}^{\ell+1}]\overline{\partial\phi}\right)\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\sigma_k(\tau)| \\ &+ [L_{gz}+L_{gzz}\bar{\phi}\rho_{fm}^{\ell+1}]\overline{\partial\phi}|\Delta\sigma_{\ell+1}(\tau)| \\ ={}& c_e\rho_{fm}^{\ell+1}|\Delta e_\ell(\tau)| + c_u\rho_{fm}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\sigma_k(\tau)| + c_{\ell+1}|\Delta\sigma_{\ell+1}(\tau)|,\end{aligned}$$

where we logically bounded $\rho_{fm}^{\ell+1} \leq \frac{1/2-\lambda}{c_e}$ to define the generic constants independently of the specific observer gain:

$$\begin{aligned}c_e &= L_{gzz}L_{gz}\bar{\phi}, \\ c_u &= [L_{gzz}L_\varphi + L_{gzo}][L_{gz}+1]\bar{\phi} + \left[L_{gz} + L_{gzz}\bar{\phi}\frac{1/2-\lambda}{c_e}\right]\overline{\partial\phi}, \\ c_{\ell+1} &= \left[L_{gz} + L_{gzz}\bar{\phi}\frac{1/2-\lambda}{c_e}\right]\overline{\partial\phi}.\end{aligned}$$

Using this bound in (83) and multiplying both sides by e^τ yields:

$$\begin{aligned}e^\tau|\Delta e_\ell(\tau)| \leq{}& e^{\tau_0}|\Delta e_\ell(\tau_0)| + c_e\rho_{fm}^{\ell+1}\int_{\tau_0}^{\tau}e^s|\Delta e_\ell(s)|ds \\ &+ \int_{\tau_0}^{\tau}e^s\left(c_u\rho_{fm}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\sigma_k(s)| + c_{\ell+1}|\Delta\sigma_{\ell+1}(s)|\right)ds.\end{aligned}$$

Applying Grönwall's inequality allows us to explicitly solve for the error magnitude:

$$e^{\tau}|\Delta e_\ell(\tau)| \leq \left[e^{\tau_0} + c_e\rho_{fm}^{\ell+1}\int_{\tau_0}^{\tau} e^{\tau_0}e^{c_e\rho_{fm}^{\ell+1}(\tau-r)}dr\right]|\Delta e_\ell(\tau_0)|$$
$$+ c_e\rho_{fm}^{\ell+1}\int_{\tau_0}^{\tau} e^{c_e\rho_{fm}^{\ell+1}(\tau-r)}\left(\int_{\tau_0}^{r} e^{s}\left(c_u\rho_{fm}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\mathscr{o}_k(s)| + c_{\ell+1}|\Delta\mathscr{o}_{\ell+1}(s)|\right)ds\right)dr$$
$$+ \int_{\tau_0}^{\tau} e^{s}\left(c_u\rho_{fm}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\mathscr{o}_k(s)| + c_{\ell+1}|\Delta\mathscr{o}_{\ell+1}(s)|\right)ds.$$

This elegantly simplifies into a single consolidated convolution:

$$|\Delta e_\ell(\tau)| \leq e^{-(1-c_e\rho_{fm}^{\ell+1})(\tau-\tau_0)}|\Delta e_\ell(\tau_0)|$$
$$+ \int_{\tau_0}^{\tau} e^{-(1-c_e\rho_{fm}^{\ell+1})(\tau-s)}\left(c_u\rho_{fm}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\mathscr{o}_k(s)| + c_{\ell+1}|\Delta\mathscr{o}_{\ell+1}(s)|\right)ds.$$

Since both ϖ and φ_ℓ are bounded on $\mathbb{R}$, the approximation error $|\Delta e_\ell|$ is bounded. We can thus safely let the initial time $\tau_0 \to -\infty$, causing the initial condition term to vanish. Evaluating at the current time $\tau = 0$ and utilizing the fading memory assumption (81), we obtain:

$$|\Delta e_\ell(0)| \leq \int_{-\infty}^{0} e^{(1-c_e\rho_{fm}^{\ell+1})s}\left(c_u\rho_{fm}K\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta o_k(0)| + c_{\ell+1}e^{\lambda s}|\Delta\mathscr{o}_{\ell+1}(s)|\right)e^{-\lambda s}ds$$
$$\leq \frac{1}{1-c_e\rho_{fm}^{\ell+1}-\lambda}\left[c_u\rho_{fm}K\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta o_k(0)| + c_{\ell+1}\sup_{s\leq 0}e^{\lambda s}|\Delta\mathscr{o}_{\ell+1}(s)|\right].$$

By enforcing the strict threshold $\rho_{fm} < \left(\frac{1/2-\lambda}{c_e}\right)^{\frac{1}{\ell+1}}$, we guarantee that the integration denominator $1 - c_e\rho_{fm}^{\ell+1} - \lambda > 1/2$. Thus, the fractional multiplier is strictly bounded by 2. This yields the final expression (82) with the constants explicitly given as:

$$L_{e\boldsymbol{\mathscr{o}}} = 2c_uK = 2\left([L_{gzz}L_\varphi + L_{gzo}][L_{gz}+1]\bar{\phi} + \left[L_{gz} + L_{gzz}\bar{\phi}\frac{1/2-\lambda}{c_e}\right]\overline{\partial\phi}\right)K,$$
$$L_{e\mathscr{o}_{\ell+1}} = 2c_{\ell+1} = 2\left[L_{gz} + L_{gzz}\bar{\phi}\frac{1/2-\lambda}{c_e}\right]\overline{\partial\phi}.$$

□

Proposition E.6 (General Input Derivative Reconstruction Bound)**.** *Suppose Assumptions 3.1, 3.2, and 3.6 hold. Let $\lambda \in [0, 1/2)$ and $K > 0$, and let $c_e, L_{inj}, \varepsilon_{inj}, L_{e\boldsymbol{\mathscr{o}}}, L_{e\mathscr{o}_{\ell+1}}$ be the constants defined previously. If two admissible inputs $\mathscr{o}_a(\tau)$ and $\mathscr{o}_b(\tau)$ take values in $\mathfrak{Y}(\rho_{fm})$ where:*

$$\rho_{fm} < \min\left\{1, \varepsilon_{inj}, \left(\frac{1/2-\lambda}{c_e}\right)^{\frac{1}{\ell+1}}, \frac{1}{2\sqrt{\ell+1}L_{inj}L_{e\boldsymbol{\mathscr{o}}}}\right\}, \tag{84}$$

and satisfy the fading memory property (81), then the following bound holds:

$$\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta\mathscr{o}_k(0)| \leq 2\sqrt{\ell+1}L_{inj}\left[|\Delta\varpi(0)| + L_{e\mathscr{o}_{\ell+1}}\sup_{s\leq 0}e^{\lambda s}|\Delta\mathscr{o}_{\ell+1}(s)|\right]. \tag{85}$$

Proof. By definition, $\Delta e_\ell(\tau) = \Delta\varpi(\tau) - \Delta\varphi_\ell(\tau)$. Since the inputs evaluated up to order ℓ belong to $\mathfrak{Y}(\varepsilon_{inj})$ (because $\rho_{fm} \leq \varepsilon_{inj}$), Proposition E.3 applies:

$$|\Delta\mathscr{o}(\tau)| + |\Delta\boldsymbol{\mathscr{o}}_\ell(\tau)| \leq L_{inj}|\Delta\varphi_\ell(\tau)| \leq L_{inj}\left(|\Delta\varpi(\tau)| + |\Delta e_\ell(\tau)|\right).$$

Evaluating at $\tau = 0$ and substituting the bound for $|\Delta e_\ell(0)|$ from Corollary E.5 yields:

$$|\Delta\mathscr{o}(0)| + |\Delta\boldsymbol{\mathscr{o}}_\ell(0)| \leq L_{inj}|\Delta\varpi(0)|$$
$$+ L_{inj}L_{e\boldsymbol{\mathscr{o}}}\rho_{fm}\sum_{k=0}^{\ell}\rho_{fm}^{\ell-k}|\Delta o_k(0)| + L_{inj}L_{e\mathscr{o}_{\ell+1}}\sup_{s\leq 0}e^{\lambda s}|\Delta\mathscr{o}_{\ell+1}(s)|.$$

Because $\rho_{fm} < 1$, we can use the Cauchy-Schwarz inequality to bound the weighted sum on the left hand side:

$$\sum_{k=0}^{\ell} \rho_{fm}^{\ell-k} |\Delta o_k(0)| \leq \sqrt{\sum_{k=0}^{\ell} \rho_{fm}^{2(\ell-k)}} \sqrt{\sum_{k=0}^{\ell} |\Delta o_k(0)|^2} \leq \sqrt{\ell+1} \left(|\Delta \mathscr{o}(0)| + |\Delta \boldsymbol{\mathscr{o}}_\ell(0)| \right).$$

Combining these inequalities, we obtain:

$$\left[\frac{1}{\sqrt{\ell+1}} - L_{inj} L_{e\boldsymbol{\mathscr{o}}} \rho_{fm} \right] \sum_{k=0}^{\ell} \rho_{fm}^{\ell-k} |\Delta o_k(0)| \leq L_{inj} |\Delta \varpi(0)| + L_{inj} L_{e\mathscr{o}_{\ell+1}} \sup_{s \leq 0} e^{\lambda s} |\Delta \mathscr{o}_{\ell+1}(s)|.$$

Thanks to the strict threshold (84), the bracketed term is strictly greater than $\frac{1}{2\sqrt{\ell+1}}$. Multiplying by $2\sqrt{\ell+1}$ directly yields (85). □

E.5 Synthesis: Proof of Theorem 3.7

We are now ready to close the proof of Theorem 3.7. Recall from Appendix D that $\boldsymbol{T}_\kappa$ defined in equation (50) is uniquely defined via the bounded solution of the latent dynamics. By construction, this specific mapping already satisfies the invariance, attractivity, and regularity properties established in Theorem 3.4. We now prove that for a sufficiently large gain κ, this exact same mapping is also Lipschitz injective.

We consider the trajectories $\Phi_f(x_{a0}, t)$ and $\Phi_f(x_{b0}, t)$ initialized in the set $\mathcal{X}$ where Assumption 3.5 holds (recall that the modification given in (38) preserve the satisfaction of this assumption) and denote $\Delta x(t) = \Phi_f(x_{a0}, t) - \Phi_f(x_{b0}, t)$. We analyze the observations in the observer time scale $\tau = \kappa t$. The observation derivatives are given by scaled Lie derivatives:

$$\mathscr{o}_{x_0,i}(\tau) = \frac{d^i \mathscr{o}_{x_0}}{d\tau^i}(\tau) = \frac{1}{\kappa^i} \mathcal{L}_f^i h \left(\Phi_f \left(x_0, \frac{\tau}{\kappa} \right) \right). \tag{86}$$

1. Verification of the Assumptions. We define the upper bound for the physical measurement derivatives on $\mathcal{X}$:

$$\bar{\mathscr{o}}' = \sup_{x \in \mathcal{X}} \max \left\{ |\mathcal{L}_f h(x)|, |\mathcal{L}_f^2 h(x)|^{1/2}, \ldots, |\mathcal{L}_f^{\ell+1} h(x)|^{1/(\ell+1)} \right\}.$$

Using (86), the magnitude function (63) satisfies:

$$\rho \left(\boldsymbol{\mathscr{o}}_{x_0,\ell+1}(\tau) \right) \leq \frac{\bar{\mathscr{o}}'}{\kappa}.$$

By choosing the gain κ large enough such that $\kappa \geq \bar{\mathscr{o}}'/\varepsilon_{ad}$, the input is guaranteed to be admissible. We set $\rho_{fm} = \bar{\mathscr{o}}'/\kappa$.

Next, we verify the fading memory property (81). The differential observability on $\mathcal{X}$ (Assumption 3.5) gives

$$c_{obs} |\Delta x(0)| = c_{obs} |x_{a0} - x_{b0}| \leq \sum_{i=0}^{\ell} |\mathcal{L}_f^i h(x_{a0}) - \mathcal{L}_f^i h(x_{b0})| = \sum_{i=0}^{\ell} |\Delta \mathscr{o}_i(0)|$$

Scaling this with $\kappa \geq 1$ yields:

$$c_{obs} |\Delta x(0)| \leq \sum_{i=0}^{\ell} \kappa^i |\Delta \mathscr{o}_i(0)|. \tag{87}$$

By the global Lipschitz property of the Lie derivatives (with constant $L_H = \sum_{i=0}^{\ell} \mathrm{Lip}(\mathcal{L}_f^i h)$), we have $\kappa^j |\Delta \mathscr{o}_j(s)| \leq \mathrm{Lip}(\mathcal{L}_f^j h) |\Delta x(s/\kappa)|$. Because the physical flow (recall that the modification given in (38)) is governed by a Lipschitz vector field with constant L_f, the backward-in-time distance is bounded by:

$$|\Delta x(s/\kappa)| \leq e^{-\frac{L_f}{\kappa} s} |\Delta x(0)| \qquad \forall s \leq 0. \tag{88}$$

Setting $\lambda = L_f/\kappa$, this implies:

$$e^{\lambda s} \sum_{i=0}^{\ell} \kappa^i |\Delta \mathscr{o}_i(s)| \leq L_H |\Delta x(0)| \leq \frac{L_H}{c_{obs}} \sum_{i=0}^{\ell} \kappa^i |\Delta \mathscr{o}_i(0)| \qquad \forall s \leq 0.$$

Multiplying by $\kappa^{-\ell}$ and substituting $\rho_{fm} = \bar{\sigma}'/\kappa$, we satisfy (81) with $K = \frac{L_H}{c_{obs}} \max\{(\bar{\sigma}')^\ell, (\bar{\sigma}')^{-\ell}\}$. All conditions for Proposition E.6 are met provided κ is chosen large enough to enforce (84).

2. Combination of Observability and Asymptotic Bounds. Starting from the scaled observability inequality (87) and pulling out $\kappa^{-\ell}$:

$$\begin{aligned}\frac{\min\{1,(\bar{\sigma}')^\ell\}}{\kappa^\ell} c_{obs}|\Delta x(0)| &\le \frac{1}{\kappa^\ell}\sum_{i=0}^{\ell}\kappa^i(\bar{\sigma}')^{\ell-i}|\Delta o_i(0)| \\ &= \rho_{fm}^\ell \sum_{i=0}^{\ell}\rho_{fm}^{-i}|\Delta o_i(0)| = \sum_{i=0}^{\ell}\rho_{fm}^{\ell-i}|\Delta o_i(0)|. \end{aligned} \tag{89}$$

Applying Proposition E.6, and recalling (with (50)) that $\Delta\varpi(0) = \boldsymbol{T}_\kappa(x_{a0}) - \boldsymbol{T}_\kappa(x_{b0})$, we obtain:

$$\sum_{i=0}^{\ell}\rho_{fm}^{\ell-i}|\Delta o_i(0)| \le 2\sqrt{\ell+1}L_{inj}\left[|\boldsymbol{T}_\kappa(x_{a0}) - \boldsymbol{T}_\kappa(x_{b0})| + R_{rem}\right], \tag{90}$$

where $R_{rem} = L_{eo_{\ell+1}} \sup_{s\le 0} e^{\lambda s}|\Delta o_{\ell+1}(s)|$. Using the Lipschitz constant of the $(\ell+1)$-th Lie derivative and (88):

$$\begin{aligned} e^{\lambda s}|\Delta o_{\ell+1}(s)| &= e^{\frac{L_f}{\kappa}s}\frac{1}{\kappa^{\ell+1}}|\mathcal{L}_f^{\ell+1}h(\Phi_f(x_{a0}, s/\kappa)) - \mathcal{L}_f^{\ell+1}h(\Phi_f(x_{b0}, s/\kappa))| \\ &\le \frac{\mathrm{Lip}(\mathcal{L}_f^{\ell+1}h)}{\kappa^{\ell+1}} e^{\frac{L_f}{\kappa}s}|\Delta x(s/\kappa)| \\ &\le \frac{\mathrm{Lip}(\mathcal{L}_f^{\ell+1}h)}{\kappa^{\ell+1}}|\Delta x(0)|. \end{aligned}$$

Substituting this bound into R_{rem} and combining with (89) yields:

$$\begin{aligned}\frac{\min\{1,(\bar{\sigma}')^\ell\}c_{obs}}{\kappa^\ell}|\Delta x(0)| &\le 2\sqrt{\ell+1}L_{inj}|\boldsymbol{T}_\kappa(x_{a0}) - \boldsymbol{T}_\kappa(x_{b0})| \\ &\quad + 2\sqrt{\ell+1}L_{inj}L_{eo_{\ell+1}}\frac{\mathrm{Lip}(\mathcal{L}_f^{\ell+1}h)}{\kappa^{\ell+1}}|\Delta x(0)|. \end{aligned}$$

Multiplying the entire equation by κ^ℓ and grouping the terms proportional to $|\Delta x(0)|$:

$$\begin{aligned}\left(\min\{1,(\bar{\sigma}')^\ell\}c_{obs} - \frac{2\sqrt{\ell+1}L_{inj}L_{eo_{\ell+1}}\mathrm{Lip}(\mathcal{L}_f^{\ell+1}h)}{\kappa}\right)|\Delta x(0)| \\ \le 2\sqrt{\ell+1}L_{inj}\kappa^\ell|\boldsymbol{T}_\kappa(x_{a0}) - \boldsymbol{T}_\kappa(x_{b0})|. \end{aligned}$$

If we select the minimal gain $\underline{\kappa}$ such that:

$$\underline{\kappa} > \frac{4\sqrt{\ell+1}L_{inj}L_{eo_{\ell+1}}\mathrm{Lip}(\mathcal{L}_f^{\ell+1}h)}{\min\{1,(\bar{\sigma}')^\ell\}c_{obs}},$$

then the bracketed coefficient on the left is strictly positive and bounded below by $\frac{1}{2}\min\{1,(\bar{\sigma}')^\ell\}c_{obs}$. Dividing both sides by this strictly positive constant yields:

$$|x_{a0} - x_{b0}| = |\Delta x(0)| \le c_T\kappa^\ell|\boldsymbol{T}_\kappa(x_{a0}) - \boldsymbol{T}_\kappa(x_{b0})|,$$

where $c_T = \frac{4\sqrt{\ell+1}L_{inj}}{\min\{1,(\bar{\sigma}')^\ell\}c_{obs}}$. Since $\boldsymbol{T}_\kappa$ is the mapping constructed in Theorem 3.4, this completely proves that it satisfies both the conclusions of Theorem 3.4 and the Lipschitz injectivity bound, concluding the proof of Theorem 3.7. □

F Decoder Approximation and Complexity (Theorems 4.1 and 4.2)

This appendix provides the proofs for the convergence of the ideal observer (Theorem 4.1) and the neural decoder complexity bound (Theorem 4.2). A fundamental prerequisite for both proofs is to formally extend the exact inverse map from the embedded manifold to the entire latent operating region.

F.1 Global Extension of the Inverse Map (McShane-Whitney)

From Theorem 3.7, the map $\boldsymbol{T}_\kappa : \mathcal{X} \to \mathbb{R}^m$ is Lipschitz injective on $\mathcal{X}$ for $\kappa \geq \underline{\kappa}$, satisfying:

$$|x_a - x_b| \leq c_T \kappa^\ell |\boldsymbol{T}_\kappa(x_a) - \boldsymbol{T}_\kappa(x_b)| \qquad \forall x_a, x_b \in \mathcal{X}.$$

This defines a strict left-inverse $\boldsymbol{T}_\kappa^{\text{inv}} : \boldsymbol{T}_\kappa(\mathcal{X}) \to \mathcal{X}$. To ensure that our decoder can process any state within the latent operating region $\mathcal{Z}$, we must extend this inverse map to the entire space $\mathbb{R}^m$ while preserving its Lipschitz continuity.

We construct the extended map, denoted $\boldsymbol{T}_\kappa^{-1} : \mathbb{R}^m \to \mathbb{R}^n$, component by component using the McShane-Whitney extension theorem. For each dimension $i \in \{1, \dots, n\}$, we define:

$$\boldsymbol{T}_\kappa^{-1}(z)_i = \sup_{\chi \in \mathcal{X}} \left\{ \chi_i - c_T \kappa^\ell |z - \boldsymbol{T}_\kappa(\chi)| \right\}. \tag{91}$$

Exact Inverse on $\mathcal{X}$: If z is on the embedded manifold, i.e., $z = \boldsymbol{T}_\kappa(x)$ for some $x \in \mathcal{X}$, then the injectivity bound directly implies:

$$\chi_i - x_i \leq |\chi - x| \leq c_T \kappa^\ell |\boldsymbol{T}_\kappa(\chi) - \boldsymbol{T}_\kappa(x)| \qquad \forall \chi \in \mathcal{X}.$$

Rearranging yields $\chi_i - c_T \kappa^\ell |\boldsymbol{T}_\kappa(x) - \boldsymbol{T}_\kappa(\chi)| \leq x_i$ for all $\chi \in \mathcal{X}$. The supremum over χ in (91) is therefore upper-bounded by x_i. However, taking $\chi = x$ exactly achieves this bound (since the distance term vanishes). Thus, $\boldsymbol{T}_\kappa^{-1}(\boldsymbol{T}_\kappa(x))_i = x_i$ for all i, proving that $\boldsymbol{T}_\kappa^{-1}(\boldsymbol{T}_\kappa(x)) = x$ on $\mathcal{X}$.

Global Lipschitz Continuity: Let $z_a, z_b \in \mathbb{R}^m$ be two arbitrary points. Without loss of generality, assume $\boldsymbol{T}_\kappa^{-1}(z_b)_i \geq \boldsymbol{T}_\kappa^{-1}(z_a)_i$. Because $\mathcal{X}$ is compact, these suprema are finite, and we have:

$$\begin{aligned} 0 &\leq \boldsymbol{T}_\kappa^{-1}(z_b)_i - \boldsymbol{T}_\kappa^{-1}(z_a)_i \\ &= \sup_{\chi_b \in \mathcal{X}} \left\{ \chi_{bi} - c_T \kappa^\ell |z_b - \boldsymbol{T}_\kappa(\chi_b)| \right\} - \sup_{\chi_a \in \mathcal{X}} \left\{ \chi_{ai} - c_T \kappa^\ell |z_a - \boldsymbol{T}_\kappa(\chi_a)| \right\}. \end{aligned}$$

Since $\sup_{\chi_a} \{\dots\}$ bounds the expression for any specific choice, we can substitute $\chi_a = \chi_b$ in the negative term to obtain an upper bound:

$$\begin{aligned} \boldsymbol{T}_\kappa^{-1}(z_b)_i - \boldsymbol{T}_\kappa^{-1}(z_a)_i &\leq \sup_{\chi_b \in \mathcal{X}} \left\{ \chi_{bi} - c_T \kappa^\ell |z_b - \boldsymbol{T}_\kappa(\chi_b)| - \left(\chi_{bi} - c_T \kappa^\ell |z_a - \boldsymbol{T}_\kappa(\chi_b)| \right) \right\} \\ &\leq c_T \kappa^\ell \sup_{\chi_b \in \mathcal{X}} \left\{ |z_a - \boldsymbol{T}_\kappa(\chi_b)| - |z_b - \boldsymbol{T}_\kappa(\chi_b)| \right\}. \end{aligned}$$

By the reverse triangle inequality, this evaluates to:

$$|\boldsymbol{T}_\kappa^{-1}(z_b)_i - \boldsymbol{T}_\kappa^{-1}(z_a)_i| \leq c_T \kappa^\ell |z_a - z_b| \qquad \forall i \in \{1, \dots, n\}. \tag{92}$$

Summing the squares over all n components yields the Euclidean Lipschitz bound for the vector map:

$$|\boldsymbol{T}_\kappa^{-1}(z_b) - \boldsymbol{T}_\kappa^{-1}(z_a)| \leq \sqrt{n} c_T \kappa^\ell |z_b - z_a| := L_{inv} |z_b - z_a|. \tag{93}$$

F.2 Proof of Theorem 4.1

Let $z(t) = \Phi_z(z_0, 0, \kappa t; x_0)$ be the latent trajectory starting at $z_0 \in \mathcal{Z}$ and driven by the physical observation sequence the initial condition of which x_0 is in $\mathcal{X}$ assumed to be forward invariant. The ideal estimate is defined as $\hat{x}(t) = \boldsymbol{T}_\kappa^{-1}(z(t))$. Since the physical trajectory $x(t)$ remains in $\mathcal{X}$ for all $t \geq 0$, we have the exact identity $x(t) = \boldsymbol{T}_\kappa^{-1}(\boldsymbol{T}_\kappa(x(t)))$. The estimation error is therefore:

$$|\hat{x}(t) - x(t)| = |\boldsymbol{T}_\kappa^{-1}(z(t)) - \boldsymbol{T}_\kappa^{-1}(\boldsymbol{T}_\kappa(x(t)))|.$$

Using the global Lipschitz property (93) of $\boldsymbol{T}_\kappa^{-1}$:

$$|\hat{x}(t) - x(t)| \leq L_{inv} |z(t) - \boldsymbol{T}_\kappa(x(t))|.$$

Applying the exponential attractivity bound from Theorem 3.4 yields:

$$|\hat{x}(t) - x(t)| \leq L_{inv} e^{-\kappa t} |z_0 - \boldsymbol{T}_\kappa(x_0)| = \sqrt{n} c_T \kappa^\ell e^{-\kappa t} |z_0 - \boldsymbol{T}_\kappa(x_0)|.$$

Setting $c_T' = \sqrt{n} c_T$ gives the bound (15). Since both z_0 and $\boldsymbol{T}_\kappa(x_0)$ belong to the compact set $\mathcal{Z}$ (a ball of radius R_z), their distance is bounded by $2R_z$, so that

$$|\hat{x}(t) - x(t)| \leq 2 R_z c_T' \kappa^\ell e^{-\kappa t}. \tag{94}$$

For any finite $t_o > 0$ and any $\epsilon > 0$, the exponential term $e^{-\kappa t_o}$ dominates the polynomial term κ^ℓ for sufficiently large κ. Thus, we can tune κ to ensure $|\hat{x}(t) - x(t)| \leq \epsilon$ for all $t \geq t_o$, proving tunability.

F.3 Proof of Theorem 4.2 (MLP Decoder Complexity)

In practice, the theoretical decoder $\boldsymbol{T}_\kappa^{-1}$ is replaced by an MLP $\gamma_\Omega(z)$. We decompose the overall estimation error using the triangle inequality:

$$|\gamma_\Omega(z(t)) - x(t)| \leq \underbrace{|\gamma_\Omega(z(t)) - \boldsymbol{T}_\kappa^{-1}(z(t))|}_{e_{approx}(t)} + \underbrace{|\boldsymbol{T}_\kappa^{-1}(z(t)) - x(t)|}_{e_{obs}(t)}.$$

We must bound the sum by ϵ for $t \geq t_o$. First, using Theorem 4.1, we choose a structural gain κ^* large enough so that the dynamic observer error satisfies:

$$e_{obs}(t) \leq 2R_z c_T' (\kappa^*)^\ell e^{-\kappa^* t_o} \leq \frac{\epsilon}{2} \qquad \forall t \geq t_o.$$

With κ^* fixed, we must construct a neural network γ_Ω that approximates the Lipschitz function $\boldsymbol{T}_\kappa^{-1}$ on the compact set $\mathcal{Z}$ with a uniform error bounded by $\epsilon/2$ in the Euclidean norm.

We invoke the quantitative approximation results from Yarotsky [58, Theorem 1], which states that any 1-Lipschitz scalar function $f : [0,1]^m \to \mathbb{R}$ can be approximated with precision $\delta \in (0,1)$ using a ReLU network containing $\mathcal{W}_{scalar}$ weights, satisfying:

$$\mathcal{W}_{scalar} \leq c_m \left(\delta^{-m} \log(1/\delta) + 1\right),$$

where c_m is a constant depending only on the input dimension m.

Since our extended inverse map $\boldsymbol{T}_\kappa^{-1}$ outputs an n-dimensional vector, bounding the Euclidean error by $\epsilon/2$ is guaranteed if each of the n scalar coordinate functions is approximated with a precision of $\epsilon/(2\sqrt{n})$. We approximate these n coordinate functions by constructing n parallel scalar networks (or equivalently, a single multi-head network), bounding the total weight complexity by $\mathcal{W} \leq n\mathcal{W}_{scalar}$.

To apply Yarotsky's theorem to each coordinate function, we perform a spatial and an amplitude scaling to map their domains to the unit hypercube and normalize their Lipschitz constants to 1:

- **Spatial Scaling:** $\mathcal{Z}$ is contained in the hypercube $[-R_z, R_z]^m$. We map the standard hypercube $[0,1]^m$ to this domain via the affine transformation $x \mapsto 2R_z x - R_z \mathbf{1}_m$. Composing the coordinate functions with this mapping scales their effective Lipschitz constants with respect to x by a factor of $2R_z$.
- **Amplitude Scaling:** Each resulting composed coordinate function has a Lipschitz constant bounded by $2R_z L_{inv} = 2R_z c_T' (\kappa^*)^\ell$. To normalize them into 1-Lipschitz functions as required by the theorem, we must divide their outputs by this exact factor.

Consequently, achieving an absolute coordinate error of $\epsilon/(2\sqrt{n})$ for the original functions is mathematically equivalent to achieving a precision δ for the fully normalized functions, where δ is scaled by the same Lipschitz factor:

$$\delta = \frac{\epsilon/(2\sqrt{n})}{2R_z c_T' (\kappa^*)^\ell} = \frac{\epsilon}{4\sqrt{n} R_z c_T' (\kappa^*)^\ell}.$$

Substituting this specific required precision δ into Yarotsky's complexity bound, and multiplying by n for the total network, yields:

$$\mathcal{W} \leq n c_m \left(\left(\frac{4\sqrt{n} R_z c_T' (\kappa^*)^\ell}{\epsilon} \right)^m \log \left(\frac{4\sqrt{n} R_z c_T' (\kappa^*)^\ell}{\epsilon} \right) + 1 \right).$$

By grouping all the constants that are independent of the tunable gain κ^* and the precision ϵ (namely $n, 4, R_z, c_T', c_m$) into a single global constant C_{net}, we obtain the simplified asymptotic complexity bound:

$$\mathcal{W} \leq C_{net} \left(\frac{(\kappa^*)^\ell}{\epsilon} \right)^m \log \left(\frac{(\kappa^*)^\ell}{\epsilon} \right).$$

Thus, there exists a network architecture γ_Ω with this bounded number of weights satisfying $e_{approx}(t) \leq \epsilon/2$. The sum of the dynamic error and the approximation error guarantees the global precision ϵ, concluding the proof.

F.4 Proof of Proposition 4.3 (Robustness to Measurement Disturbances)

Throughout the proof, we fix $\kappa \geq \underline{\kappa}$, a physical initial condition $x_0 \in \mathcal{X}$ and a latent initial condition $z_0 \in \mathcal{Z}$. We denote by $x(t) = \Phi_f(x_0, t)$ the physical trajectory, which remains in $\mathcal{X}$ for all $t \geq 0$ by Assumption 3.1, and by $o(t) = h(x(t)) \in \mathcal{O}$ the nominal observation. The disturbance $\delta : [0, +\infty) \to \mathbb{R}^p$ is Lebesgue measurable (it is unknown and need not be continuous), with $|\delta(t)| \leq \|\delta\|_\infty \leq \bar{d}$ for almost all $t \geq 0$: $\|\delta\|_\infty$ is the amplitude of the actual disturbance, whereas $\bar{d}$ is the admissible threshold defined in Proposition 4.3. The quantity $d(t)$ is its fading-memory version defined in Proposition 4.3. We compare two latent trajectories starting from the same initial condition $z_0 \in \mathcal{Z}$:

- the *nominal* latent trajectory $z(t)$, solution of $\frac{dz}{dt} = \kappa g_\Omega(z, o(t))$, which is the one studied in Theorems 3.4 to 4.1;
- the *perturbed* latent trajectory $z_d(t)$, solution of

$$\frac{dz_d}{dt} = \kappa g_\Omega(z_d, o(t) + \delta(t)), \qquad z_d(0) = z_0, \tag{95}$$

  which is the one actually implemented, and from which the estimate $\hat{x}(t) = \gamma_\Omega(z_d(t))$ is computed.

Preliminaries. Since δ is only measurable, the right-hand side of (95) is measurable in t, and it is continuously differentiable in z_d by Assumption 3.1. Its solutions are therefore understood in the sense of Carathéodory, namely absolutely continuous functions satisfying (95) for almost all t. As the considered input $o(t) + \delta(t)$ takes its values in the compact set

$$\mathcal{O}^{d_0} = \{o \in \mathbb{R}^p : \mathrm{dist}(o, \mathcal{O}) \leq d_0\},$$

the right-hand side of (95) is bounded and Lipschitz in z_d on any compact set, uniformly in t. By Carathéodory's theorem (see e.g. [19, Chapter I.5]), (95) admits a unique solution defined on a maximal interval $[0, t_{\max})$, and $t_{\max} = +\infty$ if this solution remains in a compact set.

The set $\mathcal{O}^{d_0}$ is compact since $\mathcal{O}$ is, so that the constant $\bar{L}_{go}$ defined in Proposition 4.3 is finite as the maximum of a continuous function on the compact set $\mathcal{Z} \times \mathcal{O}^{d_0}$. Since $\mathcal{O} \subset \mathcal{O}^{d_0}$, it is larger than or equal to the constant L_{go} of (42). If $\bar{L}_{go} = 0$, we use the convention $\bar{d} = d_0$. The corrupted observation then takes its values in $\mathcal{O}^{\bar{d}} = \{o \in \mathbb{R}^p : \mathrm{dist}(o, \mathcal{O}) \leq \bar{d}\} \subset \mathcal{O}^{d_0}$. Recall also, from Appendix C, the notation $\bar{g} = \max_{o \in \mathcal{O}} |g_\Omega(0, o)|$, and that the radius of $\mathcal{Z}$ satisfies $R_z \geq 2\bar{g}$.

We then establish two useful inequalities: for all $z \in \mathcal{Z}$, all $o \in \mathcal{O}$ and all $\delta_0 \in \mathbb{R}^p$ with $|\delta_0| \leq \bar{d}$,

$$|g_\Omega(z, o + \delta_0) - g_\Omega(z, o)| \leq \bar{L}_{go} |\delta_0|, \tag{96}$$

$$z^\top g_\Omega(z, o + \delta_0) \leq |z| \left(\bar{g} + \bar{L}_{go} |\delta_0| - |z|\right). \tag{97}$$

To establish (96), note that for all $s \in [0, 1]$, $\mathrm{dist}(o + s\delta_0, \mathcal{O}) \leq s|\delta_0| \leq \bar{d} \leq d_0$, so that the segment $[o, o + \delta_0]$ is contained in $\mathcal{O}^{d_0}$. Hence, by the integral mean value theorem,

$$|g_\Omega(z, o + \delta_0) - g_\Omega(z, o)| = \left| \int_0^1 \frac{\partial g_\Omega}{\partial o}(z, o + s\delta_0)\, \delta_0 \, ds \right| \leq \bar{L}_{go} |\delta_0|.$$

To establish (97), we decompose

$$z^\top g_\Omega(z, o + \delta_0) = z^\top \left[g_\Omega(z, o) - g_\Omega(0, o)\right] + z^\top g_\Omega(0, o) + z^\top \left[g_\Omega(z, o + \delta_0) - g_\Omega(z, o)\right].$$

Since $o \in \mathcal{O}$ and the segment $[0, z]$ is contained in the ball $\mathcal{Z}$, the first term is bounded by $-|z|^2$ exactly as in Appendix C (integral mean value theorem and Assumption 3.2). The second term is bounded by $|z|\bar{g}$ by the Cauchy-Schwarz inequality, and the third one by $|z|\bar{L}_{go}|\delta_0|$ by the Cauchy-Schwarz inequality and (96). This gives (97).

Step 1: forward invariance of $\mathcal{Z}$ for the perturbed latent dynamics. Recall that $\mathcal{O}^{\bar{d}}$ contains the values of $o(t) + \delta(t)$ for almost all t. Since $\mathcal{O}$ is compact, any $o' \in \mathcal{O}^{\bar{d}}$ can be written $o' = o + \delta_0$ with $o \in \mathcal{O}$ and $|\delta_0| = \mathrm{dist}(o', \mathcal{O}) \leq \bar{d}$. Applying (97) on the boundary $\partial\mathcal{Z} = \{|z| = R_z\}$, and using $\bar{g} \leq R_z/2$ and $\bar{L}_{go}\bar{d} \leq R_z/4$, we obtain

$$z^\top g_\Omega(z, o') \leq R_z \left(\frac{R_z}{2} + \frac{R_z}{4} - R_z\right) = -\frac{R_z^2}{4}, \qquad \forall z \in \partial\mathcal{Z},\ \forall o' \in \mathcal{O}^{\bar{d}}. \tag{98}$$

Since (98) holds for all o' in the compact set $\mathcal{O}^{\bar{d}}$, in which the input $o(t)+\delta(t)$ takes its values for almost all t, Lemma B.2, applied with $G=\kappa g_\Omega, \mathcal{U}=\mathcal{O}^{\bar{d}}$, $R=R_z$ and $u(t)=o(t)+\delta(t)$, shows that z_d is defined on $[0,+\infty)$ and remains in $\mathcal{Z}$. This proves that $\mathcal{Z}$ is forward invariant for the perturbed latent dynamics (95), which is the first claim of Proposition 4.3.

Step 2: the perturbed and nominal latent trajectories remain close. By Proposition 3.3 and Step 1, $z(t)$ and $z_d(t)$ belong to $\mathcal{Z}$ for all $t\geq 0$. Let $\Delta(t)=z_d(t)-z(t)$, which is absolutely continuous and satisfies $\Delta(0)=0$. For almost all $t\geq 0$, we have

$$\frac{d\Delta}{dt}(t)=\kappa\left[g_\Omega(z_d(t),o(t))-g_\Omega(z(t),o(t))\right]+\kappa\left[g_\Omega(z_d(t),o(t)+\delta(t))-g_\Omega(z_d(t),o(t))\right].$$

As in Appendix C, the integral mean value theorem allows to write the first term as $A(t)\Delta(t)$, with

$$A(t)=\kappa\int_0^1\frac{\partial g_\Omega}{\partial z}\left(z(t)+s\Delta(t),o(t)\right)ds.$$

Since $\mathcal{Z}$ is convex, $z(t)+s\Delta(t)\in\mathcal{Z}$ for all $s\in[0,1]$, and since $o(t)\in\mathcal{O}$, Assumption 3.2 gives $A(t)+A(t)^\top\leq -2\kappa I_m$. We denote the second term by $B(t)$. Since $z_d(t)\in\mathcal{Z}$, $o(t)\in\mathcal{O}$ and $|\delta(t)|\leq\bar{d}$, inequality (96) gives, for almost all $t\geq 0$,

$$|B(t)|\leq\kappa\bar{L}_{go}|\delta(t)|.$$

Hence Δ is absolutely continuous and satisfies $\frac{d\Delta}{dt}=A(t)\Delta+B(t)$ for almost all t, with A continuous, B measurable and essentially bounded, $A(t)+A(t)^\top\leq -2\kappa I_m$ and $\Delta(0)=0$. Lemma B.1, with $a_+=-\kappa$, gives

$$|\Delta(t)|\leq\int_0^t e^{-\kappa(t-s)}|B(s)|\,ds\leq\kappa\bar{L}_{go}\int_0^t e^{-\kappa(t-s)}|\delta(s)|\,ds.$$

By definition of $d(t)$, $e^{-\kappa(t-s)/2}|\delta(s)|\leq d(t)$ for almost all $s\in[0,t]$, so that $e^{-\kappa(t-s)}|\delta(s)|\leq e^{-\kappa(t-s)/2}d(t)$, and

$$|z_d(t)-z(t)|\leq\kappa\bar{L}_{go}\,d(t)\int_0^t e^{-\kappa(t-s)/2}\,ds\leq 2\bar{L}_{go}\,d(t),\qquad\forall t\geq 0.\tag{99}$$

More generally, any rate $\lambda\in(0,\kappa)$ in place of $\kappa/2$ in the definition of $d(t)$ gives the same bound with the factor 2 replaced by $\kappa/(\kappa-\lambda)$. Note that (99) does not depend on the initial condition z_0, and that $d(t)\leq\|\delta\|_\infty$.

Step 3: from the latent error to the estimation error. We first bound the distance between the perturbed latent state and the image of the physical state by $\boldsymbol{T}_\kappa$. By the triangle inequality,

$$|z_d(t)-\boldsymbol{T}_\kappa(x(t))|\leq|z_d(t)-z(t)|+|z(t)-\boldsymbol{T}_\kappa(x(t))|.$$

The first term is bounded by (99). Since $z(t)$ is the nominal latent trajectory starting from $z_0\in\mathcal{Z}$, the attractivity property of Theorem 3.4, which holds for $\kappa\geq\underline{\kappa}$ by Theorem 3.7, bounds the second term by $e^{-\kappa t}|z_0-\boldsymbol{T}_\kappa(x_0)|$. Hence

$$|z_d(t)-\boldsymbol{T}_\kappa(x(t))|\leq e^{-\kappa t}|z_0-\boldsymbol{T}_\kappa(x_0)|+2\bar{L}_{go}\,d(t),\qquad\forall t\geq 0.\tag{100}$$

We now go back to the physical state. Since $\kappa\geq\underline{\kappa}$, Theorem 3.7 applies, and the extended inverse $\boldsymbol{T}_\kappa^{-1}:\mathbb{R}^m\to\mathbb{R}^n$ constructed in Appendix F is defined on the whole latent space and satisfies the global Lipschitz property (93) with constant $L_{inv}=\sqrt{n}\,c_T\kappa^\ell=c_T'\kappa^\ell$. Moreover, it is an exact left inverse on $\mathcal{X}$, namely $\boldsymbol{T}_\kappa^{-1}(\boldsymbol{T}_\kappa(x))=x$ for all $x\in\mathcal{X}$. Since $z_d(t)\in\mathcal{Z}$ by Step 1, the decoder satisfies $|\gamma_\Omega(z_d(t))-\boldsymbol{T}_\kappa^{-1}(z_d(t))|\leq\epsilon_{dec}$; this is where Step 1 is essential, the decoder being only guaranteed to be accurate on $\mathcal{Z}$. Since $x(t)\in\mathcal{X}$ for all $t\geq 0$, the triangle inequality gives

$$\begin{aligned}|\hat{x}(t)-x(t)|&\leq|\gamma_\Omega(z_d(t))-\boldsymbol{T}_\kappa^{-1}(z_d(t))|+|\boldsymbol{T}_\kappa^{-1}(z_d(t))-\boldsymbol{T}_\kappa^{-1}(\boldsymbol{T}_\kappa(x(t)))|\\&\leq\epsilon_{dec}+c_T'\kappa^\ell\,|z_d(t)-\boldsymbol{T}_\kappa(x(t))|.\end{aligned}$$

Combining with (100) gives (18), which concludes the proof of Proposition 4.3.

Finally, the decoder constructed in the proof of Theorem 4.2 satisfies $\epsilon_{dec}\leq\epsilon/2$, with a gain such that $2R_zc_T'\kappa^\ell e^{-\kappa t_o}\leq\epsilon/2$. Since $|z_0-\boldsymbol{T}_\kappa(x_0)|\leq 2R_z$, (18) then gives $|\gamma_\Omega(z_d(t))-x(t)|\leq\epsilon+2c_T'\kappa^\ell\bar{L}_{go}\,d(t)$ for all $t\geq t_o$, as stated after Proposition 4.3.

G Discrete-Time Implementation and Euler Error (Theorem 5.3)

In this section, we explicitly extend our continuous-time analysis to the discrete-time residual RNN implementation and prove Theorem 5.3.

G.1 RNN Structure as a Residual Network

The theoretical continuous-time observer is defined by $\dot{z} = \kappa g_\Omega(z, o)$. Its discrete implementation via an explicit Euler scheme is given by:

$$z_{k+1} = z_k + \delta\kappa g_\Omega(z_k, o_k) = \mathrm{RNN}_\Omega(z_k, o_k). \tag{101}$$

This equation corresponds exactly to the structure of a Residual RNN, by setting z_k as the hidden state traversing the sequence layers.

Example: Let $\sigma(\cdot) = \tanh(\cdot)$ be the activation function. Since $|\tanh(v)| \leq 1$ element-wise, the nonlinearity is bounded by $\sqrt{m}$ in Euclidean norm. We explicitly define the compact operating set $\mathcal{Z}$ as the closed ball of radius $R_z > \sqrt{m}$:

$$\mathcal{Z} = \{z \in \mathbb{R}^m \mid |z| \leq R_z\}.$$

Consider the base dynamics:

$$g_\Omega(z, o) = -z + \tanh(Wz + Uo + b).$$

For any z on the boundary of $\mathcal{Z}$ (i.e., $|z| = R_z$), the linear dissipative term $-z$ strictly dominates the bounded term $\tanh(\dots)$, ensuring that the continuous vector field points inward. Thus, $\mathcal{Z}$ is forward invariant. The discrete update becomes:

$$z_{k+1} = (1 - \delta\kappa)z_k + \delta\kappa \tanh(Wz_k + Uo_k + b).$$

The weights $\Omega = (W, U, b)$ are then chosen to ensure the strict monotonicity condition (see (7)) for the function $z \in \mathcal{Z} \mapsto \tanh(Wz + Uo + b)$.

The physical state estimate is finally given by the MLP Decoder:

$$\hat{x}_k = \gamma_\Omega(z_k) \approx \boldsymbol{T}_\kappa^{-1}(z_k). \tag{102}$$

G.2 Proof of Tunability of the Residual RNN

To prove Theorem 5.3, we decompose the total discrete estimation error into two components: the continuous observer tracking error (which decreases with κ) and the numerical discretization error (which increases with κ but decreases with δ).

Let $z(t)$ be the exact solution of the continuous observer $\dot{z} = \kappa g_\Omega(z, o(t))$ initialized at z_0. The overall estimation error using the exact left-inverse map at discrete time $t_k = k\delta$ is:

$$\begin{aligned} |\hat{x}_k - x(t_k)| &= |\boldsymbol{T}_\kappa^{-1}(z_k) - \boldsymbol{T}_\kappa^{-1}(\boldsymbol{T}_\kappa(x(t_k)))| \\ &\leq c_T \kappa^\ell |z_k - \boldsymbol{T}_\kappa(x(t_k))|, \end{aligned} \tag{103}$$

where we used the global Lipschitz property of the extended inverse $\boldsymbol{T}_\kappa^{-1}$ (established in Appendix F). We introduce the exact continuous solution $z(t_k)$ to separate the error via the triangle inequality:

$$|z_k - \boldsymbol{T}_\kappa(x(t_k))| \leq \underbrace{|z_k - z(t_k)|}_{e_{num,k}\text{ (Numerical Error)}} + \underbrace{|z(t_k) - \boldsymbol{T}_\kappa(x(t_k))|}_{e_{obs}(t_k)\text{ (Observer Error)}}. \tag{104}$$

Step 1: Continuous Observer Error Bound

According to Theorem 3.4, the exact continuous observer error decays exponentially:

$$e_{obs}(t_k) \leq e^{-\kappa t_k} |z_0 - \boldsymbol{T}_\kappa(x_0)| \leq C_0 e^{-\kappa t_k}, \tag{105}$$

where $C_0 = 2R_z$ is the maximum distance in $\mathcal{Z}$.

Step 2: Numerical Error Bound (Euler Discretization)

We analyze the propagation of the global numerical error defined as $e_{num,k} = z_k - z(t_k)$.

a) **Stability (Discrete-time Contraction).** Let g_Ω be contracting (satisfying Assumption 3.2) and L_z-Lipschitz with respect to the latent state. Consider two discrete trajectories initialized at $z_{a,k}$ and $z_{b,k}$, driven by the same observation o_k. Let $\Delta z_k = z_{a,k} - z_{b,k}$ and $\Delta g_k = g_\Omega(z_{a,k}, o_k) - g_\Omega(z_{b,k}, o_k)$. The squared norm of their difference evolves as:

$$\begin{aligned}|z_{a,k+1} - z_{b,k+1}|^2 &= |(z_{a,k} - z_{b,k}) + \delta\kappa(g_\Omega(z_{a,k}, o_k) - g_\Omega(z_{b,k}, o_k))|^2 \\ &= |\Delta z_k|^2 + 2\delta\kappa\langle \Delta z_k, \Delta g_k\rangle + \delta^2\kappa^2|\Delta g_k|^2.\end{aligned}$$

Using the strong monotonicity property $\langle \Delta z_k, \Delta g_k\rangle \le -|\Delta z_k|^2$ and the Lipschitz property $|\Delta g_k| \le L_z|\Delta z_k|$:

$$|z_{a,k+1} - z_{b,k+1}|^2 \le (1 - 2\delta\kappa + \delta^2\kappa^2 L_z^2)|\Delta z_k|^2.$$

To ensure discrete-time stability, we enforce the condition $\delta\kappa L_z^2 \le 1$. Under this constraint, the coefficient satisfies $1 - 2\delta\kappa + \delta^2\kappa^2 L_z^2 \le (1 - \frac{1}{2}\delta\kappa)^2$. Taking the square root yields the discrete contraction inequality:

$$|\Delta z_{k+1}| \le \rho|\Delta z_k|, \quad \text{where } \rho = 1 - \frac{\delta\kappa}{2}. \tag{106}$$

b) **Consistency (Local Truncation Error).** We define the local truncation error ξ_k as the defect of the Euler step with respect to the exact continuous solution:

$$z(t_{k+1}) = z(t_k) + \delta\kappa g_\Omega(z(t_k), o(t_k)) + \xi_k.$$

Using a second-order Taylor expansion of the exact solution $z(t)$ at $t = t_k$:

$$z(t_{k+1}) = z(t_k) + \delta\dot{z}(t_k) + \frac{\delta^2}{2}\ddot{z}(\tau_k) \quad \text{for some } \tau_k \in [t_k, t_{k+1}].$$

Identifying terms with $\dot{z}(t_k) = \kappa g_\Omega(z(t_k), o(t_k))$, we obtain $|\xi_k| = \frac{\delta^2}{2}|\ddot{z}(\tau_k)|$. Differentiation of the dynamics yields:

$$\ddot{z}(t) = \frac{d}{dt}(\kappa g_\Omega(z, o(t))) = \kappa^2\frac{\partial g_\Omega}{\partial z}g_\Omega(z, o(t)) + \kappa\frac{\partial g_\Omega}{\partial o}\dot{o}(t).$$

By Assumption 3.1, the physical state $x(t)$ remains in the forward-invariant compact set $\mathcal{X}$. Consequently, both the physical vector field f and the observation Jacobian $\frac{\partial h}{\partial x}$ are uniformly bounded, which implies that the time derivative $\dot{o}(t) = \frac{\partial h}{\partial x}(x(t))f(x(t))$ is globally bounded for all $t \ge 0$. Since g_Ω and its partial derivatives are also bounded on the compact operating domain $\mathcal{Z} \times \mathcal{O}$, there exists a global constant $M > 0$ (independent of t and δ) such that for any gain $\kappa \ge 1$:

$$|\ddot{z}(t)| \le M\kappa^2.$$

Consequently, the local error is uniformly bounded by:

$$|\xi_k| \le \frac{M}{2}\kappa^2\delta^2.$$

c) **Global Bound (Finite Time).** The global numerical error $e_{num,k+1}$ satisfies the recurrence:

$$\begin{aligned}e_{num,k+1} &= z_{k+1} - z(t_{k+1}) \\ &= [z_k + \delta\kappa g_\Omega(z_k, o_k)] - [z(t_k) + \delta\kappa g_\Omega(z(t_k), o_k) + \xi_k]\,.\end{aligned}$$

Using the contraction property derived in step a):

$$|e_{num,k+1}| \le \rho|e_{num,k}| + |\xi_k|.$$

Assuming an exact initial condition ($e_{num,0} = 0$), expanding the recurrence gives:

$$|e_{num,k}| \le \sum_{j=0}^{k-1}\rho^{k-1-j}|\xi_j| \le \frac{M}{2}\kappa^2\delta^2\sum_{j=0}^{k-1}\rho^j.$$

Since $\rho < 1$, we can upper-bound the finite sum by the geometric series:

$$\sum_{j=0}^{k-1}\rho^j < \sum_{j=0}^{\infty}\rho^j = \frac{1}{1-\rho} = \frac{1}{\delta\kappa/2}.$$

Substituting this back into the error bound yields a result valid for all $k \ge 0$:

$$|e_{num,k}| \le \left(\frac{M}{2}\kappa^2\delta^2\right)\left(\frac{2}{\delta\kappa}\right) = M\kappa\delta. \tag{107}$$

Step 3: Synthesis of Tunability

Combining the Lipschitz bound (103), the continuous error (105), and the discrete numerical error (107), the total estimation error at step k is strictly bounded by:

$$|\hat{x}_k - x(t_k)| \leq c_T \kappa^\ell \left(\underbrace{C_0 e^{-\kappa t_k}}_{\text{Transient Error}} + \underbrace{M\kappa\delta}_{\text{Numerical Error}} \right).$$

Let us formally prove tunability for a fixed target precision $\epsilon > 0$ and observation time $t_o > 0$:

- **Choice of the continuous gain κ:** We choose a structural gain $\kappa \geq \underline{\kappa}$ large enough such that, for all $t_k \geq t_o$, the continuous transient error is sufficiently damped:
$$c_T C_0 \kappa^\ell e^{-\kappa t_o} \leq \frac{\epsilon}{2}.$$
This is always achievable because the exponential decay $e^{-\kappa t_o}$ dominates the polynomial growth κ^ℓ as $\kappa \to \infty$. We now fix this chosen gain κ.
- **Choice of the step size δ:** With κ fixed, we choose a maximum sampling step δ^* small enough to simultaneously satisfy the numerical stability condition ($\delta^* \leq \frac{1}{\kappa L_z^2}$) and strictly suppress the numerical truncation error:
$$c_T \kappa^\ell (M\kappa\delta^*) \leq \frac{\epsilon}{2} \implies \delta^* \leq \frac{\epsilon}{2M c_T \kappa^{\ell+1}}.$$

For any $\delta \leq \delta^*$, the sum of both errors is strictly bounded by ϵ for all $t_k \geq t_o$. Thus, there exists a configuration pair (κ, δ) enabling the achievement of arbitrary precision. The Residual RNN architecture is therefore formally tunable, concluding the proof of Theorem 5.3. □

H More details on the Robot pose estimation experiments (Section 6)

H.1 Supervision and losses

The standard per-step loss,

$$\mathcal{L}_{\text{step}} = \frac{1}{T} \sum_{t=1}^{T} \left\| \hat{y}_t - y_t \right\|^2, \tag{108}$$

treats each timestep independently. AMCL noise at step t contributes gradient signal only at step t. Consistent drift—the primary cause of trajectory-level error—is not amplified relative to per-step noise.

We propose a **trajectory integration loss** that directly penalises accumulated drift. The chunk is divided into non-overlapping windows of length H. Within each window, both the corrected and target body-frame displacements are integrated kinematically, and position MSE is computed on the resulting trajectories:

$$\mathcal{L}_{\text{traj}}(H) = \frac{1}{\lfloor T/H \rfloor} \sum_{k=0}^{\lfloor T/H \rfloor - 1} \text{MSE}\Big(\text{Integrate}\big(\hat{d}_{kH:kH+H}\big),\ \text{Integrate}\big(d^{\text{amcl}}_{kH:kH+H}\big) \Big) \tag{109}$$

The integration is implemented with `cumsum` over the heading increments, making it fully differentiable without a Python loop. The gradient at correction $\hat{y}_t$ therefore flows from all position errors at times $t, t+1, \ldots, kH + H - 1$: consistent drift that compounds forward is amplified, while zero-mean AMCL noise at individual steps tends to cancel across future loss terms.

The window length H controls the trade-off between signal and noise: short windows ($H \lesssim 20$) carry too little drift signal (AMCL and odometry agree at sub-metre scales); long windows ($H \gtrsim 100$) accumulate enough AMCL noise in the integrated target to overwhelm the gradient signal. Note that $H = 1$ recovers $\mathcal{L}_{\text{step}}$ exactly.

Training. Training uses sliding-window chunks of length 300 timesteps (stride 150). The GRU hidden state is reset to zero at each chunk boundary. Chunks containing AMCL relocalization jumps (consecutive AMCL displacement > 0.5 m) are excluded. We use Adam with initial learning rate

4×10^{-4}, cosine annealing with minimum learning rate 4×10^{-5} over up to 2000 epochs, batch size 1024, and gradient clipping at norm 1.0.

Checkpoint selection. Validation loss (per-step MSE on held-out chunks) and trajectory RMSE are anti-correlated after the training optimum: a model that continues to minimise per-step loss is fitting AMCL noise at individual timesteps, producing erratic corrections that accumulate into larger trajectory errors. We therefore select checkpoints by trajectory RMSE evaluated every 10 epochs, not by validation loss.

H.2 Network architectures: "Contracting GRU"

The standard GRU provides no guarantee on the long-term behaviour of the hidden state: two copies of the network initialised at different hidden states may not converge under the same input sequence. Motivated by the KKL observer framework of this paper, we introduce a **"Contracting GRU"**, which replaces the GRU cell with a provably contracting residual update.

Architecture. Let $z_k \in \mathbb{R}^m$ denote the hidden state at step k and $o_k \in \mathbb{R}^9$ the input. The update rule is

$$r_k = \sigma(W_r z_k + U_r o_k + b_r), \tag{110}$$
$$n_k = \tanh(W_n z_k + U_n o_k + b_n), \tag{111}$$
$$g_k = -\alpha\, z_k + r_k \odot n_k, \tag{112}$$
$$z_{k+1} = z_k + \delta\kappa\, g_k, \tag{113}$$

where r_k is a GRU-style reset gate, n_k is a candidate state, and g_k is the monotonic velocity field. The scalar $\alpha > 0$ is a fixed dissipation coefficient; δ and κ are the sampling period and gain inherited from the KKL framework. The reset gate allows g to depend on the current input in a richer way than the plain KKL cell, while keeping the contraction analysis tractable.

Contraction guarantee. To apply the tools of this paper, we define the associated continuous-time latent dynamics given by

$$g_\Omega(z, o) = -\alpha\, z + r_\Omega(z, o) \odot n_\Omega(z, o) \tag{114}$$

where Ω stores the weights $W_r, W_n, U_r, U_n, b_r, b_n$ and

$$r_\Omega(z, o) = \sigma(W_r z + U_r o + b_r) \quad , \quad n_\Omega(z, o) = \tanh(W_n z + U_n o + b_n).$$

The Jacobian of g_Ω with respect to z is

$$\frac{\partial g_\Omega}{\partial z} = -\alpha I + \operatorname{diag}(r_\Omega)\, D_n\, W_n + \operatorname{diag}(n_\Omega) \operatorname{diag}(r_\Omega \odot (1 - r_\Omega))\, W_r,$$

where $D_n = \operatorname{diag}(\operatorname{sech}^2(\cdot))$. Using $\|D_n\|_{\text{op}} \le 1$, $r_i(1 - r_i) \le 1/4$, and $|n_i| \le 1$, the symmetric part satisfies

$$\frac{\partial g_\Omega}{\partial z} + \left(\frac{\partial g_\Omega}{\partial z}\right)^\top \le \left(-2\alpha + 2\|W_n\|_2 + \tfrac{1}{2}\|W_r\|_2\right) I. \tag{115}$$

We enforce $\|W_n\|_2 \le 1$ and $\|W_r\|_2 \le 1$ throughout training via spectral normalisation. Setting $\alpha = 2.5$ gives an upper bound of $-2.5\, I$ on the symmetric part, so Assumption 3.2 of the main paper is globally satisfied by construction, namely for any values of the latent state and inputs (z, o).

Empirical contraction constant. The bound (115) is an a priori certificate; we also measured the quantity our theory actually relies on. To avoid any confusion between conventions, we express everything in terms of the largest eigenvalue $\lambda_{\max}$ of the symmetric part $\frac{1}{2}\left(\frac{\partial g_\Omega}{\partial z} + \left(\frac{\partial g_\Omega}{\partial z}\right)^\top\right)$ of the latent Jacobian. In this convention, Assumption 3.2 requires $\lambda_{\max} \le -1$, and the a priori bound (115) with $\alpha = 2.5$ gives $\lambda_{\max} \le -1.25$. Computed by automatic differentiation along all validation trajectories (8 trajectories, 26,888 time steps, latent dimension 128, observation dimension 9), $\lambda_{\max}$ for the first layer of the trained Contracting GRU averages -2.18 (standard deviation 0.015, worst case -2.16 over the whole dataset); the second, stacked layer behaves similarly (mean -2.19). The monotonicity condition therefore holds strictly at 100% of the time steps, with a margin of about 0.9 with respect to the a priori bound, and of about 1.2 with respect to the requirement of Assumption 3.2. The discrete residual RNN inherits the corresponding guarantees through the small-step analysis of Appendix G: the step size $\delta\kappa = 1/3$ used in our experiments (Section H.4) keeps the discrete update in the regime where this continuous-time monotonicity is preserved.

Stacking layers. A second contracting cell can be stacked on top of the first by using $z_k^{(0)}$ as the input to layer 1 in place of o_k:

$$z_{k+1}^{(0)} = z_k^{(0)} + \delta\kappa\, g_k^{(0)}, \tag{116}$$

$$z_{k+1}^{(1)} = z_k^{(1)} + \delta\kappa\, g_k^{(1)}, \tag{117}$$

where $g^{(0)}$ uses $(z_k^{(0)}, o_k)$ and $g^{(1)}$ uses $(z_k^{(1)}, z_k^{(0)})$. In fact, denoting g_Ω the single-layer latent base map given by (114) and globally verifying Assumption 3.2, we claim that the two-layer architecture corresponding to a map g_Ω^{tot} built as

$$g_\Omega^{\mathrm{tot}}(z, o) = (g_{\Omega_0}(z^0, o), g_{\Omega_1}(z^1, z^0))$$

where $z = (z^0, z^1)$ stacks the latent state of each cell, $\Omega = (\Omega_0, \Omega_1)$ stores the weights of each layer, still falls into the scope of this paper. This is done by operating a change of coordinates on the latent states and a scaling in κ. Indeed, we have

$$\frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o) = \begin{pmatrix} \frac{\partial g_{\Omega_0}}{\partial z}(z^0, o) & 0 \\ \frac{\partial g_{\Omega_1}}{\partial o}(z^1, z^0) & \frac{\partial g_{\Omega_1}}{\partial z}(z^1, z^0) \end{pmatrix}$$

which implies

$$\frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o) + \frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o)^\top = \begin{pmatrix} \frac{\partial g_{\Omega_0}}{\partial z}(z^0, o) + \frac{\partial g_{\Omega_0}}{\partial z}(z^0, o)^\top & \frac{\partial g_{\Omega_1}}{\partial o}(z^1, z^0)^\top \\ \frac{\partial g_{\Omega_1}}{\partial o}(z^1, z^0) & \frac{\partial g_{\Omega_1}}{\partial z}(z^1, z^0) + \frac{\partial g_{\Omega_1}}{\partial z}(z^1, z^0)^\top \end{pmatrix}$$

From the global monotonicity of g_Ω,

$$\frac{\partial g_{\Omega_0}}{\partial z}(z^0, o) + \frac{\partial g_{\Omega_0}}{\partial z}(z^0, o)^\top \leq -2I \quad , \qquad \frac{\partial g_{\Omega_1}}{\partial z}(z^1, z^0) + \frac{\partial g_{\Omega_1}}{\partial z}(z^1, z^0)^\top \leq -2I$$

which is not enough to imply the monotonicity of g_Ω^{tot}. However, notice that $\frac{\partial g_\Omega}{\partial o}$ in (114) is bounded and consider $\mu \geq 1$ such that

$$\mu \geq \max_{(z,o)} \left| \frac{\partial g_{\Omega_1}}{\partial o}(z, o) \right|.$$

Then, defining $P = \begin{pmatrix} \mu^2 I_m & 0 \\ 0 & I_m \end{pmatrix}$ (with I_m the identity matrix of size $m \times m$), we have

$$P\frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o) + \frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o)^\top P$$
$$= \begin{pmatrix} \mu^2 \left[\frac{\partial g_{\Omega_0}}{\partial z}(z^0, o) + \frac{\partial g_{\Omega_0}}{\partial z}(z^0, o)^\top \right] & \frac{\partial g_{\Omega_1}}{\partial o}(z^1, z^0)^\top \\ \frac{\partial g_{\Omega_1}}{\partial o}(z^1, z^0) & \frac{\partial g_{\Omega_1}}{\partial z}(z^1, z^0) + \frac{\partial g_{\Omega_1}}{\partial z}(z^1, z^0)^\top \end{pmatrix}$$

so that, for all $\xi = (\xi_0, \xi_1) \in \mathbb{R}^{2m}$,

$$\begin{aligned} \begin{bmatrix} \xi_0 \\ \xi_1 \end{bmatrix}^\top \left[P\frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o) + \frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o)^\top P \right] \begin{bmatrix} \xi_0 \\ \xi_1 \end{bmatrix} & \\ \leq -2\mu^2\|\xi_0\|^2 + 2\left| \frac{\partial g_{\Omega_1}}{\partial o}(z^1, z^0) \right| \|\xi_0\|\|\xi_1\| - 2\|\xi_1\|^2 & \\ \leq -2\mu^2\|\xi_0\|^2 + 2\mu\|\xi_0\|\|\xi_1\| - 2\|\xi_1\|^2 & \\ \leq -\mu^2\|\xi_0\|^2 - \|\xi_1\|^2 = -\xi^\top P \xi, & \end{aligned}$$

hence

$$P\frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o) + \frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}(z, o)^\top P \leq -P \qquad \forall (z, o).$$

(the last equality follows from $P = \mathrm{diag}(\mu^2 I_m, I_m)$). Defining the new latent coordinates $\tilde{z} = P^{1/2} z$ and the new scaling parameter $\tilde{\kappa} = \frac{\kappa}{2}$, the latent continuous-time dynamics can be written as $\dot{\tilde{z}} = \tilde{\kappa}\tilde{g}(\tilde{z}, o)$ where

$$\tilde{g}_\Omega^{\mathrm{tot}}(\tilde{z}, o) = 2P^{1/2} g_\Omega^{\mathrm{tot}}(P^{-1/2}\tilde{z}, o)$$

which verifies, since $\frac{\partial \tilde{g}_\Omega^{\mathrm{tot}}}{\partial \tilde{z}} + \frac{\partial \tilde{g}_\Omega^{\mathrm{tot}}}{\partial \tilde{z}}^\top = 2P^{-1/2}\big(P\frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z} + \frac{\partial g_\Omega^{\mathrm{tot}}}{\partial z}^\top P\big)P^{-1/2}$,

$$\frac{\partial \tilde{g}_\Omega^{\mathrm{tot}}}{\partial \tilde{z}}(\tilde{z}, o) + \frac{\partial \tilde{g}_\Omega^{\mathrm{tot}}}{\partial \tilde{z}}(\tilde{z}, o)^\top \leq -2I_{2m} \qquad \forall (\tilde{z}, o).$$

In other words, the two-layered contracting GRU is indeed globally monotonic, modulo a change of coordinates on z and a rescaling of κ. This reasoning can be reproduced to handle multi-layered architecture where each cell is globally monotonic and with bounded derivatives with respect to the input o.

H.3 Dataset details

Table 1: Dataset statistics per building: we collected data in 14 floors of 4 buildings distributed over 2 continents. Distance is path length along the odometry trajectory. Floors indicates distinct floor levels covered in each building.

Split	Building	Continent	Floors	Runs	Steps	Duration	Distance
Train	Building-17	A	7	51	83 466	7h 43m 42s	14.1 km
Train	Building-GF	A	3	11	19 105	1h 46m 08s	4.5 km
Train	Building-Q (1F, 2F)	B	2	16	43 140	3h 59m 39s	3.8 km
Val	Building-Q (3F)	B	1	8	26 888	2h 29m 22s	2.0 km
Test	Building-CC	B	1	14	24 113	2h 13m 57s	3.1 km
Total			14	100	196 712	18h 12m 49s	27.4 km

The validation floor (Building-Q) is entirely unseen during training. The full test building is unseen during training and validation. The recordings from Buildings-17 and Building-GF are from a large multi-floor office building and introduce substantially more diverse motion patterns.

In these data, $\delta = 1/3s$. Also, we took $\kappa = 1$.

H.4 Training hyper-parameters

Batch size	1024
Learning rate	10^{-4}
Minimum LR	10^{-5}
Scheduler	Cosine annealing
Optimizer	Adam
Gradient clipping	max norm 1.0
Epochs	2000
Chunk length	300 steps (100 s)
Chunk stride	150 steps (50 s)
Traj. loss horizon	$H = 50$ steps (17 s, $\approx$5 m)
Min. speed filter	0.05 m/s
Hidden dimension	128
GRU layers	2
Input dimension	9
Gain / step (Contracting GRU)	$\kappa = 1$, $\delta = 1/3$, $\alpha = 2.5$

H.5 Empirical filter rank analysis

We empirically check whether trained models satisfy the Filter Rank and Capacity Condition of Assumption 3.6 along realistic robot trajectories.

We consider the involved biases are visible from the input signal o and its first time derivative, i.e., the differential observability condition in Assumption 3.5 holds at order $\ell = 1$ (velocity biases). However, notice that the physical system is not autonomous in this experiment due to trajectory control. As usual in control theory, inputs may be assumed to be generated by an autonomous system of sufficiently large dimension. Hence, we may need more time derivatives of o to recover all the missing information, i.e., $\ell > 1$. We evaluate the matrix $C_\ell(o)$ at sampled operating points and compute its column rank as a function of ℓ.

Setup. Let $f_\Omega(z, o) = z + \delta\kappa\, g_\Omega(z, o)$ be the one-step recurrence of the trained encoder, with joint hidden state $z \in \mathbb{R}^m$ obtained by concatenating the per-layer hidden vectors of the two-layer architecture (so $m = 256$, twice the per-layer dimension of 128). For each sampled operating point $o^* \in \mathcal{O}$ drawn from validation trajectories:

1. we find the equilibrium z^* such that $f_\Omega(z^*, o^*) = z^*$ by iterating the recurrence with constant input until $\|z_{k+1} - z_k\|_2 < 10^{-7}$. The contraction property of Assumption 3.2 guarantees this for the Contracting GRU;
2. we compute the local Jacobians by reverse-mode automatic differentiation,
$$A(o^*) = \frac{\partial f_\Omega}{\partial z}(z^*, o^*) \in \mathbb{R}^{m\times m}, \qquad B(o^*) = \frac{\partial f_\Omega}{\partial o}(z^*, o^*) \in \mathbb{R}^{m\times p}; \tag{118}$$
3. we build, for each $\ell \in \{1, \dots, 15\}$,
$$C_\ell(o^*) = \left[\, B(o^*) \,\middle|\, A(o^*)B(o^*) \,\middle|\, \dots \,\middle|\, A(o^*)^\ell B(o^*) \,\right] \in \mathbb{R}^{m\times p(\ell+1)}, \tag{119}$$
and compute its numerical column rank as $r_\ell(o^*) = |\{i : \sigma_i(C_\ell) > \tau\, \sigma_{\max}(C_\ell)\}|$ with $\tau = 10^{-6}$.

Result. We discover that the Filter Rank and Capacity Condition is fulfilled up to $\ell = 4$.

I Compute resources

The trained models are small and can even be trained on a laptop CPU in a couple of hours. We trained them on V100 GPUs and did not bother our A100 and H100 GPUs. Training one single model is typically done in a couple of minutes (under half an hour), but it took longer to generate the hyper-parameter sweeps of the paper.

Real experiments (pose estimation). Figure 4 comparing 3 models on 5 training seeds each took 2.5h to calculate on a single V100. The Contracting GRU took a bit longer (3x or 4x longer) since the implementation is less efficient than the optimized PyTorch GRU.

Simulated experiments. Figures 5 and 6 required training 144 models (6 values of κ times 8 decoder widths times 3 seeds for each model) which took around 30 hours.

J More details on the simulation experiments (Section 6)

Let us define a compact forward invariant set $\mathcal{X} \subset \mathbb{R}^3$ of trajectories of interest on which the differential observability Assumption 3.5 is satisfied. Let $\mathcal{X}_3$ be an arbitrary compact set in $(0, +\infty)$. Let $r_0 = \sqrt{\min \mathcal{X}_3}$ and $r_1 = \sqrt{\max \mathcal{X}_3}$. Consider the function $V : \mathbb{R}^2 \to \mathbb{R}$ defined by $V(x) = (x_1^2 + x_2^2)$. Let $\mathcal{D}_0 = \{(x_1, x_2) \in \mathbb{R}^2 \mid V(x_1, x_2) < r_0^2\}$ and $\mathcal{D}_1 = \{(x_1, x_2) \in \mathbb{R}^2 \mid V(x_1, x_2) \leq r_1^2\}$. Define $\mathcal{X} = (\mathcal{D}_1 \setminus \mathcal{D}_0) \times \mathcal{X}_3$. Then $\mathcal{X}$ is compact and forward invariant. Indeed, for all solutions to (27), $\frac{d}{dt}V(x_1, x_2) = 2(x_3 - x_1^2 - x_2^2)x_2^2$ is non-negative if $(x_1, x_2) \in \mathcal{D}_0$ and non-positive if $(x_1, x_2) \in \mathbb{R}^2 \setminus \mathcal{D}_1$.

Now, we prove that Assumption 3.5 is satisfied on $\mathcal{X}$, with $\ell = 3$. Computing the successive Lie derivatives of $h(x) = x_1$ with respect to the vector field f of (27), we obtain for all $x \in \mathcal{X}$:
$$\mathcal{L}_f h(x) = x_2, \qquad \mathcal{L}_f^2 h(x) = (x_3 - x_1^2 - x_2^2)x_2 - x_1.$$
Hence
$$x_3 x_2 = (\mathcal{L}_f^2 h + h + (h^2 + (\mathcal{L}_f h)^2)\mathcal{L}_f h)(x).$$
Moreover
$$\mathcal{L}_f(\mathcal{L}_f^2 h + h + (h^2 + (\mathcal{L}_f h)^2)\mathcal{L}_f h)(x) = x_3 \mathcal{L}_f^2 h(x).$$
Hence
$$x_3(x_3x_2 - x_1) = (\mathcal{L}_f + h^2 + (\mathcal{L}_f h)^2)(\mathcal{L}_f^2 h + h + (h^2 + (\mathcal{L}_f h)^2)\mathcal{L}_f h)(x).$$
The above equations define a polynomial map $P : \mathbb{R}^4 \to \mathbb{R}^4$ such that, for all $x \in \mathbb{R}^3$, $(x_1, x_2, x_3x_2, x_3(x_3x_2 - x_1)) = P(h(x), \mathcal{L}_f h(x), \mathcal{L}_f^2 h(x), \mathcal{L}_f^3 h(x))$.

Moreover, since $0 < x_3 \leq r_1^2$ and $2r_1^2|x_1x_2| \leq \frac{1}{2}x_1^2 + 2r_1^4 x_2^2$, for all $x \in \mathcal{X}$,
$$(2r_1^4 + \frac{1}{2})x_2^2 + (x_3x_2 - x_1)^2 \geq (2r_1^4 + \frac{1}{2})x_2^2 + x_1^2 - 2r_1^2|x_1x_2| \geq \frac{1}{2}(x_1^2 + x_2^2) \geq \frac{1}{2}r_0^2.$$
Hence,
$$x_2^2 + (x_3x_2 - x_1)^2 \geq \min\left\{1, \frac{1}{2r_1^4 + \frac{1}{2}}\right\}\frac{1}{2}r_0^2.$$

Thus the polynomial map $I : \mathcal{X} \to \mathbb{R}^4$ defined by $I(x_1, x_2, x_3) = (x_1, x_2, x_3x_2, x_3(x_3x_2 - x_1))$ admits a globally Lipschitz left-inverse $J : I(\mathcal{X}) \to \mathbb{R}^3$ defined by $J(x_1, x_2, a_1, a_2) = (x_1, x_2, \frac{x_2a_1+(a_1-x_1)a_2}{x_2^2+(a_1-x_1)^2})$.

To conclude, for all $x \in \mathcal{X}$, $x = J(P(h(x), \mathcal{L}_f h(x), \mathcal{L}_f^2 h(x), \mathcal{L}_f^3 h(x)))$. Since $J \circ P$ is globally Lipschitz over the compact set $h(\mathcal{X}) \times \mathcal{L}_f h(\mathcal{X}) \times \mathcal{L}_f^2 h(\mathcal{X}) \times \mathcal{L}_f^3 h(\mathcal{X})$, system (27) is differentially observable of order $\ell = 3$. In other words, (10) is verified, with c_{obs} the Lipschitz constant of $J \circ P$.

K About Assumption 3.6

This appendix provides a formal justification for the Filter Rank and Capacity Condition, demonstrating that the left-invertibility requirement on the matrix C_ℓ is a weak condition.

To illustrate how the Filter Rank and Capacity Condition can be structurally satisfied, let us consider a scenario where the physical observation is scalar ($p = 1$), the required observability index is $\ell = 2$, and the latent space dimension is $m = 4$. Aligning with standard deep learning conventions where the input feeds into the first layer, we design the base latent dynamics as a lower-triangular causal cascade:

$$\begin{aligned}
\dot{z}_1 &= -\alpha_1 z_1 + \sigma_1(w_{11} z_1 + u_1 o + b_1) \\
\dot{z}_2 &= -\alpha_2 z_2 + \sigma_2(u_2 z_1 + w_{22} z_2 + b_2) \\
\dot{z}_3 &= -\alpha_3 z_3 + \sigma_3(u_3 z_2 + w_{33} z_3 + b_3) \\
\dot{z}_4 &= -\alpha_4 z_4 + \sigma_4(u_4 z_3 + w_{44} z_4 + b_4)
\end{aligned}$$

where the activation functions σ_i are strictly monotonic (hence $\sigma_i' > 0$ everywhere) and the coupling weights u_i are strictly non-zero. By evaluating the Jacobians at the equilibrium, we obtain the matrices $A \in \mathbb{R}^{4\times4}$ and $B \in \mathbb{R}^{4\times1}$:

$$A = \begin{bmatrix} -\alpha_1 + \sigma_1' w_{11} & 0 & 0 & 0 \\ \sigma_2' u_2 & -\alpha_2 + \sigma_2' w_{22} & 0 & 0 \\ 0 & \sigma_3' u_3 & -\alpha_3 + \sigma_3' w_{33} & 0 \\ 0 & 0 & \sigma_4' u_4 & -\alpha_4 + \sigma_4' w_{44} \end{bmatrix}, \quad B = \begin{bmatrix} \sigma_1' u_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

To highlight the algebraic rank properties, let 0 denote exactly zero, $\times$ denote a strictly non-zero element, and $\star$ represent an arbitrary scalar. The matrices exhibit the following structured sparsity pattern:

$$A = \begin{bmatrix} \star & 0 & 0 & 0 \\ \times & \star & 0 & 0 \\ 0 & \times & \star & 0 \\ 0 & 0 & \times & \star \end{bmatrix}, \quad B = \begin{bmatrix} \times \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

This causal structure naturally maps to a lower-triangular rank progression. The required $(4, 3)$ matrix $C_2 = \begin{bmatrix} B & AB & A^2B \end{bmatrix}$ becomes:

$$C_2 = \begin{bmatrix} \sigma_1' u_1 & \star & \star \\ 0 & \sigma_1' u_1 \sigma_2' u_2 & \star \\ 0 & 0 & \sigma_1' u_1 \sigma_2' u_2 \sigma_3' u_3 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \times & \star & \star \\ 0 & \times & \star \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{bmatrix}$$

Because the upper 3×3 submatrix of C_2 is upper-triangular, if the u_i's are non zero, C_2 maintains full column rank uniformly over the observation domain. This explicitly guarantees the left-invertibility required by Assumption 3.6.

While this specific cascaded architecture provides a constructive proof of existence, we conjecture that this rank requirement is not a restrictive architectural constraint. In fact, left-invertibility likely constitutes a generic topological property for a broad class of over-parameterized latent dynamics.

However, establishing such a general result involves complex differential topology arguments that fall beyond the scope of this paper. A first result that we can give is the following.

Proposition K.1 (Genericity for Single-Layer Analytic RNNs)**.** *Consider the latent dynamics of a single-layer RNN evaluated at a fixed observation point $o^* \in \mathbb{R}^p$. For simplicity of exposition, assume $p = 1$:*

$$\dot{z} = -\alpha z + \sigma(Wz + Uo^* + b), \tag{120}$$

where $z \in \mathbb{R}^m$, the parameter vector is $\Theta = (W, U, b) \in \mathbb{R}^{m\times m} \times \mathbb{R}^m \times \mathbb{R}^m$, and $\alpha > 0$. Assume that:

1. ***Analyticity:*** *The activation function σ is real-analytic and $\sigma'(x) > 0$ everywhere.*
2. ***Dimensionality:*** *The latent dimension satisfies $m \geq \ell + 1$.*

Then, the filter observability matrix $C_\ell(\Theta, o^*) \in \mathbb{R}^{m\times(\ell+1)}$ has full column rank for almost all parameter configurations Θ (i.e., outside a set of Lebesgue measure zero).

Proof. Let $v(\Theta) = Wz_{eq} + Uo^* + b$ denote the pre-activation state at the equilibrium z_{eq}. The local Jacobians evaluated at this equilibrium are:

$$A(\Theta) = -\alpha I_m + \Sigma'(v(\Theta))W, \tag{121}$$

$$B(\Theta) = \Sigma'(v(\Theta))U, \tag{122}$$

where $\Sigma'(v) = \mathrm{diag}(\sigma'(v_1), \ldots, \sigma'(v_m))$. Because σ is strictly monotonic, $\sigma'(v_i) > 0$ for all i.

The matrix $C_\ell(\Theta)$ has full column rank $\ell + 1$ if and only if at least one of its $(\ell + 1) \times (\ell + 1)$ minors has a non-zero determinant. Let $F(\Theta)$ be the sum of the squares of all such maximal minors. The matrix is rank-deficient if and only if $F(\Theta) = 0$.

Because σ is a real-analytic function, its derivative σ' is also real-analytic. Consequently, the entries of the matrices $A(\Theta)$ and $B(\Theta)$ in (121) and (122) are real-analytic functions of the parameters Θ. It follows that the function $F(\Theta)$ is real-analytic on $\mathbb{R}^N$.

If an analytic function on a connected open set is not identically zero, its zero set has Lebesgue measure zero [35, Proposition 0]. Therefore, to prove that $C_\ell(\Theta)$ is generically full rank, it is sufficient to construct a single parameter configuration $\Theta^* = (W^*, U^*, b^*)$ for which $F(\Theta^*) \neq 0$.

We construct this configuration as follows:

1. Choose $W^* = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ such that the eigenvalues λ_i are pairwise distinct.
2. Choose $U^* = [1, 1, \ldots, 1]^\top$.
3. Choose b^* such that $v_1 = v_2 = \cdots = v_m$. This ensures that the derivative is a strictly positive constant $c = \sigma'(v_i) > 0$ for all i.

Under the configuration Θ^*, the matrix $\Sigma'(v)$ simplifies to cI_m. Thus, the Jacobian A^* is diagonal:

$$A^* = \mathrm{diag}(a_1, \ldots, a_m), \quad \text{where } a_i = -\alpha + c\lambda_i. \tag{123}$$

Because the λ_i are pairwise distinct and $c > 0$, the eigenvalues a_i are also pairwise distinct. The vector B^* simplifies to $c[1, 1, \ldots, 1]^\top$.

The observability matrix $C_\ell(\Theta^*)$ takes the explicit form:

$$C_\ell(\Theta^*) = c \begin{bmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^\ell \\ 1 & a_2 & a_2^2 & \ldots & a_2^\ell \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_m & a_m^2 & \ldots & a_m^\ell \end{bmatrix} \tag{124}$$

Because $m \geq \ell + 1$, we can extract the square submatrix $M \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$ formed by the first $\ell + 1$ rows of $C_\ell(\Theta^*)$. Up to the non-zero scaling factor c, M is precisely a Vandermonde matrix generated by the elements $(a_1, \ldots, a_{\ell+1})$.

Since the values a_i are pairwise distinct, the determinant of the Vandermonde matrix is non-zero. Consequently, the maximal minor corresponding to M is non-zero, which implies $F(\Theta^*) > 0$.

Because $F(\Theta)$ is not identically zero, the set $\{\Theta \in \mathbb{R}^N \mid F(\Theta) = 0\}$ has Lebesgue measure zero. We conclude that $C_\ell(\Theta, o^*)$ has full column rank for almost all parameters Θ. $\square$

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: We provided all proofs and added experimental corroboration.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Given in Section 7

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All proofs are given in the supplementary material

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: [Yes]

   Justification: The paper describes all details required to implement models and training algorithms and is thus reproducible in that sense. However, for the part of the experiments which deal with a real robot application, the models have been trained on an internal large-scale dataset, which we unfortunately have not yet made public.

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: [No]

   Justification: This is mainly a theory paper. However, the part of the experiments which deal with a real robot application, the models have been trained on an internal large-scale dataset, which we unfortunately have not yet made public. Code will be released at some point but could unfortunately not do this before submission.

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: [Yes]

   Justification: The training hyper-parameters have been given in Section H.4

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: [Yes]

   Justification: This is mainly a theoretical paper, but for the experimental part we trained the models with multiple seeds and we report plots with means $\pm$ standard-deviation.

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: [Yes]

Justification: This is explained in Section I

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: there is nothing to justify.

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 is highly theoretical work and while we could invent societal impact, we are not sure that this is a worthwhile exercise.

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 is highly theoretical work and we are not sure how safeguards apply.

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 only uses internal datasets of the team.

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: We do not introduce new datasets.

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: We did not use crowd sourcing.

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: We did not use research with or on human subjects.

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: LLMs were only used for coding assistance and for improving writing.