NEURAL ODES MEET CONCURRENT LEARNING: STABLE ONLINE LEARNING WITH LYAPUNOV GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural ODEs learn dynamics from trajectory losses, but their adjoint gradients lack the regressor-times-parameter-error structure on which Lyapunov analyses of online adaptation rest, so training on streaming data comes without stability guarantees. We show that this structure is in fact present: the adjoint gradient decomposes exactly into a positive semi-definite trajectory operator acting on the parameter error plus a nonlinear perturbation with explicit, horizon-dependent bounds. A quadratic Lyapunov function then certifies online Neural ODE training over sliding windows under computable gain and horizon conditions, and the same certificate extends to stored data: its drift branch recovers concurrent learning, and its trajectory branch yields NODE-CL, a stored-segment Gauss–Newton method built on batched forward sensitivities that needs no state-derivative estimates. On four DeepMind Control Suite domains, NODE-CL attains the lowest median prediction error on three under velocity measurement noise, where observer-based concurrent learning degrades by up to $8\times$; with clean measurements it is best on the pendulum and within a factor of 1.6 of the best stored-data baseline on the cartpole and reacher.

1 INTRODUCTION

Two communities learn unknown dynamics from data with neural networks and rarely talk to each other. In machine learning, a neural ordinary differential equation (Neural ODE) is fitted by minimizing a loss over predicted trajectories with adjoint gradients (Chen et al., 2018); the loss is informative because a trajectory mismatch integrates the modeling error through the dynamics, but training is offline, with no guarantee about the parameter flow on streaming data. In adaptive control, the same network is updated online from the instantaneous prediction error, and a Lyapunov argument certifies that the estimation error stays bounded (Sanner & Slotine, 1992; Lewis et al., 1999; Patil et al., 2022), because that residual is, to first order, a known regressor times the parameter error. Trajectory losses lack that structure: the predicted trajectory depends on the parameters through an entire solve, so the adjoint gradient is not a regressor times a parameter error, and least-squares gains inject an indefinite term into any Lyapunov derivative. Concurrent learning (Chowdhary & Johnson, 2011), which replays stored samples, sits in between: certified, but tied to the instantaneous residual and hence to a state-derivative observer.

This paper proposes NODE-CL, a stored-segment Gauss–Newton method for online Neural ODE identification that needs no state-derivative estimates and comes with a Lyapunov certificate, and shows that the same certificate covers both communities' methods (Figure 1). The result that makes this possible is a trajectory-error decomposition, under which the adjoint gradient decomposes into a positive semi-definite trajectory operator acting on the parameter estimation error plus a perturbation with explicit, window-length dependent constants. A quadratic Lyapunov function then suffices: online Neural ODE training over sliding windows is certified under computable gain and horizon conditions, the window can be replaced by stored segments, and a Gauss–Newton gain from batched forward sensitivities can be added without leaving the certificate. The drift-based formulation is the foundation: with a least-squares gain driven by the regressor of the gradient decomposition, the indefinite gain-dynamics term is dominated by the gradient term, and the stored-set version of that law contains the stack structure of concurrent learning as its constant-gain instance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

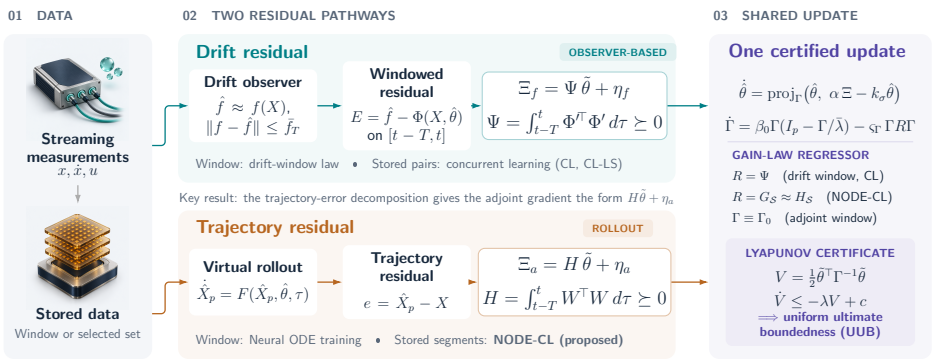

Figure 1: Streaming data (01) feed two residual pathways (02), an instantaneous drift residual formed with an observer and a trajectory residual formed by rolling the model out from stored initial states; each gradient decomposes into a positive semi-definite operator on the parameter error plus a bounded perturbation, so both pathways share one certified update (03).

Contributions. First, NODE-CL: on four DeepMind Control Suite domains it is the most robust of the compared methods to velocity measurement noise and, with clean data, is best on the pendulum and within a factor of 1.6 of the best stored-data baseline on the cartpole and reacher, against concurrent-learning baselines derived from Hart et al. (2025a) (CL-LS and CL) and controlled Adam replay, with horizon, observer, and gain ablations. Second, the trajectory-error decomposition, which gives the adjoint gradient of a nonlinear-in-parameter trajectory model the regressor form needed for Lyapunov analysis, with explicit horizon-dependent constants. Third, two further certified online methods that the same analysis yields, drift-window and adjoint-window identification. Computational costs and limitations are reported in Appendix K.

Related work. Neural ODEs (Chen et al., 2018; Dupont et al., 2019; Rubanova et al., 2019; Kidger et al., 2020; Massaroli et al., 2020; Kidger, 2022) are commonly trained offline with adjoint or checkpointed gradients (Zhuang et al., 2020), and long horizons are handled by multiple shooting (Turan & Jäschke, 2022; Massaroli et al., 2021) or by regularizing the flow (Finlay et al., 2020). Existing Lyapunov-based neural adaptive laws (Sanner & Slotine, 1992; Lewis et al., 1999; Patil et al., 2022; Joshi & Chowdhary, 2019; Sun et al., 2022) use instantaneous residuals; composite adaptation and concurrent learning add recorded data (Slotine & Li, 1989; Chowdhary & Johnson, 2011; Kamalapurkar et al., 2017). Least-squares DNN laws appear in Makumi et al. (2025); Hart et al. (2025a). Our drift branch adds certified gain bounds and a gain-metric projection, while the trajectory branch removes the derivative observer.

The NODE-CL gain recursion is related to covariance updates in extended-Kalman-filter and recursive Gauss–Newton training (Singhal & Wu, 1989; Puskorius & Feldkamp, 1994; Haykin, 2001); the new ingredient is the trajectory-error decomposition that places this update inside a Lyapunov certificate. This differs from certifying properties of the learned dynamics themselves, such as stability of the learned flow (Kolter & Manek, 2019; Kang et al., 2021; Rodriguez et al., 2022; Luo et al., 2025) or of a constraint or attractor set (White et al., 2023; Sochopoulos et al., 2024), and from regret guarantees for instantaneous nonlinear adaptation (Boffi & Slotine, 2021; Boffi et al., 2021) and meta-learned adaptive controllers (Richards et al., 2021; Shi et al., 2021; O'Connell et al., 2022); the certificate here concerns the online parameter flow. Classical identification and sparse regression use different model classes (Ljung, 1999; Brunton et al., 2016); the projection follows Lavretsky & Wise (2013); Cai et al. (2006). Section 2 formulates the problem, Section 3 develops the methods, and Section 4 reports the experiments.

2 Problem Formulation

Consider the second-order nonlinear system

$$\ddot{x} = f(x, \dot{x}) + g(x, \dot{x})u, \tag{1}$$

where $x, \dot{x} \in \mathbb{R}^n$ are measurable states, $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is an unknown drift to be identified, $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$ is a known input matrix, and $u \in \mathbb{R}^m$ is a known external input. Let $X \triangleq [x^\top \ \dot{x}^\top]^\top \in \mathbb{R}^{2n}$ and $B \triangleq [0_{n\times n} \ I_n]^\top \in \mathbb{R}^{2n\times n}$; we write $f(X)$ and $g(X)$ for $f(x, \dot{x})$ and $g(x, \dot{x})$. The following assumptions describe the identification experiment and the regularity of the system.

Assumption 1. The external input is bounded, $\|u(t)\| \le \bar{u}$ for all $t \ge t_0$, and the resulting state trajectory is bounded, $\|X(t)\| \le \bar{X}$ for all $t \ge t_0$ (standard in identification, where boundedness is enforced by a stabilizing baseline input, as in Section 4).

Assumption 2. The drift f is continuously differentiable, and g is twice continuously differentiable. Consequently, on $\{X : \|X\| \le \bar{X}\}$ there exist constants $c_{fx}, c_{fv}, \bar{f}, \bar{g} > 0$ such that $\|\partial f/\partial x\| \le c_{fx}$, $\|\partial f/\partial \dot{x}\| \le c_{fv}$, $\|f\| \le \bar{f}$, and $\|g\| \le \bar{g}$.

Let $\Phi : \mathbb{R}^{2n} \times \mathbb{R}^p \to \mathbb{R}^n$ denote a DNN with parameters $\theta \in \mathbb{R}^p$ and $\Phi'(X, \theta) \triangleq \partial\Phi/\partial\theta \in \mathbb{R}^{n\times p}$. Let $\Theta \triangleq \{\theta \in \mathbb{R}^p : \|\theta\| \le \bar{\theta}\}$ denote the admissible parameter set. Fix a margin $\varpi > 0$ and define the compact convex domain $\Omega \triangleq \{X \in \mathbb{R}^{2n} : \|X\| \le \bar{X} + \varpi\}$, which contains the trajectory and a tube in which the virtual predictions below remain. Since Φ is a composition of smooth layers, it is twice continuously differentiable on any compact set, so there exist constants $\bar{v}, c_{\theta\theta}, c_{X\theta} > 0$ such that, on $\Omega \times \Theta$, $\|\Phi'(X, \theta)\| \le \bar{v}$, $\left\|\partial^2\Phi/\partial\theta^2\right\| \le c_{\theta\theta}$, and $\left\|\partial^2\Phi/\partial X\partial\theta\right\| \le c_{X\theta}$.

Let μ denote a finite Borel measure on Ω, and define the ideal parameters as $\theta^* = \arg\min_{\theta\in\Theta} \int_\Omega \|f(X) - \Phi(X, \theta)\|^2 \, d\mu(X) + \rho \|\theta\|^2$ with a regularization weight $\rho > 0$.[1] The drift is then modeled as

$$f(X) = \Phi(X, \theta^*) + \varepsilon(X), \qquad \max_{X\in\Omega} \|\varepsilon(X)\| \le \bar{\varepsilon}, \tag{2}$$

where $\bar{\varepsilon}$ can be made small by an expressive architecture and a large enough $\bar{\theta}$; every ultimate bound below grows with $\bar{\varepsilon}$.

The objective is an online law for $\hat{\theta}(t)$ under which the estimation error $\tilde{\theta}(t) \triangleq \theta^* - \hat{\theta}(t)$ is uniformly ultimately bounded, with an exponential rate into a residual set determined by the leakage, approximation, and observer errors, which shrinks under a window persistence-of-excitation condition stated in the appendix. Section 3 develops the adaptation laws that achieve this objective.

3 Method: From Local Residuals to NODE-CL

We develop the method in three stages, each building on the previous one. First, we define a drift residual, the mismatch between an observer-based drift estimate and the model over a sliding window, and derive a gradient-based estimator with a certified least-squares gain. Second, we replace the drift residual by a trajectory residual, the mismatch between a model rollout and the measured trajectory, which removes the observer, and show that its adjoint gradient admits the same certified estimator. Third, we apply concurrent learning to the trajectory regressor: NODE-CL replays stored trajectory segments instead of a sliding window and rescales their directions with a Gauss–Newton gain. Section 3.1 presents the first stage.

3.1 Stage I: proposed drift-window method

We introduce the drift-window method to convert streaming state and input measurements into approximate drift labels and to adapt the model over a sliding window of those labels. Let $\hat{f}$ denote a drift estimate and $\tilde{f} \triangleq f(X) - \hat{f}$ its error. Dynamic state-derivative estimators are standard in concurrent learning (Kamalapurkar et al., 2017; Hart et al., 2025a). Throughout, the estimate is kept by a projection (introduced in Section 3.1.2) inside $\Theta_\epsilon \triangleq \{\theta : \|\theta\| \le \bar{\theta} + \epsilon_\theta\}$, the inflation of Θ by a margin $\epsilon_\theta > 0$, so $\left\|\tilde{\theta}(t)\right\| \le \bar{\tilde{\theta}} \triangleq 2\bar{\theta} + \epsilon_\theta$; the constants $\bar{v}, c_{\theta\theta}, c_{X\theta}$ are taken so that the bounds of Section 2 hold on $\Omega \times \Theta_\epsilon$. The analysis below requires only an estimator error envelope, rather than a particular construction. For completeness, Appendix A specifies the observer used in the

[1] The reader is referred to Hart et al. (2025b) for identifiability conditions under which this regularized minimizer is unique.

experiments and proves the bound below. For a window length $T > 0$ and $t \geq t_0 + T$, the resulting bound on the observer error over the window is

$$\sup_{\tau \in [t-T,t]} \left\| \tilde{f}(\tau) \right\| \leq \bar{f}_T(t) \triangleq \|z_o(t_0)\| \, e^{-k_o(t-T-t_0)} + \bar{z}_o. \tag{3}$$

where z_o is the observer error state, $k_o > 0$ its decay rate, and $\bar{z}_o$ its steady-state residual, which decreases with the observer gain (all given explicitly in Appendix A). Section 3.1.1 defines the windowed loss and decomposes its gradient, Section 3.1.2 introduces the gain and projection, and Section 3.1.3 states the adaptation law and its guarantee.

3.1.1 Window loss and gradient decomposition

To make the update less sensitive to the error of any single sample, we accumulate the residual over a sliding window rather than at a single instant, and we show that the gradient of this windowed loss has the regressor form needed for the Lyapunov analysis. Fix a window length $T > 0$. For $t \geq t_0 + T$, the windowed drift prediction error and loss are defined as

$$E(\tau; t) \triangleq \hat{f}(\tau) - \Phi(X(\tau), \hat{\theta}(t)), \tag{4}$$

$$\mathcal{E}_f(t) \triangleq \tfrac{1}{2} \int_{t-T}^{t} \|E(\tau; t)\|^2 \, d\tau, \tag{5}$$

where the pairs $(X(\tau), \hat{f}(\tau))$ are buffered and $\hat{\theta}(t)$ is the current estimate. Since only Φ depends on $\hat{\theta}$, the negative gradient of Equation (5) is

$$\Xi_f(t) \triangleq -\nabla_{\hat{\theta}} \mathcal{E}_f(t) = \int_{t-T}^{t} \Phi'(X(\tau), \hat{\theta}(t))^\top E(\tau; t) \, d\tau. \tag{6}$$

Expanding $\Phi(X(\tau), \theta^*)$ about $\hat{\theta}(t)$ in each sample gives the following decomposition of Ξ_f, which is the form every Lyapunov argument below relies on.

Proposition 1. *Let Assumptions 1 and 2 hold, and let $\hat{\theta}(t) \in \Theta_\epsilon$. Then, for $t \geq t_0 + T$,*

$$\Xi_f(t) = \Psi(t) \, \tilde{\theta}(t) + \eta_f(t), \tag{7}$$

where the windowed regressor

$$\Psi(t) \triangleq \int_{t-T}^{t} \Phi'(X(\tau), \hat{\theta}(t))^\top \Phi'(X(\tau), \hat{\theta}(t)) \, d\tau \tag{8}$$

is positive semi-definite with $\|\Psi(t)\| \leq T\bar{v}^2$, and the perturbation satisfies

$$\|\eta_f(t)\| \leq c_{\eta 1} \left\| \tilde{\theta}(t) \right\|^2 + c_{\eta 2} \, \bar{f}_T(t) + c_{\eta 3}, \tag{9}$$

with $c_{\eta 1} \triangleq \frac{1}{2} T \bar{v} \, c_{\theta\theta}$, $c_{\eta 2} \triangleq T\bar{v}$, and $c_{\eta 3} \triangleq T\bar{v} \, \bar{\varepsilon}$.

The proof is given in Appendix E; the three terms in Equation (9) are the Taylor remainder, the observer error, and the approximation error, each scaled by T. Section 3.1.2 introduces the gain and projection used in the update law.

3.1.2 Least-squares gain and projection

The adaptation law also requires a least-squares gain for excited directions and a projection that keeps the estimate inside Θ_ϵ, where the smoothness constants of Φ hold. The gain evolves as

$$\dot{\Gamma} = \beta_0 \, \Gamma \left(I_p - \Gamma / \bar{\lambda} \right) - \varsigma_\Gamma(t) \, \Gamma \, \Psi(t) \, \Gamma, \qquad \Gamma(t_0) = \Gamma_0, \tag{10}$$

where $\beta_0 > 0$ is the forgetting rate and $\bar{\lambda} > 0$ is a cap. The Lipschitz gate $\varsigma_\Gamma(t) = \varsigma(\lambda_{\min}(\Gamma(t)))$ is zero below the floor $\underline{\lambda}$, one above $2\underline{\lambda}$, and linear between them, with $0 < \underline{\lambda} < \bar{\lambda}/2$; set $\Psi = 0$ during the initial window. Driving the gain by the same regressor as Equation (7) preserves $\underline{\lambda} I_p \preceq \Gamma(t) \preceq \bar{\lambda} I_p$ (proved in Appendix C). The gain-metric projection $\mathrm{proj}_\Gamma(\theta, y)$ equals Γy inside Θ and, in the margin between Θ and Θ_ϵ, removes the component of Γy that points outward (defined in Appendix D); it keeps $\hat{\theta}(t) \in \Theta_\epsilon$ and satisfies, for every $\theta^* \in \Theta$,

$$\tilde{\theta}^\top \Gamma^{-1} \mathrm{proj}_\Gamma(\hat{\theta}, y) \geq \tilde{\theta}^\top y \tag{11}$$

(proved in Appendix D). This metric compatibility is the property used in every Lyapunov proof. Section 3.1.3 states the adaptation law and its guarantee.

3.1.3 ADAPTATION LAW AND GUARANTEE

The drift-based windowed adaptation law is designed as

$$\dot{\hat{\theta}} = \mathrm{proj}_{\Gamma}\Big(\hat{\theta},\ \alpha\,\Xi_f(t) - k_\sigma\,\hat{\theta}\Big), \tag{12}$$

where $\alpha > 0$ is the learning gain and $k_\sigma > 0$ is a sigma-modification gain; set $\Xi_f = 0$ during the initial window. This is a projected, least-squares preconditioned gradient flow on Equation (5). Existence and uniqueness of solutions are addressed in Appendix D. The following proposition certifies this law.

Proposition 2. *Let Assumptions 1 and 2 hold, let $\hat{\theta}(t_0) \in \Theta_\epsilon$, let $\underline{\lambda} I_p \preceq \Gamma_0 \preceq \bar{\lambda} I_p$, and let the gains satisfy $\alpha > \frac{1}{2}$ and $k_\sigma > \alpha\, T\, \bar{v}\, c_{\theta\theta}\, \bar{\tilde{\theta}}$. Suppose the drift estimate satisfies Equation* (3)*. Then, under the gain dynamics in Equation* (10) *and the adaptation law in Equation* (12)*, the parameter estimation error satisfies, for all $t \geq t_1 \triangleq t_0 + T$,*

$$\left\|\tilde{\theta}(t)\right\| \leq \sqrt{\tfrac{\bar{\lambda}}{\underline{\lambda}}\left\|\tilde{\theta}(t_1)\right\|^2 e^{-\lambda_d (t-t_1)} + \tfrac{2\bar{\lambda}}{\lambda_d}\, c_d(t_1)}, \qquad \limsup_{t\to\infty}\left\|\tilde{\theta}(t)\right\| \leq \sqrt{\tfrac{2\bar{\lambda}}{\lambda_d}\, c_d^{\infty}}, \tag{13}$$

where $k_1 \triangleq \frac{k_\sigma}{2} - \alpha c_{\eta 1}\bar{\tilde{\theta}} > 0$, $\lambda_d \triangleq 2k_1\underline{\lambda}$,

$$c_d(t) \triangleq k_\sigma \bar{\theta}^2 + \tfrac{\alpha^2}{k_\sigma}\left(c_{\eta 2}\,\bar{f}_T(t) + c_{\eta 3}\right)^2, \qquad c_d^{\infty} \triangleq k_\sigma \bar{\theta}^2 + \tfrac{\alpha^2}{k_\sigma}\left(c_{\eta 2}\,\bar{z}_o + c_{\eta 3}\right)^2. \tag{14}$$

The proof is given in Appendix F: in $V = \frac{1}{2}\tilde{\theta}^\top \Gamma^{-1}\tilde{\theta}$ the gain dynamics produce $\frac{1}{2}\varsigma_\Gamma \tilde{\theta}^\top \Psi \tilde{\theta}$ and the projected gradient $-\alpha\tilde{\theta}^\top \Psi\tilde{\theta}$, so $\alpha > \frac{1}{2}$ leaves a sink; $k_\sigma\bar{\theta}^2$ is the usual sigma-modification tradeoff. Without excitation no law can localize θ beyond the projection set, and Equation (13) reflects this (see the remark in Appendix F).

Two further consequences are stated and proved in Appendix F: the sink retained in the Lyapunov argument bounds the time-averaged first-order functional error without excitation, and it improves the parameter rate under a window persistence-of-excitation condition. The analysis also applies to noncontiguous stored drift pairs with uniformly bounded observer errors (Appendix J); this stored-set result supplies the guarantee used by the adapted CL-LS baseline. Section 3.2 replaces the drift residual by a trajectory residual.

3.2 STAGE II: PROPOSED ADJOINT-WINDOW METHOD

We now replace the drift residual by a trajectory residual, the mismatch between a model rollout and the measured trajectory, which needs no derivative estimate. The difficulty is that its adjoint gradient has no immediate regressor form, and this section derives one: Section 3.2.1 defines the rollout and decomposes its adjoint gradient, and Section 3.2.2 states the update and its guarantee.

3.2.1 VIRTUAL SYSTEM, WINDOW LOSS, AND ADJOINT GRADIENT

Define the model vector field $F : \mathbb{R}^{2n} \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^{2n}$ as $F(\chi, \theta, \tau) \triangleq [\chi_v^\top,\ (\Phi(\chi,\theta) + g(\chi)\,u_{\mathrm{rec}}(\tau))^\top]^\top$, where $\chi = [\chi_x^\top\ \chi_v^\top]^\top$ and u_{rec} is the recorded input. At time $t \geq t_0 + T$, the virtual predicted trajectory $\hat{X}_p(\cdot; t) : [t-T, t] \to \mathbb{R}^{2n}$ is generated by

$$\frac{d\hat{X}_p}{d\tau} = F(\hat{X}_p, \hat{\theta}(t), \tau), \qquad \hat{X}_p(t-T; t) = X(t-T), \tag{15}$$

and the trajectory window loss and prediction error are defined as

$$\mathcal{E}_T(t) \triangleq \tfrac{1}{2}\int_{t-T}^{t} \|e(\tau; t)\|^2\, d\tau, \qquad e(\tau; t) \triangleq \hat{X}_p(\tau; t) - X(\tau). \tag{16}$$

Let $L_F, L_{FX} > 0$ bound $\|\partial F/\partial\chi\|$ and $\left\|\partial^2 F/\partial\chi^2\right\|$ on $\Omega \times \Theta_\epsilon$, and let $\kappa_T \triangleq (e^{L_F T} - 1)/L_F$. The virtual trajectory remains in Ω under the window condition

$$\kappa_T\left(\bar{v}\,\bar{\tilde{\theta}} + \bar{\varepsilon}\right) < \varpi, \tag{17}$$

(proved in Appendix H), which also gives $\|e(\tau;t)\| < \varpi$.

By the adjoint sensitivity method (Chen et al., 2018), the gradient of Equation (16) with respect to $\hat{\theta}$ is

$$\begin{aligned} \nabla_{\hat{\theta}} \mathcal{E}_T(t) &= \int_{t-T}^{t} \left(\frac{\partial F}{\partial \theta}(\hat{X}_p, \hat{\theta}(t), \tau)\right)^\top \nu(\tau)\, d\tau, \\ \frac{d\nu}{d\tau} &= -\left(\frac{\partial F}{\partial \chi}(\hat{X}_p, \hat{\theta}(t), \tau)\right)^\top \nu - e(\tau;t), \qquad \nu(t) = 0_{2n}, \end{aligned} \tag{18}$$

where $\nu : [t-T, t] \to \mathbb{R}^{2n}$ is the adjoint state (derived in Appendix G).

Since $\partial F/\partial\theta = B\,\Phi'(\hat{X}_p, \theta)$, partitioning $\nu = [\nu_x^\top\ \nu_v^\top]^\top$ gives the trajectory gradient signal

$$\Xi_a(t) \triangleq -\nabla_{\hat{\theta}} \mathcal{E}_T(t) = -\int_{t-T}^{t} \Phi'(\hat{X}_p(\tau;t), \hat{\theta}(t))^\top \nu_v(\tau)\, d\tau. \tag{19}$$

The main result of the paper, stated next, shows that the trajectory gradient has the same structure as the drift gradient: a positive semi-definite operator acting on the parameter error plus a bounded perturbation. It follows from an exact representation of the rollout error, obtained by variation of constants applied to a mean-value linearization, which yields a trajectory regressor $W(\tau)$ and a remainder $d(\tau)$ with $e = -(W\tilde{\theta} + d)$; their definitions and all constants are given in Appendix H.

Theorem 1. *Let Assumptions 1 and 2 and Equation* (17) *hold, and let $\hat{\theta}(t) \in \Theta_\epsilon$. Let W and d be the explicitly constructed quantities in Appendix H. Then $e(\tau;t) = -\Big(W(\tau)\,\tilde{\theta}(t) + d(\tau)\Big)$ for all $\tau \in [t-T, t]$, and the adjoint gradient signal in Equation* (19) *decomposes as*

$$\Xi_a(t) = H(t)\,\tilde{\theta}(t) + \eta_a(t), \tag{20}$$

where $H(t) \triangleq \int_{t-T}^{t} W(\tau)^\top W(\tau)\, d\tau \succeq 0$ with $\|H(t)\| \le T\bar{w}^2$, and

$$\|\eta_a(t)\| \le b_1 \left\|\tilde{\theta}\right\|^2 + b_2 \left\|\tilde{\theta}\right\| + b_3, \tag{21}$$

with $\bar{w} \triangleq \bar{v}\kappa_T$ and constants $b_1, b_2, b_3 \ge 0$ defined in Appendix H, where $b_1 = O(T\kappa_T^2)$, b_2 is proportional to $\bar{\varepsilon}$, and b_3 vanishes with $\bar{\varepsilon}$.

The proof is given in Appendix H. Every constant in Equation (21) carries a factor $T\kappa_T$, so $b_1, b_2, b_3 \to 0$ as $T \to 0$. No observer error appears because Equation (16) compares the rollout with measured states. The corresponding cost is the Gronwall growth of the constants with T. Section 3.2.2 uses this decomposition to certify the online update.

3.2.2 Certified online update

The trajectory-based windowed adaptation law is designed as

$$\dot{\hat{\theta}} = \mathrm{proj}_{\Gamma_0}\Big(\hat{\theta},\ \alpha\,\Xi_a(t) - k_\sigma\,\hat{\theta}\Big), \tag{22}$$

with a constant symmetric positive definite gain Γ_0 satisfying $\underline{\lambda} I_p \preceq \Gamma_0 \preceq \bar{\lambda} I_p$, and $\Xi_a \triangleq 0_p$ on $[t_0, t_0 + T]$. Forming H online would require extra solves, so the trajectory law uses a constant gain. This gain contributes no additional Lyapunov term, and therefore $\alpha > \frac{1}{2}$ is not required. The resulting guarantee is stated next.

Proposition 3. *Let Assumptions 1 and 2 hold, let $\hat{\theta}(t_0) \in \Theta_\epsilon$, and let the window length and gains satisfy Equation* (17) *and $k_\sigma > 2\alpha\left(b_1\bar{\tilde{\theta}} + b_2\right)$. Then, for the adaptation law in Equation* (22) *with the adjoint gradient in Equation* (19)*, the parameter estimation error satisfies, for all $t \ge t_1 = t_0 + T$,*

$$\left\|\tilde{\theta}(t)\right\| \le \sqrt{\tfrac{\bar{\lambda}}{\underline{\lambda}} \left\|\tilde{\theta}(t_1)\right\|^2 e^{-\lambda_a (t-t_1)} + \tfrac{\bar{\lambda}\, c_a}{k_{1a}\,\underline{\lambda}}}, \tag{23}$$

where $k_{1a} \triangleq \frac{k_\sigma}{2} - \alpha\left(b_1\bar{\tilde{\theta}} + b_2\right) > 0$, $\lambda_a \triangleq 2k_{1a}\underline{\lambda}$, and $c_a \triangleq k_\sigma\bar{\theta}^2 + \alpha^2 b_3^2/k_\sigma$.

The proof is given in Appendix I. Since b_3 vanishes with $\bar{\varepsilon}$, the residual $\alpha^2 b_3^2/k_\sigma$ vanishes with the approximation error. Moreover, $b_1, b_2 \to 0$ as $T \to 0$, so the gain condition can always be met by shortening the window. For $L_F T \le 1$, it holds whenever $T \le (k_\sigma/(2\alpha C_b))^{1/3}$ for a computable constant C_b (Appendix I). Section 3.3 extends the trajectory law to stored segments.

3.3 Stage III: proposed stored-segment NODE-CL

We now replace the sliding window of Stage II by stored trajectory segments, in the spirit of concurrent learning. A sliding window retains only the most recently visited region of the state space, and a constant gain does not compensate for the scale of the trajectory operator H, which grows as T^3 for short windows since $\|H\| \le T\bar{w}^2$ and $\bar{w} = \bar{v}\kappa_T \approx \bar{v}T$. NODE-CL addresses these limitations by retaining selected trajectory segments and adapting the gain. Let $\mathcal{S}(t) = \{(X_j(\cdot), u_j(\cdot))\}_{j=1}^{N_s}$ be N_s stored segments, each a measured state and input record over $[t_j, t_j + T_s]$ with $t_j + T_s \le t$ and membership piecewise constant in t; let $\hat{X}_{p,j}$ solve Equation (15) from $X_j(t_j)$ with the current $\hat{\theta}(t)$, $e_j \triangleq \hat{X}_{p,j} - X_j$, and $S_j \triangleq \partial \hat{X}_{p,j}/\partial\hat{\theta}$. Define

$$\Xi_{\mathcal{S}} \triangleq -\frac{1}{N_s}\sum_{j=1}^{N_s}\int_{t_j}^{t_j+T_s} S_j^\top e_j \, d\tau, \qquad G_{\mathcal{S}} \triangleq \frac{1}{N_s}\sum_{j=1}^{N_s}\int_{t_j}^{t_j+T_s} S_j^\top S_j \, d\tau, \tag{24}$$

the negative gradient of the averaged segment loss and its Gauss–Newton matrix, both produced by one batched forward pass at cost $O(N_s N_w n^2 p)$ per update (N_w integration steps per segment). Applying the trajectory-error decomposition of Theorem 1 segment by segment gives $\Xi_{\mathcal{S}} = H_{\mathcal{S}}\tilde{\theta} + \eta_{\mathcal{S}}$ with $H_{\mathcal{S}} \triangleq \frac{1}{N_s}\sum_{j=1}^{N_s}\int_{t_j}^{t_j+T_s} W_j^\top W_j \, d\tau \succeq 0$ and $\|\eta_{\mathcal{S}}\| \le b_1 \left\|\tilde{\theta}\right\|^2 + b_2 \left\|\tilde{\theta}\right\| + b_3$ at segment length T_s. The computable $G_{\mathcal{S}}$ differs from $H_{\mathcal{S}}$ only through $S_j - W_j$: as shown in Appendix J, $\|G_{\mathcal{S}} - H_{\mathcal{S}}\| \le g_2 \left\|\tilde{\theta}\right\|^2 + g_1 \left\|\tilde{\theta}\right\| + g_0$ for constants g_0, g_1, g_2 built from those of Theorem 1 at T_s.

The stored-segment law is $\dot{\hat{\theta}} = \mathrm{proj}_\Gamma(\hat{\theta}, \alpha\Xi_{\mathcal{S}} - k_\sigma\hat{\theta})$ with Γ generated by Equation (10), using $G_{\mathcal{S}}$ in place of Ψ. On the excited subspace, the forgetting update equilibrates at $\Gamma \approx \beta_0 G_{\mathcal{S}}^{-1}$. The resulting projected Gauss–Newton flow is insensitive to the T_s^3 scale and can therefore use short segments.

Algorithm 3 Proposed NODE-CL: Stored-Segment Gauss–Newton Identification

Require: Segment length T_s, budget N_s, gains $\alpha, k_\sigma, \beta_0, \bar{\lambda}, \underline{\lambda}$
1: Initialize $\hat{\theta}(t_0) \in \Theta_\epsilon$, $\Gamma(t_0) = \Gamma_0$, and segment set $\mathcal{S} \leftarrow \emptyset$
2: **for** each update time t **do**
3: Form a candidate from the latest T_s of (X, u) and admit it if it improves the minimum singular value of the stacked terminal sensitivities
4: Roll out all segments in $\mathcal{S}$ with current $\hat{\theta}$; integrate sensitivities S_j and accumulate $\Xi_{\mathcal{S}}, G_{\mathcal{S}}$ by Equation (24)
5: Update Γ by Equation (10), replacing Ψ with $G_{\mathcal{S}}$
6: Update $\hat{\theta}$ by $\dot{\hat{\theta}} = \mathrm{proj}_\Gamma(\hat{\theta}, \alpha\Xi_{\mathcal{S}} - k_\sigma\hat{\theta})$
7: **end for**

NODE-CL uses short segments to control sensitivity growth and admits segments that add complementary directions within a fixed memory budget. Before the projected update, $G_{\mathcal{S}}$ rescales those directions. Segment length therefore determines a bias–conditioning tradeoff, while the admission rule remains a data-selection heuristic rather than an excitation guarantee. The Gauss–Newton rescaling distinguishes NODE-CL from replaying the same segments with a generic optimizer. The following theorem certifies NODE-CL.

Theorem 2. *Let Assumptions 1 and 2 hold, let Equation* (17) *hold at T_s, and activate NODE-CL at time t_s with a nonempty segment set, $\hat{\theta}(t_s) \in \Theta_\epsilon$, and $\underline{\lambda} I_p \preceq \Gamma(t_s) \preceq \bar{\lambda} I_p$. Suppose $\alpha > \frac{1}{2}$ and*

$$k_\sigma > 2\alpha(b_1\bar{\tilde{\theta}} + b_2) + g_2\bar{\tilde{\theta}}^2 + g_1\bar{\tilde{\theta}} + g_0. \tag{25}$$

Then $\underline{\lambda} I_p \preceq \Gamma(t) \preceq \bar{\lambda} I_p$ *for all* $t \geq t_s$*. For any* $\varphi_s \geq 0$ *such that* $H_{\mathcal{S}}(t) \succeq \varphi_s I_p$ *for all* $t \geq t_s$ *(always valid with* $\varphi_s = 0$*), let*

$$k_{1s} \triangleq \tfrac{k_\sigma}{2} - \alpha(b_1\bar{\tilde{\theta}} + b_2) - \tfrac{1}{2}(g_2\bar{\tilde{\theta}}^2 + g_1\bar{\tilde{\theta}} + g_0), \quad k_s(\varphi_s) \triangleq k_{1s} + (\alpha - \tfrac{1}{2})\varphi_s, \quad \lambda_s(\varphi_s) \triangleq 2k_s(\varphi_s)\underline{\lambda}. \tag{26}$$

The parameter error satisfies

$$\left\|\tilde{\theta}(t)\right\| \leq \sqrt{\frac{\bar{\lambda}}{\underline{\lambda}} \left\|\tilde{\theta}(t_s)\right\|^2 e^{-\lambda_s(\varphi_s)(t-t_s)} + \frac{\bar{\lambda} c_a}{\underline{\lambda} k_s(\varphi_s)}}, \qquad \limsup_{t\to\infty} \left\|\tilde{\theta}(t)\right\| \leq \sqrt{\frac{\bar{\lambda} c_a}{\underline{\lambda} k_s(\varphi_s)}}. \tag{27}$$

The proof is given in Appendix J. Thus excitation increases $k_s(\varphi_s)$, improving both the exponential rate and the certified ultimate radius; without excitation, set $\varphi_s = 0$. The ratio $\bar{\lambda}/\underline{\lambda}$ in Equation (27) arises only from converting the Lyapunov metric to the Euclidean norm. By the same bound on $G_{\mathcal{S}} - H_{\mathcal{S}}$, the computable condition $G_{\mathcal{S}} \succeq \varphi I_p$ implies $H_{\mathcal{S}} \succeq (\varphi - g_2\bar{\tilde{\theta}}^2 - g_1\bar{\tilde{\theta}} - g_0) I_p$, so the enhanced rate holds with that φ_s whenever it is positive. Segments are admitted by the singular-value maximizing rule of Chowdhary & Johnson (2011) on the stacked terminal sensitivities $S_j(t_j + T_s)$, a heuristic that does not itself guarantee excitation. The complete procedure is given in Algorithm 3; pseudocode for the drift-window and adjoint-window methods is in Appendix B. Section 4 evaluates the three methods.

4 Simulations

The experiments test whether the observer-free trajectory pathway is robust to velocity noise while staying competitive on clean data; two diagnostic systems then check individual theorem conditions. Section 4.1 reports the benchmark and Section 4.2 the diagnostics.

4.1 Benchmark suite

Methods. NODE-CL is the proposed stored-segment method evaluated here. Baselines are single-step adaptation, CL-LS (the identification form of the least-squares concurrent-learning law for DNNs of Hart et al. (2025a), the stored-set extension of Proposition 2 in Appendix J), CL (the same DNN point stack with a constant gain, i.e., the classical stack of Chowdhary & Johnson (2011) applied to the DNN Jacobian, a simplification of Hart et al. (2025a) rather than a separate prior method), and NODE-replay (Adam (Kingma & Ba, 2015) applied to exactly the NODE-CL segments). The drift- and adjoint-window methods are evaluated in the diagnostics.

Protocol. Four DeepMind Control Suite domains (Tassa et al., 2018), pendulum, cartpole, acrobot, and reacher, are used unmodified, and the full acceleration map $\ddot{x} = \Phi(z, \theta)$ with $z = [\phi(x)^\top, \dot{x}^\top, u^\top]^\top$ is identified, where ϕ maps unlimited hinge angles to $(\cos, \sin)$; no input matrix is assumed known, and Section 3.2 applies with $F = [\chi_v^\top, \Phi(z, \theta)^\top]^\top$. Each run streams 60 s of 1 kHz measurements under a multisine excitation with a weak stabilizing baseline; five evaluation seeds vary excitation, initialization, and noise; a sixth, tuning-only seed is excluded. Evaluation is open-loop prediction from eight held-out initial conditions on a held-out record: root-mean-square state error normalized by the state standard deviation at 1 s and 4 s horizons, angles wrapped. All methods use the same network (one hidden layer, 73 to 162 parameters) and, per seed and noise level, the same measurement stream; CL and CL-LS share a point stack, NODE-replay and NODE-CL a segment budget and admission rule, the observer-based methods the observer, and the projected laws the projection. Gains were selected on the tuning seed and held fixed;[2] a six-link swimmer ($p = 356$) that no method learns at this budget is reported in Appendix K.

Results (Table 1, Figure 2). With clean measurements, NODE-CL is best on every pendulum seed (0.035 against 0.18 to 0.32) but not on the other domains. On the cartpole and reacher, the three stored-set methods lie within a factor of 1.6. Under velocity noise of 3×10^{-3}, CL-LS degrades on these three domains by a factor of 3.3 to 8.4 relative to its clean error, and CL degrades by up to 2.2;

[2] As is standard for Lyapunov-based nonlinear and adaptive designs, the gain conditions in Propositions 2 and 3 and Theorem 2 are sufficient, not necessary, and conservative (worst-case Lipschitz and Taylor constants); the gains were tuned for performance and, as Figure 11(c) and Appendix K show, the sampled implementations remain stable well below the certified thresholds.

Table 1: Benchmark suite (five evaluation seeds, tuning seed excluded, 60 s): held-out 1 s prediction error normalized by the state standard deviation, median over seeds, clean (c) and with velocity noise $\sigma_v = 3 \times 10^{-3}$ (n); in parentheses, seeds on which it is best; best median in bold. Initial-model errors: 2.16, 2.08, 1.52, 1.71. CL-LS is the identification form of the DNN concurrent-learning law of Hart et al. (2025a) (stored-set extension of Proposition 2, Appendix J); CL is its constant-gain simplification; NODE-CL is the proposed method in Theorem 2.

	pendulum ($p = 73$)		cartpole ($p = 146$)		acrobot ($p = 162$)		reacher ($p = 162$)	
Method	c	n	c	n	c	n	c	n
single-step	1.643 (0)	1.650 (0)	3.287 (0)	3.145 (0)	1.621 (1)	**1.232** (2)	1.509 (0)	1.516 (0)
CL (constant gain)	0.315 (0)	0.609 (0)	0.623 (0)	1.363 (0)	3.615 (0)	4.385 (0)	1.007 (0)	0.976 (0)
CL-LS (Hart et al., 2025a)	0.195 (0)	1.642 (0)	**0.251** (4)	1.984 (0)	1.409 (0)	2.015 (0)	**0.064** (2)	0.215 (0)
NODE-replay (control)	0.183 (0)	0.141 (3)	0.311 (0)	0.741 (1)	**1.002** (2)	2.129 (1)	0.077 (2)	0.088 (2)
NODE-CL (proposed)	**0.035** (5)	**0.038** (2)	0.381 (1)	**0.302** (4)	1.545 (2)	1.279 (2)	0.069 (1)	**0.075** (3)

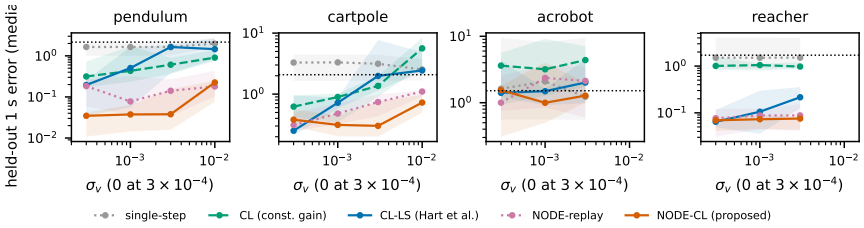

Figure 2: Held-out 1 s error against velocity noise per domain (five seeds, median and range; dotted, initial model). The 10^{-2} level was not run for the acrobot and reacher.

$k_f = 200$ was tuned on clean data rather than per noise level. NODE-CL has the lowest median on all three domains and wins nine of fifteen seeds, the expected advantage of trajectory residuals when derivative estimates are corrupted by noise. The advantage is not uniform: on the pendulum NODE-replay wins three seeds at 3×10^{-3} despite a higher median, and at 10^{-2} NODE-CL is ahead of NODE-replay on the cartpole (0.73 against 1.10) but behind on the pendulum (0.23 against 0.18). On the acrobot, NODE-CL improves on the initial model at 10^{-3} and 3×10^{-3} (1.00 and 1.28 against 1.52) but not with clean data (1.55), whereas NODE-replay does so only with clean data (1.00). Section 4.2 tests individual conditions of the analysis on two diagnostic systems.

4.2 Diagnostic studies and ablations

Two diagnostic systems, a teacher network (S1, $p = 25$) with known θ^* and a two-link manipulator (S2, $p = 86$), test distinct predictions of the analysis. **Excited subspace:** in S1, the component of the parameter error in the excited subspace contracts to 0.27 of its initial value while the ratio for the full parameter error remains near 0.80, consistent with the excitation-free functional bound of Appendix F. **Horizon:** the adjoint-window error increases about 3.7-fold from $T = 0.8$ to 1.6 s, consistent with the horizon dependence in Equation (21). **Sufficient gains:** sampled implementations remain stable below the conservative thresholds on α and leakage. Full curves, seed ranges, and condition checks are in Appendix K. Section 5 concludes.

5 Conclusion

We showed that online Neural ODE training and concurrent learning fit in one Lyapunov-certified framework for adaptive identification: drift and trajectory residuals both yield a gradient of the form regressor times parameter error plus a bounded perturbation, so one projected least-squares update is certified for windowed and stored data, and its stored-segment instance, NODE-CL, brings a Gauss–Newton gain to concurrent learning without state-derivative estimates. On the pendulum, cartpole, and reacher, NODE-CL is best at $\sigma_v = 3 \times 10^{-3}$, where the observer-based CL-LS degrades sharply; with clean data it is best on the pendulum and within a factor of 1.6 of the best method on the cartpole and reacher, and on the acrobot the advantage is not uniform.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 2, and complete proofs of every lemma, theorem, and corollary are given in Appendices A and C to J. NODE-CL pseudocode is given in Algorithm 3, the proposed drift-window and adjoint-window algorithms in Appendix B, and every gain, window length, sampling rate, architecture, excitation signal, and evaluation metric used in the simulations is listed in Section 4 and Appendix K. The supplementary material contains the complete simulation code, the resumable experiment drivers for the main comparison, the noise sweep, and the ablations, the analysis scripts that generate the tables and figures from the run records, the saved summary files from which the tables and Figure 2 are reproduced, and the numerical checks described in Appendix K; the raw run records are omitted from the archive for size and are regenerated by the drivers. The simulations are deterministic given the seed.

AI USE STATEMENT

In this work, generative AI tools were used for implementing the simulation code, for interpreting simulation results, and for drafting and editing the text and formatting references. The two illustrative icons in Figure 1 (a sensor module and a memory stack) are AI-generated conceptual images; no reported experimental data or numerical results were produced by generative AI, and the research questions were determined by the author(s). Selected benchmark derivatives and numerical updates were checked against the mathematical formulation, including finite-difference checks of the trajectory gradient and Gauss–Newton matrix and numerical checks of the projection inequality, the gain matrix bounds, the observer bound, and the drift gradient decomposition on simulated data; all AI-assisted text was reviewed. Responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI, rests with the author(s).

REFERENCES

Nicholas M. Boffi and Jean-Jacques E. Slotine. Implicit regularization and momentum algorithms in nonlinearly parameterized adaptive control and prediction. *Neural Computation*, 33(3):590–673, 2021.

Nicholas M. Boffi, Stephen Tu, and Jean-Jacques E. Slotine. Regret bounds for adaptive nonlinear control. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control (L4DC)*, volume 144 of *Proceedings of Machine Learning Research*, pp. 471–483. PMLR, 2021.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Zhijun Cai, Marcio S. de Queiroz, and Darren M. Dawson. A sufficiently smooth projection operator. *IEEE Transactions on Automatic Control*, 51(1):135–139, 2006.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Girish Chowdhary and Eric Johnson. Theory and flight-test validation of a concurrent-learning adaptive controller. *Journal of Guidance, Control, and Dynamics*, 34(2):592–607, 2011.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural ODEs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M. Oberman. How to train your neural ODE: The world of Jacobian and kinetic regularization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3154–3164. PMLR, 2020.

Rebecca G. Hart, Omkar Sudhir Patil, Zachary I. Bell, and Warren E. Dixon. Concurrent learning for system identification and control using Lyapunov-based deep neural networks. *IEEE Control Systems Letters*, 9:2957–2962, 2025a.

Rebecca G. Hart, Omkar Sudhir Patil, Zachary I. Bell, and Warren E. Dixon. System identification and control using Lyapunov-based deep neural networks without persistent excitation: A concurrent learning approach. *arXiv preprint arXiv:2505.10678*, 2025b.

Simon Haykin (ed.). *Kalman Filtering and Neural Networks*. Wiley, 2001.

Girish Joshi and Girish Chowdhary. Deep model reference adaptive control. In *IEEE Conference on Decision and Control*, pp. 4601–4608, 2019.

Rushikesh Kamalapurkar, Benjamin Reish, Girish Chowdhary, and Warren E. Dixon. Concurrent learning for parameter estimation using dynamic state-derivative estimators. *IEEE Transactions on Automatic Control*, 62(7):3594–3601, 2017.

Qiyu Kang, Yang Song, Qinxu Ding, and Wee Peng Tay. Stable neural ODE with Lyapunov-stable equilibrium points for defending against adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14925–14937, 2021.

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 3rd edition, 2002.

Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2022.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

J. Zico Kolter and Gaurav Manek. Learning stable deep dynamics models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Eugene Lavretsky and Kevin A. Wise. *Robust and Adaptive Control with Aerospace Applications*. Springer, 2013.

Frank L. Lewis, Sarangapani Jagannathan, and Aydin Yesildirek. *Neural Network Control of Robot Manipulators and Nonlinear Systems*. Taylor & Francis, 1999.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2nd edition, 1999.

Chaoyang Luo, Yan Zou, Wanying Li, and Nanjing Huang. FxTS-Net: Fixed-time stable learning framework for neural ODEs. *Neural Networks*, 185:107219, 2025.

Wanjiku Makumi, Omkar Sudhir Patil, and Warren E. Dixon. Lyapunov-based adaptive deep system identification for approximate dynamic programming. *Automatica*, 180:112462, 2025.

Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Dissecting neural ODEs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3952–3963, 2020.

Stefano Massaroli, Michael Poli, Sho Sonoda, Taiji Suzuki, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Differentiable multiple shooting layers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16532–16544, 2021.

Michael O'Connell, Guanya Shi, Xichen Shi, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural-Fly enables rapid learning for agile flight in strong winds. *Science Robotics*, 7(66):eabm6597, 2022.

Omkar Sudhir Patil, Duc M. Le, Max L. Greene, and Warren E. Dixon. Lyapunov-derived control and adaptive update laws for inner and outer layer weights of a deep neural network. *IEEE Control Systems Letters*, 6:1855–1860, 2022.

Gintaras V. Puskorius and Lee A. Feldkamp. Neurocontrol of nonlinear dynamical systems with Kalman filter trained recurrent networks. *IEEE Transactions on Neural Networks*, 5(2):279–297, 1994.

Spencer M. Richards, Navid Azizan, Jean-Jacques Slotine, and Marco Pavone. Adaptive-control-oriented meta-learning for nonlinear systems. In *Proceedings of Robotics: Science and Systems*, 2021.

Ivan Dario Jimenez Rodriguez, Aaron D. Ames, and Yisong Yue. LyaNet: A Lyapunov framework for training neural ODEs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18687–18703. PMLR, 2022.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Robert M. Sanner and Jean-Jacques E. Slotine. Gaussian networks for direct adaptive control. *IEEE Transactions on Neural Networks*, 3(6):837–863, 1992.

Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, and Yisong Yue. Meta-adaptive nonlinear control: Theory and algorithms. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10013–10025, 2021.

Sharad Singhal and Lance Wu. Training multilayer perceptrons with the extended Kalman algorithm. In *Advances in Neural Information Processing Systems*, volume 1, 1989.

Jean-Jacques E. Slotine and Weiping Li. Composite adaptive control of robot manipulators. *Automatica*, 25(4):509–519, 1989.

Andreas Sochopoulos, Michael Gienger, and Sethu Vijayakumar. Learning deep dynamical systems using stable neural ODEs. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 11163–11170, 2024.

Runhan Sun, Max L. Greene, Duc M. Le, Zachary I. Bell, Girish Chowdhary, and Warren E. Dixon. Lyapunov-based real-time and iterative adjustment of deep neural networks. *IEEE Control Systems Letters*, 6:193–198, 2022.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Evren Mert Turan and Johannes Jäschke. Multiple shooting for training neural differential equations on time series. *IEEE Control Systems Letters*, 6:1897–1902, 2022.

Alistair White, Niki Kilbertus, Maximilian Gelbrecht, and Niklas Boers. Stabilized neural differential equations for learning dynamics with explicit constraints. In *Advances in Neural Information Processing Systems*, volume 36, pp. 12929–12950, 2023.

Juntang Zhuang, Nicha Dvornek, Xiaoxiao Li, Sekhar Tatikonda, Xenophon Papademetris, and James Duncan. Adaptive checkpoint adjoint method for gradient estimation in neural ODE. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11639–11649. PMLR, 2020.

A Drift Observer and Error Bound

The experiments instantiate the drift estimate with the following state-derivative observer. Let $\hat{v} \in \mathbb{R}^n$ denote an auxiliary filter state and $\tilde{v} \triangleq \dot{x} - \hat{v}$ its measurable error. For gains $\alpha_o, k_f > 0$, define

$$\dot{\hat{v}} = \alpha_o \tilde{v} + \hat{f} + g(X)u, \tag{28}$$

$$\hat{f}(t) = \hat{f}(t_0) + k_f\left(\tilde{v}(t) - \tilde{v}(t_0)\right) + \int_{t_0}^{t} \left(k_f \alpha_o + 1\right) \tilde{v}(\tau)\, d\tau. \tag{29}$$

Differentiating Equation (29) and using Equations (1) and (28), the errors $\tilde{v}$ and $\tilde{f} = f(X) - \hat{f}$ satisfy $\dot{\tilde{v}} = \tilde{f} - \alpha_o \tilde{v}$ and $\dot{\tilde{f}} = \dot{f} - k_f \tilde{f} - \tilde{v}$. Let $z_o \triangleq [\tilde{v}^\top \ \tilde{f}^\top]^\top$.

Lemma 3. *Let Assumptions 1 and 2 hold. Then, for all $t \geq t_0$,*

$$\|z_o(t)\| \leq \|z_o(t_0)\|\, e^{-k_o(t-t_0)} + \bar{z}_o, \tag{30}$$

where $k_o \triangleq \min\{\alpha_o, k_f/2\}$, $\bar{z}_o \triangleq d_f/\sqrt{2k_f k_o}$, and $d_f \triangleq c_{fx}\bar{X} + c_{fv}(\bar{f} + \bar{g}\bar{u})$.

Proof. Consider the observer Lyapunov function $V_o = \frac{1}{2}\tilde{v}^\top\tilde{v} + \frac{1}{2}\tilde{f}^\top\tilde{f}$, which satisfies $V_o = \frac{1}{2}\|z_o\|^2$. Taking the time derivative along the observer error dynamics yields

$$\dot{V}_o = \tilde{v}^\top(\tilde{f} - \alpha_o\tilde{v}) + \tilde{f}^\top(\dot{f} - k_f\tilde{f} - \tilde{v}) = -\alpha_o\|\tilde{v}\|^2 - k_f\left\|\tilde{f}\right\|^2 + \tilde{f}^\top\dot{f}, \tag{31}$$

where the cross terms cancel. Under Assumptions 1 and 2, the drift derivative along the trajectory satisfies

$$\left\|\dot{f}\right\| = \left\|\tfrac{\partial f}{\partial x}\dot{x} + \tfrac{\partial f}{\partial \dot{x}}\left(f + gu\right)\right\| \leq c_{fx}\bar{X} + c_{fv}\left(\bar{f} + \bar{g}\bar{u}\right) = d_f. \tag{32}$$

Using Young's inequality, $\tilde{f}^\top\dot{f} \leq \frac{k_f}{2}\left\|\tilde{f}\right\|^2 + \frac{d_f^2}{2k_f}$, and therefore

$$\dot{V}_o \leq -k_o\|z_o\|^2 + c_o = -2k_oV_o + c_o, \tag{33}$$

where $k_o = \min\{\alpha_o, k_f/2\}$ and $c_o = d_f^2/(2k_f)$. The comparison lemma (Khalil, 2002) yields $V_o(t) \leq V_o(t_0)e^{-2k_o(t-t_0)} + \frac{c_o}{2k_o}$, and using $\|z_o\| = \sqrt{2V_o}$ together with $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ gives Equation (30) with $\bar{z}_o = \sqrt{c_o/k_o} = d_f/\sqrt{2k_fk_o}$. ■

B Pseudocode for the Proposed Drift-Window and Adjoint-Window Methods

The two windowed laws of Section 3 are given in pseudocode here, in the form implemented in the experiments.

Algorithm 1 Proposed Drift-Window Identification

Require: T, gains $\alpha_o, k_f, \alpha, k_\sigma, \beta_0, \bar{\lambda}, \underline{\lambda}$

Initialize $\hat{\theta}(t_0) \in \Theta_\epsilon$, $\Gamma(t_0) = \Gamma_0$, $\hat{v}(t_0) = \dot{x}(t_0)$, $\hat{f}(t_0) = 0$, empty circular buffer
for each time step t **do**
  Measure $x, \dot{x}$; record u
  Update observer Equation (28)–Equation (29); push $(X(t), \hat{f}(t))$ to buffer
  if $t \geq t_0 + T$ **then**
    $E(\tau) \leftarrow \hat{f}(\tau) - \Phi(X(\tau), \hat{\theta})$ over the buffer
    $\Xi_f \leftarrow$ quadrature of $\Phi'(X(\tau), \hat{\theta})^\top E(\tau)$
    $\Psi \leftarrow$ quadrature of $\Phi'(X(\tau), \hat{\theta})^\top \Phi'(X(\tau), \hat{\theta})$
  else
    $\Xi_f \leftarrow 0_p$; $\Psi \leftarrow 0_{p\times p}$
  end if
  Update Γ by Equation (10) with gate ς_Γ
  Update $\hat{\theta}$ by Equation (12)
end for

Algorithm 2 Proposed Adjoint-Window Identification

Require: T, gains α, k_σ, constant Γ_0
1: Initialize $\hat{\theta}(t_0) \in \Theta_\epsilon$, empty circular buffer
2: **for** each time step t **do**
3: Measure $x, \dot{x}$; record u; push $(X(t), u(t))$ to buffer
4: **if** $t \geq t_0 + T$ **then**
5: Forward: integrate Equation (15) from $X(t-T)$
6: Backward: integrate the adjoint equation in Equation (18) from $\nu(t) = 0$
7: $\Xi_a \leftarrow -$ quadrature of $\Phi'(\hat{X}_p, \hat{\theta})^\top \nu_v$
8: **else**
9: $\Xi_a \leftarrow 0_p$
10: **end if**
11: Update $\hat{\theta}$ by Equation (22)
12: **end for**

C Gain-Matrix Bounds

This appendix proves that the gated forgetting update keeps the gain matrix inside the certified interval.

Lemma 4. *If $\underline{\lambda} I_p \preceq \Gamma_0 \preceq \bar{\lambda} I_p$, then the solution of Equation* (10) *satisfies $\underline{\lambda} I_p \preceq \Gamma(t) \preceq \bar{\lambda} I_p$ for all $t \geq t_0$, and $\frac{d}{dt}\Gamma^{-1} = -\beta_0 \left(\Gamma^{-1} - \bar{\lambda}^{-1} I_p\right) + \varsigma_\Gamma \Psi$ with $\Gamma^{-1} - \bar{\lambda}^{-1} I_p \succeq 0$.*

The right-hand side of Equation (10) is locally Lipschitz in Γ (it is polynomial in Γ, and ς_Γ is Lipschitz in Γ through $\lambda_{\min}$), and Ψ is piecewise continuous, so solutions are unique and absolutely continuous, and Nagumo's theorem applies to the closed convex sets $\mathcal{K}_u \triangleq \{\Gamma : \Gamma \preceq \bar{\lambda} I_p\}$ and $\mathcal{K}_\ell \triangleq \{\Gamma : \Gamma \succeq \underline{\lambda} I_p\}$: each set is forward invariant if, at every boundary point, the vector field lies in the tangent cone, which for these sets holds if $v^\top \dot{\Gamma} v \leq 0$ (respectively ≥ 0) for every unit vector v in the extremal eigenspace.

For the upper bound, let $\Gamma \in \partial\mathcal{K}_u$, so $\lambda_{\max}(\Gamma) = \bar{\lambda}$, and let v be any unit vector with $\Gamma v = \bar{\lambda} v$. Then $v^\top \dot{\Gamma} v = \beta_0 \bar{\lambda}(1 - \bar{\lambda}/\bar{\lambda}) - \varsigma_\Gamma v^\top \Gamma \Psi \Gamma v \leq 0$. Since $\Gamma_0 \in \mathcal{K}_u$, forward invariance yields $\Gamma(t) \preceq \bar{\lambda} I_p$ for all t, hence $\Gamma^{-1} - \bar{\lambda}^{-1} I_p \succeq 0$. The expression for $\frac{d}{dt}\Gamma^{-1}$ follows from $\frac{d}{dt}\Gamma^{-1} = -\Gamma^{-1}\dot{\Gamma}\Gamma^{-1}$.

For the lower bound, let $\Gamma \in \partial\mathcal{K}_\ell$, so $\lambda_{\min}(\Gamma) = \underline{\lambda}$, and let v be any unit vector with $\Gamma v = \underline{\lambda} v$. Then $\varsigma_\Gamma = \varsigma(\underline{\lambda}) = 0$, so $v^\top \dot{\Gamma} v = \beta_0 \underline{\lambda}(1 - \underline{\lambda}/\bar{\lambda}) > 0$ since $\underline{\lambda} < \bar{\lambda}$. Since $\Gamma_0 \in \mathcal{K}_\ell$, forward invariance yields $\Gamma(t) \succeq \underline{\lambda} I_p$ for all t. ■

D Gain-Metric Projection

This appendix defines the gain-metric projection used in every adaptation law and proves its two properties. The projection operator is defined as follows. Let $p_c(\theta) \triangleq (\|\theta\|^2 - \bar{\theta}^2)/(2\epsilon_\theta \bar{\theta} + \epsilon_\theta^2)$, so that $p_c(\theta) \leq 0$ if and only if $\theta \in \Theta$ and $p_c(\theta) = 1$ if and only if $\|\theta\| = \bar{\theta} + \epsilon_\theta$. For a symmetric positive definite Γ, the projection operator is defined as

$$\mathrm{proj}_\Gamma(\theta, y) \triangleq \begin{cases} \Gamma y - p_c(\theta) \dfrac{\Gamma \nabla p_c \nabla p_c^\top \Gamma\, y}{\nabla p_c^\top \Gamma \nabla p_c}, & \text{if } p_c(\theta) > 0 \text{ and } \nabla p_c^\top \Gamma y > 0, \\ \Gamma y, & \text{otherwise,} \end{cases} \tag{34}$$

where $\nabla p_c = \nabla p_c(\theta)$; cf. the smooth projection operators in Lavretsky & Wise (2013); Cai et al. (2006). The operator removes only the component of Γy that points outward across the boundary of Θ_ϵ, and it does so in the Γ metric, which yields the following properties.

Lemma 5. *Let $\Gamma(t)$ be continuous, symmetric, and positive definite for each t, and let $\hat{\theta}$ evolve by $\dot{\hat{\theta}} = \mathrm{proj}_\Gamma(\hat{\theta}, y(t))$ for a piecewise continuous y. If $\hat{\theta}(t_0) \in \Theta_\epsilon$, then (i) $\hat{\theta}(t) \in \Theta_\epsilon$ for all $t \geq t_0$, and (ii) for every $\theta^* \in \Theta$, Equation* (11) *holds.*

The projection is continuous and locally Lipschitz across both switching surfaces. Together with the smoothness of Φ, this makes the right-hand sides of the proposed adaptation laws locally Lipschitz in $\hat{\theta}$ between data-set updates; hence their Carathéodory solutions exist and are unique.

For (i), on the boundary $p_c(\hat{\theta}) = 1$, the outward normal component of the flow satisfies $\nabla p_c^\top \dot{\hat{\theta}} = (1 - p_c(\hat{\theta}))\, \nabla p_c^\top \Gamma y = 0$ when the activation condition $\mathcal{A}$ holds, and $\nabla p_c^\top \dot{\hat{\theta}} = \nabla p_c^\top \Gamma y \le 0$ otherwise; hence the sublevel set $\{p_c \le 1\} = \Theta_\epsilon$ is forward invariant.

For (ii), when $\mathcal{A}$ does not hold, Equation (11) holds with equality. When $\mathcal{A}$ holds,

$$\tilde{\theta}^\top \Gamma^{-1} \mathrm{proj}_\Gamma(\hat{\theta}, y) - \tilde{\theta}^\top y = -\, p_c(\hat{\theta})\, \frac{\left(\tilde{\theta}^\top \nabla p_c\right) \left(\nabla p_c^\top \Gamma y\right)}{\nabla p_c^\top \Gamma \nabla p_c}. \tag{35}$$

By convexity of p_c and $p_c(\theta^*) \le 0 < p_c(\hat{\theta})$, it follows that $\tilde{\theta}^\top \nabla p_c = (\theta^* - \hat{\theta})^\top \nabla p_c(\hat{\theta}) \le p_c(\theta^*) - p_c(\hat{\theta}) < 0$. Since $p_c(\hat{\theta}) > 0$, $\nabla p_c^\top \Gamma y > 0$, and $\nabla p_c^\top \Gamma \nabla p_c > 0$, the right-hand side is nonnegative, which establishes Equation (11). ■

E Proof of the Windowed Gradient Decomposition

This appendix proves the windowed gradient decomposition of Proposition 1. Using $\hat{f} = f - \tilde{f}$ and Equation (2),

$$E(\tau; t) = \Phi(X(\tau), \theta^*) + \varepsilon(X(\tau)) - \tilde{f}(\tau) - \Phi(X(\tau), \hat{\theta}(t)). \tag{36}$$

Since Θ_ϵ is convex and contains both $\hat{\theta}(t)$ and θ^*, Taylor's theorem with the Lagrange form of the remainder yields

$$\Phi(X(\tau), \theta^*) - \Phi(X(\tau), \hat{\theta}(t)) = \Phi'(X(\tau), \hat{\theta}(t))\, \tilde{\theta}(t) + R(\tau), \qquad \|R(\tau)\| \le \tfrac{1}{2} c_{\theta\theta} \left\|\tilde{\theta}(t)\right\|^2, \tag{37}$$

where the remainder bound follows from the bound on $\partial^2 \Phi / \partial \theta^2$. Substituting into Equation (6) gives Equation (7) with Ψ as in Equation (8) and

$$\eta_f(t) = \int_{t-T}^{t} \Phi'(X(\tau), \hat{\theta}(t))^\top \left[R(\tau) + \varepsilon(X(\tau)) - \tilde{f}(\tau)\right] d\tau. \tag{38}$$

The matrix Ψ is an integral of Gram matrices and is therefore positive semi-definite with $\|\Psi\| \le T\bar{v}^2$. Using $\|\Phi'\| \le \bar{v}$, the definition of $\bar{f}_T$, and integrating over the window of length T yields Equation (9). ■

F Proof of the Drift-Window Guarantee and Its Consequences

This appendix proves Proposition 2 and states and proves its two consequences, the excitation-dependent rate and the excitation-free functional bound. The windowed excitation condition and the corresponding result are stated here. Since $\Phi'(X, \theta) \in \mathbb{R}^{n \times p}$, the integrand of Ψ has rank at most n, so a full-rank Ψ requires the Jacobian row space to rotate through all of $\mathbb{R}^p$ within each window.

Definition 6. The closed loop satisfies window persistence of excitation (WPE) if there exist $\varphi > 0$ and $t^* \ge t_1$ such that $\Psi(t) \succeq \varphi I_p$ for all $t \ge t^*$. Since $\Psi(t)$ is evaluated at $\hat{\theta}(t)$, this is a condition on the signals generated by the closed loop rather than on the state trajectory alone.

Under this condition the rate of Proposition 2 improves as follows.

Theorem 7. *Under the conditions of Proposition 2 and Definition 6, the bounds in Equation* (13) *hold for $t \ge t^*$ with t_1 replaced by t^* and the rate λ_d replaced by $\lambda_e \triangleq 2\left(k_1 + \left(\alpha - \frac{1}{2}\right)\varphi\right) \underline{\lambda}$.*

Without any excitation, the same Lyapunov argument still bounds the time-averaged first-order functional error.

Proposition 4. *Under the conditions of Proposition 2,* $\limsup_{S \to \infty} \frac{1}{S} \int_{t_1}^{t_1+S} \tilde{\theta}(r)^\top \Psi(r)\, \tilde{\theta}(r)\, dr \le c_d^\infty / (\alpha - \frac{1}{2})$.

The quantity $\tilde{\theta}(r)^\top \Psi(r)\tilde{\theta}(r)$ is the first-order functional error of the current model accumulated along the recent trajectory. Thus, up to the Taylor remainder and approximation error, Proposition 4 bounds the time-averaged squared mismatch between $\Phi(\cdot,\hat{\theta})$ and f without requiring excitation. For $p \gg n$, a bounded trajectory typically excites only a parameter subspace, making this functional statement more informative than a full-parameter claim.

Consider the Lyapunov function $V = \frac{1}{2}\tilde{\theta}^\top P(t)\tilde{\theta}$ with $P \triangleq \Gamma^{-1}$, which, by Lemma 4, satisfies the sandwich inequality

$$\tfrac{1}{2\bar{\lambda}} \left\|\tilde{\theta}\right\|^2 \le V \le \tfrac{1}{2\underline{\lambda}} \left\|\tilde{\theta}\right\|^2. \tag{39}$$

Since $\dot{P} = -P\dot{\Gamma}P$, the update Equation (10) gives $\dot{P} = -\beta_0(P - \bar{\lambda}^{-1}I_p) + \varsigma_\Gamma \Psi$. Using $\dot{\tilde{\theta}} = -\dot{\hat{\theta}}$ and the projection property Equation (11) with $y = \alpha\Xi_f - k_\sigma\hat{\theta}$,

$$\dot{V} = -\tilde{\theta}^\top P\dot{\hat{\theta}} + \tfrac{1}{2}\tilde{\theta}^\top \dot{P}\tilde{\theta} \le -\alpha\tilde{\theta}^\top\Xi_f + k_\sigma\tilde{\theta}^\top\hat{\theta} - \tfrac{\beta_0}{2}\tilde{\theta}^\top(P - \bar{\lambda}^{-1}I_p)\tilde{\theta} + \tfrac{\varsigma_\Gamma}{2}\tilde{\theta}^\top\Psi\tilde{\theta}. \tag{40}$$

Substituting Equation (7), using $\hat{\theta} = \theta^* - \tilde{\theta}$, $0 \le \varsigma_\Gamma \le 1$, $\Psi \succeq 0$, and $P - \bar{\lambda}^{-1}I_p \succeq 0$ (Lemma 4),

$$\dot{V} \le -\left(\alpha - \tfrac{1}{2}\right)\tilde{\theta}^\top\Psi\tilde{\theta} - k_\sigma \left\|\tilde{\theta}\right\|^2 + k_\sigma\tilde{\theta}^\top\theta^* - \alpha\tilde{\theta}^\top\eta_f. \tag{41}$$

Using Young's inequality in the form $ab \le \frac{a^2}{4} + b^2$, the sigma-modification cross term is bounded as $k_\sigma\tilde{\theta}^\top\theta^* \le \frac{k_\sigma}{4}\left\|\tilde{\theta}\right\|^2 + k_\sigma\bar{\theta}^2$. Using Equation (9) and the projection bound $\left\|\tilde{\theta}\right\| \le \bar{\tilde{\theta}}$,

$$\begin{aligned}\alpha\left|\tilde{\theta}^\top\eta_f\right| &\le \alpha c_{\eta 1}\bar{\tilde{\theta}}\left\|\tilde{\theta}\right\|^2 + \alpha\left(c_{\eta 2}\bar{f}_T(t) + c_{\eta 3}\right)\left\|\tilde{\theta}\right\| \\ &\le \left(\alpha c_{\eta 1}\bar{\tilde{\theta}} + \tfrac{k_\sigma}{4}\right)\left\|\tilde{\theta}\right\|^2 + \tfrac{\alpha^2}{k_\sigma}\left(c_{\eta 2}\bar{f}_T(t) + c_{\eta 3}\right)^2,\end{aligned} \tag{42}$$

where Young's inequality with parameter $k_\sigma/2$ was applied to the linear term. Collecting terms in Equation (41) yields

$$\dot{V} \le -\left(\alpha - \tfrac{1}{2}\right)\tilde{\theta}^\top\Psi\tilde{\theta} - k_1\left\|\tilde{\theta}\right\|^2 + c_d(t), \tag{43}$$

with $k_1 = \frac{k_\sigma}{2} - \alpha c_{\eta 1}\bar{\tilde{\theta}}$, which is positive by the gain condition of Proposition 2 since $2\alpha c_{\eta 1}\bar{\tilde{\theta}} = \alpha T\bar{v}c_{\theta\theta}\bar{\tilde{\theta}}$, and $c_d(t)$ as in Equation (14).

Proof of Proposition 2. Dropping the first (non-positive) term in Equation (43) and using Equation (39) in the form $\left\|\tilde{\theta}\right\|^2 \ge 2\underline{\lambda}V$,

$$\dot{V} \le -\lambda_d V + c_d(t), \qquad \lambda_d = 2k_1\underline{\lambda}. \tag{44}$$

The comparison lemma gives, for $t \ge t_1$,

$$V(t) \le V(t_1)e^{-\lambda_d(t-t_1)} + \int_{t_1}^{t} e^{-\lambda_d(t-s)}c_d(s)\,ds. \tag{45}$$

Since $\bar{f}_T$ is nonincreasing by the definition of $\bar{f}_T$, c_d is nonincreasing with $\lim_{t\to\infty} c_d(t) = c_d^\infty$, so the integral is bounded by $c_d(t_1)/\lambda_d$ and its limit superior by c_d^∞/λ_d. Converting through Equation (39), $\left\|\tilde{\theta}(t)\right\|^2 \le 2\bar{\lambda}V(t)$ and $V(t_1) \le \left\|\tilde{\theta}(t_1)\right\|^2/(2\underline{\lambda})$, which yields both bounds in Equation (13). ■

Remark. Since $c_d^\infty \ge k_\sigma\bar{\theta}^2$ and $k_1 < k_\sigma/2$, the excitation-free radius in Equation (13) satisfies $\frac{2\bar{\lambda}}{\lambda_d}c_d^\infty = \frac{\bar{\lambda}}{k_1\underline{\lambda}}c_d^\infty > 2\frac{\bar{\lambda}}{\underline{\lambda}}\bar{\theta}^2$, i.e., it exceeds $\bar{\theta}\sqrt{2\bar{\lambda}/\underline{\lambda}} > 2\bar{\theta}$ (for the diagnostic gains $\bar{\lambda}/\underline{\lambda} = 20$, more than $6\bar{\theta}$), whereas the projection alone gives $\left\|\tilde{\theta}\right\| \le 2\bar{\theta} + \epsilon_\theta$. The same computation for Equation (23) gives a radius of at least $\sqrt{2}\,\bar{\theta}$. This is the usual sigma-modification tradeoff: without excitation, Propositions 2 and 3 certify invariance and the exponential rate into the residual set, Proposition 4 certifies the averaged functional error, and parameter localization requires excitation (Theorem 7).

Proof of Proposition 4. Retaining the first term in Equation (43) and dropping $-k_1 \left\|\tilde{\theta}\right\|^2$, integration over $[t_1, t_1 + S]$ gives

$$\begin{aligned}\left(\alpha - \tfrac{1}{2}\right) \int_{t_1}^{t_1+S} \tilde{\theta}(r)^\top \Psi(r) \tilde{\theta}(r)\, dr \\ \leq V(t_1) - V(t_1 + S) + \int_{t_1}^{t_1+S} c_d(r)\, dr \\ \leq \tfrac{\bar{\tilde{\theta}}^2}{2\underline{\lambda}} + \int_{t_1}^{t_1+S} c_d(r)\, dr,\end{aligned} \tag{46}$$

where $V \geq 0$ and Equation (39) with $\left\|\tilde{\theta}\right\| \leq \bar{\tilde{\theta}}$ were used. Dividing by S and letting $S \to \infty$, the first term vanishes and the Cesaro mean of the nonincreasing c_d converges to c_d^∞, which establishes the bound in Proposition 4. ■

Proof of Theorem 7. For $t \geq t^*$, $\tilde{\theta}^\top \Psi \tilde{\theta} \geq \varphi \left\|\tilde{\theta}\right\|^2$, so Equation (43) yields $\dot{V} \leq -\left(k_1 + (\alpha - \tfrac{1}{2})\varphi\right) \left\|\tilde{\theta}\right\|^2 + c_d(t) \leq -\lambda_e V + c_d(t)$, and the argument of Proposition 2 applies verbatim with λ_d replaced by λ_e. ■

G Adjoint Representation of the Trajectory Gradient

This appendix states and proves the adjoint representation of the trajectory gradient used in Section 3.2.

Theorem 8. *Let Equation* (17) *hold, let $\hat{X}_p$ satisfy Equation* (15)*, and let ν satisfy the backward equation in Equation* (18) *with $\nu(t) = 0_{2n}$. Then $\nabla_{\hat{\theta}} \mathcal{E}_T(t)$ equals the integral in Equation* (18)*.*

Let $S(\tau) = \partial \hat{X}_p(\tau; t)/\partial \hat{\theta}$ denote the sensitivity, which satisfies $S' = \frac{\partial F}{\partial \chi}(\hat{X}_p) S + \frac{\partial F}{\partial \theta}(\hat{X}_p)$ with $S(t - T) = 0$, since the initial condition $X(t - T)$ does not depend on $\hat{\theta}$. By the chain rule,

$$\nabla_{\hat{\theta}} \mathcal{E}_T = \int_{t-T}^{t} S(\tau)^\top e(\tau; t)\, d\tau. \tag{47}$$

Consider $\int_{t-T}^{t} \nu^\top \left(S' - \frac{\partial F}{\partial \chi} S - \frac{\partial F}{\partial \theta}\right) d\tau = 0$. Integrating $\nu^\top S'$ by parts and using $\nu(t) = 0$ and $S(t - T) = 0$ eliminates the boundary terms, leaving

$$\int_{t-T}^{t} \left(-\dot{\nu}^\top - \nu^\top \tfrac{\partial F}{\partial \chi}\right) S\, d\tau = \int_{t-T}^{t} \nu^\top \tfrac{\partial F}{\partial \theta}\, d\tau. \tag{48}$$

Selecting ν to satisfy the backward equation in Equation (18) makes the left integrand equal $e^\top S$, so Equation (47) equals $\int_{t-T}^{t} \left(\frac{\partial F}{\partial \theta}\right)^\top \nu\, d\tau$, which is Equation (18). ■

H Containment and Proof of the Trajectory Decomposition

This appendix proves that the virtual trajectory stays in Ω and then proves the trajectory decomposition of Theorem 1. The containment lemma referenced in Section 3.2 is stated first.

Lemma 9. *Let Assumptions 1 and 2 hold, let $\hat{\theta}(t) \in \Theta_\epsilon$, and let the window condition in Equation* (17) *hold. Then $\hat{X}_p(\tau; t) \in \Omega$ and $\|e(\tau; t)\| < \varpi$ for all $\tau \in [t - T, t]$.*

Proof of Lemma 9. Let $\tau^* \triangleq \sup\{\tau \in [t - T, t] : \|e(s; t)\| \leq \varpi \ \forall s \in [t - T, \tau]\}$, which is well defined with $\tau^* > t - T$ by continuity and $e(t - T; t) = 0$. For $\tau \leq \tau^*$, both $\hat{X}_p(\tau)$ and the segment between $X(\tau)$ and $\hat{X}_p(\tau)$ lie in Ω. On this interval,

$$\begin{aligned}\dot{e} &= F(\hat{X}_p, \hat{\theta}, \tau) - F(X, \hat{\theta}, \tau) + F(X, \hat{\theta}, \tau) - \begin{bmatrix} \dot{x} \\ f(X) + g(X)u \end{bmatrix} \\ &= F(\hat{X}_p, \hat{\theta}, \tau) - F(X, \hat{\theta}, \tau) - B\left(f(X) - \Phi(X, \hat{\theta})\right),\end{aligned} \tag{49}$$

so, using the mean value bound $\left\|F(\hat{X}_p,\hat{\theta},\tau)-F(X,\hat{\theta},\tau)\right\| \le L_F\,\|e\|$ and, by Equation (2) and the mean value theorem on the segment $[\hat{\theta},\theta^*]\subset\Theta_\epsilon$, $\left\|f(X)-\Phi(X,\hat{\theta})\right\| \le \bar{v}\left\|\tilde{\theta}\right\|+\bar{\varepsilon}\le \bar{v}\bar{\tilde{\theta}}+\bar{\varepsilon}$, the Gronwall inequality yields

$$\|e(\tau;t)\| \le \frac{e^{L_F(\tau-(t-T))}-1}{L_F}\left(\bar{v}\bar{\tilde{\theta}}+\bar{\varepsilon}\right) \le \kappa_T\left(\bar{v}\bar{\tilde{\theta}}+\bar{\varepsilon}\right) < \varpi, \tag{50}$$

by Equation (17). If $\tau^* < t$, then continuity would require $\|e(\tau^*;t)\| = \varpi$, contradicting Equation (50); hence $\tau^* = t$ and the claim follows. ■

The operators used in Theorem 1 are defined as follows. Let $X_s(\tau) \triangleq X(\tau)+s\,e(\tau;t)$ for $s\in[0,1]$ denote the segment between the measured and predicted states, and define the mean-value system matrix and its transition matrix as

$$\bar{A}(\tau) \triangleq \int_0^1 \frac{\partial F}{\partial \chi}\left(X_s(\tau),\,\hat{\theta}(t),\,\tau\right) ds, \qquad \frac{\partial}{\partial\tau}\Phi_A(\tau,s) = \bar{A}(\tau)\,\Phi_A(\tau,s), \quad \Phi_A(s,s)=I_{2n}. \tag{51}$$

Define the operators

$$W(\tau) \triangleq \int_{t-T}^{\tau} \Phi_A(\tau,s)\,B\,\Phi'(X(s),\hat{\theta}(t))\,ds, \tag{52}$$

$$d(\tau) \triangleq \int_{t-T}^{\tau} \Phi_A(\tau,s)\,B\,[R(s)+\varepsilon(X(s))]\,ds, \tag{53}$$

where $R(s) \triangleq \Phi(X(s),\theta^*)-\Phi(X(s),\hat{\theta}(t))-\Phi'(X(s),\hat{\theta}(t))\,\tilde{\theta}(t)$ denotes the Taylor remainder, which satisfies $\|R(s)\| \le \frac{1}{2}c_{\theta\theta}\left\|\tilde{\theta}(t)\right\|^2$ by the smoothness of Φ. Let $S(\tau) \triangleq \partial\hat{X}_p(\tau;t)/\partial\hat{\theta}$ denote the sensitivity of the virtual trajectory, which satisfies $S' = (\partial F/\partial\chi)(\hat{X}_p)\,S + B\,\Phi'(\hat{X}_p,\hat{\theta})$ with $S(t-T)=0$. Since $F(\hat{X}_p,\hat{\theta})-F(X,\hat{\theta}) = \bar{A}\,e$ along the segment, the error dynamics are linear in e with forcing $-B(\Phi'\tilde{\theta}+R+\varepsilon)$, which yields the following representation.

The constants in Equation (21) are $d_2 \triangleq \frac{1}{2}c_{\theta\theta}\kappa_T$, $d_0 \triangleq \kappa_T\bar{\varepsilon}$, $w_e \triangleq \bar{w}+d_2\bar{\tilde{\theta}}$, $D_S \triangleq Te^{L_FT}\,(c_{X\theta}+L_{FX}\,T\,\bar{v})$,

$$b_1 \triangleq T\left(\bar{w}\,d_2+D_S\,w_e^2\right), \qquad b_2 \triangleq 2\,T\,D_S\,w_e\,d_0, \qquad b_3 \triangleq T\,d_0\,(\bar{w}+D_S\,d_0)\,. \tag{54}$$

Proof of Theorem 1. By Lemma 9, the segment $X(\tau)+s\,e(\tau;t)$, $s\in[0,1]$, remains in Ω, so $\bar{A}$ in Equation (51) is well defined with $\left\|\bar{A}(\tau)\right\| \le L_F$, and by the fundamental theorem of calculus, $F(\hat{X}_p,\hat{\theta},\tau)-F(X,\hat{\theta},\tau) = \bar{A}(\tau)e(\tau;t)$ exactly. Therefore

$$\dot{e} = \bar{A}(\tau)e - B\left(\Phi'(X(\tau),\hat{\theta})\tilde{\theta}+R(\tau)+\varepsilon(X(\tau))\right), \tag{55}$$

using Equation (37) and Equation (2), and variation of constants with $e(t-T;t)=0$ gives the exact representation $e(\tau;t) = -\left(W(\tau)\tilde{\theta}+d(\tau)\right)$ with W and d as in Equation (52)–Equation (53). Since $\|\Phi_A(\tau,s)\| \le e^{L_F(\tau-s)}$,

$$\|W(\tau)\| \le \bar{v}\kappa_T = \bar{w}, \qquad \|d(\tau)\| \le \kappa_T\left(\tfrac{1}{2}c_{\theta\theta}\left\|\tilde{\theta}\right\|^2+\bar{\varepsilon}\right) = d_2\left\|\tilde{\theta}\right\|^2+d_0, \tag{56}$$

and consequently, using $\left\|\tilde{\theta}\right\| \le \bar{\tilde{\theta}}$,

$$\|e(\tau;t)\| \le \left(\bar{w}+d_2\bar{\tilde{\theta}}\right)\left\|\tilde{\theta}\right\|+d_0 = w_e\left\|\tilde{\theta}\right\|+d_0. \tag{57}$$

Next, the exact sensitivity S and the operator W are compared. Let $A^v(\tau) \triangleq \frac{\partial F}{\partial\chi}(\hat{X}_p(\tau),\hat{\theta},\tau)$ with transition matrix $\Phi_v(\tau,s)$, so that

$$S(\tau) = \int_{t-T}^{\tau} \Phi_v(\tau,s)B\,\Phi'(\hat{X}_p(s),\hat{\theta})\,ds.$$

Since A^v and $\bar{A}$ are evaluated at points within distance $\|e\|$ of one another inside Ω, the bound on $\partial^2\Phi/\partial X\partial\theta$ and the definition of L_{FX} give $\left\|A^v(\tau) - \bar{A}(\tau)\right\| \le L_{FX}\,\|e(\tau;t)\|$. Writing $D(\tau,s) \triangleq \Phi_v(\tau,s) - \Phi_A(\tau,s)$, the difference satisfies $\partial_\tau D = A^v D + (A^v - \bar{A})\Phi_A$ with $D(s,s) = 0$, so

$$\|D(\tau,s)\| \le \int_s^\tau e^{L_F(\tau-r)}\, L_{FX}\,\|e(r;t)\|\, e^{L_F(r-s)}\, dr \le L_{FX}\, T\, e^{L_F T} \sup_r \|e(r;t)\|\,. \tag{58}$$

Similarly, $\left\|\Phi'(\hat{X}_p(s),\hat{\theta}) - \Phi'(X(s),\hat{\theta})\right\| \le c_{X\theta}\,\|e(s;t)\|$. Combining,

$$\|S(\tau) - W(\tau)\| \le \int_{t-T}^{\tau} \left(\|D(\tau,s)\|\,\bar{v} + e^{L_F T} c_{X\theta}\,\|e(s;t)\|\right) ds \le D_S \sup_{r\in[t-T,t]} \|e(r;t)\|\,, \tag{59}$$

with $D_S = Te^{L_F T}\,(c_{X\theta} + L_{FX} T\bar{v})$.

By Theorem 8 and Equation (47), the adjoint signal satisfies exactly

$$\begin{aligned}\Xi_a &= -\int_{t-T}^{t} S(\tau)^\top e(\tau;t)\, d\tau = \int_{t-T}^{t} S(\tau)^\top \left(W(\tau)\tilde{\theta} + d(\tau)\right) d\tau \\ &= H\tilde{\theta} + \int_{t-T}^{t} \left[W^\top d + (S-W)^\top \left(W\tilde{\theta} + d\right)\right] d\tau,\end{aligned} \tag{60}$$

which is Equation (20) with η_a equal to the last integral. Using Equation (56), Equation (57), Equation (59), $\left\|W\tilde{\theta} + d\right\| = \|e\|$, and $\left\|\tilde{\theta}\right\| \le \bar{\tilde{\theta}}$ to fold cubic and quartic powers,

$$\begin{aligned}\|\eta_a\| &\le T\left[\bar{w}\left(d_2\left\|\tilde{\theta}\right\|^2 + d_0\right) + D_S\left(w_e\left\|\tilde{\theta}\right\| + d_0\right)^2\right] \\ &\le b_1\left\|\tilde{\theta}\right\|^2 + b_2\left\|\tilde{\theta}\right\| + b_3,\end{aligned} \tag{61}$$

with b_1, b_2, b_3 as in Equation (54). Finally, H is an integral of Gram matrices, hence positive semi-definite, with $\|H\| \le T\bar{w}^2$. ■

I Proof of the Adjoint-Window Guarantee and the Window-Length Condition

This appendix proves Proposition 3 and the explicit window-length condition referenced in Section 3.2. The window-length corollary is stated first.

Corollary 10. *Let $L_F T \le 1$. Then the gain condition of Proposition 3 holds whenever $T \le (k_\sigma/(2\,\alpha\,C_b))^{1/3}$, where C_b is the computable constant defined in Appendix I, obtained by majorizing $\kappa_T \le e\,T$ and $e^{L_F T} \le e$ in Equation* (54).

Consider $V_a = \frac{1}{2}\tilde{\theta}^\top\Gamma_0^{-1}\tilde{\theta}$, which satisfies $\frac{1}{2\bar{\lambda}}\left\|\tilde{\theta}\right\|^2 \le V_a \le \frac{1}{2\underline{\lambda}}\left\|\tilde{\theta}\right\|^2$. Since Γ_0 is constant, using the projection property Equation (11) with $y = \alpha\Xi_a - k_\sigma\hat{\theta}$ and the decomposition Equation (20),

$$\begin{aligned}\dot{V}_a &\le -\alpha\tilde{\theta}^\top H\tilde{\theta} - \alpha\tilde{\theta}^\top\eta_a + k_\sigma\tilde{\theta}^\top\theta^* - k_\sigma\left\|\tilde{\theta}\right\|^2 \\ &\le -\alpha\tilde{\theta}^\top H\tilde{\theta} - k_\sigma\left\|\tilde{\theta}\right\|^2 + k_\sigma\tilde{\theta}^\top\theta^* \\ &\quad + \alpha\left(b_1\bar{\tilde{\theta}} + b_2\right)\left\|\tilde{\theta}\right\|^2 + \alpha b_3\left\|\tilde{\theta}\right\|,\end{aligned} \tag{62}$$

using Equation (21) and $\left\|\tilde{\theta}\right\| \le \bar{\tilde{\theta}}$. Applying Young's inequality as in Appendix F, $k_\sigma\tilde{\theta}^\top\theta^* \le \frac{k_\sigma}{4}\left\|\tilde{\theta}\right\|^2 + k_\sigma\bar{\theta}^2$ and $\alpha b_3\left\|\tilde{\theta}\right\| \le \frac{k_\sigma}{4}\left\|\tilde{\theta}\right\|^2 + \frac{\alpha^2 b_3^2}{k_\sigma}$. Dropping the non-positive term $-\alpha\tilde{\theta}^\top H\tilde{\theta}$ yields

$$\dot{V}_a \le -k_{1a}\left\|\tilde{\theta}\right\|^2 + c_a \le -\lambda_a V_a + c_a, \tag{63}$$

with $k_{1a} = \frac{k_\sigma}{2} - \alpha(b_1\bar{\tilde{\theta}} + b_2) > 0$ by the gain condition of Proposition 3 and $\lambda_a = 2k_{1a}\underline{\lambda}$. The comparison lemma and the sandwich inequality then give Equation (23). If WPE holds for H, that is $H(t) \succeq \varphi_a I_p$, the rate improves to $2(k_{1a} + \alpha\varphi_a)\underline{\lambda}$ by retaining the dropped term. ■

Proof of Corollary 10. For $L_F T \le 1$, the majorizations $\kappa_T = (e^{L_F T} - 1)/L_F \le (e-1)T < eT$, $e^{L_F T} \le e$, and $T \le 1/L_F$ hold, the first strictly for $T > 0$. Substituting into the definitions in Equation (54) gives $\bar{w} \le e\bar{v}T$, $d_2 \le \frac{ec_{\theta\theta}}{2}T$, $d_0 \le e\bar{\varepsilon}T$, $w_e \le e\, w_1 T$ with $w_1 \triangleq \bar{v} + \frac{c_{\theta\theta}\bar{\tilde{\theta}}}{2}$, and $D_S \le eM_S T$ with $M_S \triangleq c_{X\theta} + L_{FX}\bar{v}/L_F$. Substituting into Equation (54) yields

$$b_1\bar{\tilde{\theta}} + b_2 \le \frac{e^2\bar{v}c_{\theta\theta}\bar{\tilde{\theta}}}{2}T^3 + e^3 M_S\, w_1 \left(\bar{\tilde{\theta}}\, w_1 + 2\bar{\varepsilon}\right) T^4 \le C_b\, T^3, \tag{64}$$

where the factor T^4 was majorized by T^3/L_F using $T \le 1/L_F$, and

$$C_b \triangleq \frac{e^2\,\bar{v}\,c_{\theta\theta}\,\bar{\tilde{\theta}}}{2} + \frac{e^3\,M_S\,w_1}{L_F}\left(\bar{\tilde{\theta}}\, w_1 + 2\bar{\varepsilon}\right). \tag{65}$$

Since every term of $b_1\bar{\tilde{\theta}} + b_2$ contains at least one factor κ_T (through $\bar{w}$, d_2, d_0, or w_e), and $\kappa_T < eT$ strictly, $b_1\bar{\tilde{\theta}} + b_2 < C_b T^3$ for $T > 0$. Hence the window bound in Corollary 10 implies $2\alpha\left(b_1\bar{\tilde{\theta}} + b_2\right) < 2\alpha C_b T^3 \le k_\sigma$, which is the strict gain condition of Proposition 3. ■

J Stored-Set Extension, Gauss–Newton Bound, and Proof of the NODE-CL Guarantee

This appendix states and proves the stored-set extension of Proposition 2, the bound relating the Gauss–Newton matrix to the trajectory regressor, and Theorem 2.

Proposition 5. *Let $\mathcal{D}(t) = \{(X_j, \hat{f}(\tau_j))\}_{j=1}^N$ be a set of N stored pairs recorded at times $\tau_j \le t$, with membership piecewise constant in t, and replace the integrals in Equation* (6)*, Equation* (8)*, and Equation* (10) *by $\Xi_\mathcal{D} \triangleq \frac{T}{N}\sum_{j=1}^N \Phi'(X_j,\hat{\theta})^\top(\hat{f}(\tau_j) - \Phi(X_j,\hat{\theta}))$ and $\Psi_\mathcal{D} \triangleq \frac{T}{N}\sum_{j=1}^N \Phi'(X_j,\hat{\theta})^\top\Phi'(X_j,\hat{\theta})$. Suppose every admitted sample satisfies $\left\|\tilde{f}(\tau_j)\right\| \le \delta_\mathcal{D}$, which holds with $\delta_\mathcal{D} = \|z_o(t_0)\|\, e^{-k_o(t_a - t_0)} + \bar{z}_o$ when admission starts at $t_a \ge t_0$. Let $t_s \ge t_1$ denote the time at which the stored-set law is activated, with the stack non-empty, $\hat{\theta}(t_s) \in \Theta_\epsilon$, and $\underline{\lambda} I_p \preceq \Gamma(t_s) \preceq \bar{\lambda} I_p$. Then Proposition 1, Proposition 2, and Proposition 4 hold for $t \ge t_s$ with t_1 replaced by t_s and with $\bar{f}_T(t)$ and $\bar{z}_o$ both replaced by $\delta_\mathcal{D}$, so that $c_d(t) = c_d^\infty = k_\sigma\bar{\theta}^2 + \frac{\alpha^2}{k_\sigma}(c_{\eta 2}\delta_\mathcal{D} + c_{\eta 3})^2$ is constant.*

A stored sample retains the observer error of its recording time, so the decaying window envelope becomes a uniform bound over admitted samples. The law of Hart et al. (2025a) is a stored-set drift law of this type, whereas fixing Γ recovers its constant-gain simplification, the DNN analogue of the stack of Chowdhary & Johnson (2011). The additions analyzed here are the certified gain interval, gain-metric projection, and excitation-free functional bound; the finite-excitation convergence result of Chowdhary & Johnson (2011) is complementary.

Proof of Proposition 5. The expansion in the proof of Proposition 1 is performed sample by sample, so replacing the integral over $[t-T, t]$ by the average over $\mathcal{D}(t)$ scaled by T gives $\Xi_\mathcal{D} = \Psi_\mathcal{D}\tilde{\theta} + \eta_\mathcal{D}$ with $\Psi_\mathcal{D} \succeq 0$, $\|\Psi_\mathcal{D}\| \le T\bar{v}^2$, and $\|\eta_\mathcal{D}\| \le c_{\eta 1}\left\|\tilde{\theta}\right\|^2 + c_{\eta 2}\delta_\mathcal{D} + c_{\eta 3}$, since each stored $\hat{f}(\tau_j)$ carries the observer error at its recording time and every admitted sample is bounded by $\delta_\mathcal{D}$; the bound on $\delta_\mathcal{D}$ under an admission time follows from Equation (30). The proofs of Lemma 4, Proposition 2, and Proposition 4 use only positive semi-definiteness and boundedness of the regressor and the form of the perturbation bound, with the constant c_d^∞ in place of the nonincreasing $c_d(t)$, so they apply verbatim. Membership changes make $\Psi_\mathcal{D}$ piecewise continuous in t; V is continuous and the inequality on $\dot{V}$ holds almost everywhere, so the comparison lemma applies unchanged. ■

The computable Gauss–Newton matrix is related to the trajectory regressor by the following bound.

Lemma 11. *Under the conditions of Theorem 1 at segment length T_s, $\|G_{\mathcal{S}} - H_{\mathcal{S}}\| \le g_2 \left\|\tilde{\theta}\right\|^2 + g_1 \left\|\tilde{\theta}\right\| + g_0$ with $g_2 \triangleq T_s D_S^2 w_e^2$, $g_1 \triangleq 2T_s D_S w_e(\bar{w} + D_S d_0)$, and $g_0 \triangleq T_s D_S d_0(2\bar{w} + D_S d_0)$, where $\bar{w}, w_e, d_0, D_S$ are the constants of Appendix H at T_s.*

Proof of Lemma 11. By Equation (59) and Equation (57) applied to each segment at length T_s, $\|S_j - W_j\| \le D_S(w_e \left\|\tilde{\theta}\right\| + d_0) =: \Delta$, while $\|W_j\| \le \bar{w}$ and hence $\|S_j\| \le \bar{w} + \Delta$. Hence $\left\|S_j^\top S_j - W_j^\top W_j\right\| = \left\|(S_j - W_j)^\top S_j + W_j^\top (S_j - W_j)\right\| \le \Delta(2\bar{w} + \Delta)$, and integrating over a segment of length T_s and averaging over segments gives $\|G_{\mathcal{S}} - H_{\mathcal{S}}\| \le T_s\Delta(2\bar{w} + \Delta)$; expanding Δ in powers of $\left\|\tilde{\theta}\right\|$ yields the stated constants. ■

Proof of Theorem 2. Lemma 4 holds because $G_{\mathcal{S}} \succeq 0$ and its proof uses nothing else. With $V = \frac{1}{2}\tilde{\theta}^\top \Gamma^{-1}\tilde{\theta}$ and Equation (11),

$$\dot{V} \le -\alpha\tilde{\theta}^\top H_{\mathcal{S}}\tilde{\theta} - \alpha\tilde{\theta}^\top \eta_{\mathcal{S}} + k_\sigma \tilde{\theta}^\top \theta^* - k_\sigma \left\|\tilde{\theta}\right\|^2 + \tfrac{1}{2}\varsigma_\Gamma \tilde{\theta}^\top H_{\mathcal{S}}\tilde{\theta} + \tfrac{1}{2}\varsigma_\Gamma \tilde{\theta}^\top (G_{\mathcal{S}} - H_{\mathcal{S}})\tilde{\theta},$$

where the last two terms come from $\frac{1}{2}\tilde{\theta}^\top \frac{d}{dt}(\Gamma^{-1})\tilde{\theta}$ with the term $-\frac{\beta_0}{2}\tilde{\theta}^\top(\Gamma^{-1} - \bar{\lambda}^{-1} I_p)\tilde{\theta} \le 0$ dropped. Since $0 \le \varsigma_\Gamma \le 1$, the $H_{\mathcal{S}}$ terms combine into $-(\alpha - \frac{1}{2})\tilde{\theta}^\top H_{\mathcal{S}}\tilde{\theta}$, and Lemma 11 with $\left\|\tilde{\theta}\right\| \le \bar{\tilde{\theta}}$ bounds the mismatch term by $\frac{1}{2}(g_2\bar{\tilde{\theta}}^2 + g_1\bar{\tilde{\theta}} + g_0)\left\|\tilde{\theta}\right\|^2$. Bounding the remaining terms as in the proof of Proposition 3 and using $H_{\mathcal{S}} \succeq \varphi_s I_p$ gives

$$\dot{V} \le -k_s(\varphi_s)\left\|\tilde{\theta}\right\|^2 + c_a \le -2k_s(\varphi_s)\underline{\lambda}V + c_a = -\lambda_s(\varphi_s)V + c_a.$$

The comparison lemma and $\left\|\tilde{\theta}\right\|^2/(2\bar{\lambda}) \le V \le \left\|\tilde{\theta}\right\|^2/(2\underline{\lambda})$ yield Equation (27). ■

K Simulation Details, Additional Ablations, and Verification of the Certified Inequalities

The diagnostic results are reported in Table 2 and Figures 3, 4 and 7 to 11; the full benchmark results in Table 3 and Figures 5 and 6.

Plants and excitation. In S1, the plant is $\ddot{x} = \Phi(X, \theta_{\mathrm{tr}}) + u$ with $n = m = 1$, where Φ is a single-hidden-layer $\tanh$ network with input $X \in \mathbb{R}^2$, six hidden neurons, and scalar output ($p = 25$), and θ_{tr} is drawn per seed with hidden weights and biases from $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 0.25)$ and output weights and bias from $\mathcal{N}(0, 0.64)$ and $\mathcal{N}(0, 0.04)$. The input is $u = \sum_{k=1}^{7} a_k \sin(2\pi\omega_k t + \phi_k) - x - 1.5\dot{x}$ with $\omega_k \in \{0.20, 0.31, 0.47, 0.71, 1.03, 1.51, 2.23\}$ Hz, amplitudes $a_k = 1.1\,\xi_k$ with $\xi_k \sim \mathcal{U}[0.5, 1]$, and phases $\phi_k \sim \mathcal{U}[0, 2\pi]$. The estimate is initialized at $\hat{\theta}(t_0) = \theta_{\mathrm{tr}} + \delta$ with δ Gaussian and rescaled to $\|\delta\| = 0.4\,\|\theta_{\mathrm{tr}}\|$, and $\bar{\theta} = 2\,\|\theta_{\mathrm{tr}}\|$, $\epsilon_\theta = 0.5$. In S2, the manipulator model is $M(q)\ddot{q} + C(q, \dot{q})\dot{q} + F_d\dot{q} = \tau$ with $M = [p_1 + 2p_3c_2,\ p_2 + p_3c_2;\ p_2 + p_3c_2,\ p_2]$, $C = [-p_3s_2\dot{q}_2,\ -p_3s_2(\dot{q}_1 + \dot{q}_2);\ p_3s_2\dot{q}_1,\ 0]$, $c_2 = \cos q_2$, $s_2 = \sin q_2$, $p_1 = 3.473$, $p_2 = 0.196$, $p_3 = 0.242$, and $F_d = \mathrm{diag}(5.3, 1.1)$. The input is a five-component multisine per joint with frequencies $\{0.13, 0.29, 0.53, 0.97, 1.61\}$ Hz and $\{0.17, 0.37, 0.67, 1.19, 1.87\}$ Hz, amplitudes $4\xi_k$ and $2\xi_k$, and the baseline $-3q - 3\dot{q}$. The network has input $X \in \mathbb{R}^4$, twelve hidden neurons, and two outputs ($p = 86$), initialized with entries from $\mathcal{N}(0, 0.0025)$; $\bar{\theta} = 40$ and $\epsilon_\theta = 2$. The held-out excitation uses a different draw of amplitudes and phases.

Gains and discretization (S1 and S2). The observer gains are $\alpha_o = k_f = 200$, the learning gains are $\alpha = 2$ (drift, single-step, CL) and $\alpha = 6$ (adjoint), and the gain matrix parameters of the windowed drift law are $\Gamma_0 = 5I_p$, $\bar{\lambda} = 10$, $\underline{\lambda} = 0.5$, $\beta_0 = 0.05$, and $k_\sigma = 10^{-4}$ (the CL-LS and stored-segment gains are given under *Stored-set methods* below). The plant is integrated by fourth-order Runge-Kutta at $\Delta t = 1$ ms. The observer is integrated by fourth-order Runge-Kutta over each sampling interval from the sampled velocity, with the velocity extrapolated within the interval as $\dot{x}(t_k) + (\hat{f}(t_k) + g(X(t_k))u(t_k))(\tau - t_k)$ (predictive hold); the zero-order hold variant in Figure 11(b) uses $\dot{x}(t_k)$ throughout the interval. The drift-law, single-step, and CL updates are applied by forward Euler at 1 ms in S1 and 5 ms in S2, and the adjoint and Adam updates at 10 ms

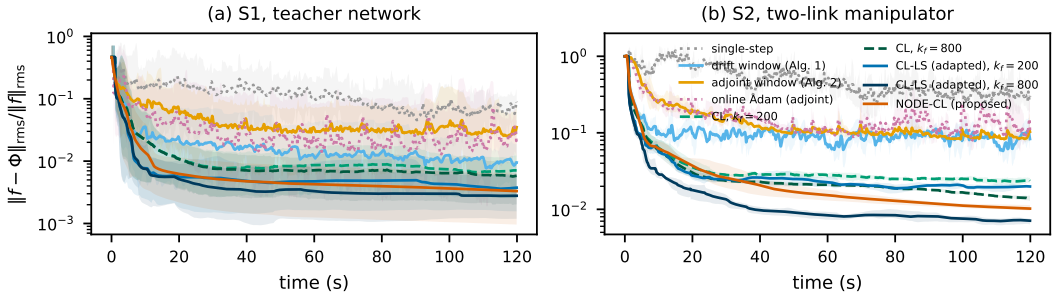

Figure 3: Normalized functional error over time (three seeds, geometric mean with seed range) for (a) S1 and (b) S2.

Table 2: Diagnostic studies S1 and S2 (three seeds, 120 s; these runs use the scalar bounded-gain forgetting of Slotine & Li (1989) in the least-squares laws): terminal normalized functional error (geometric mean, seed range); final parameter error ratio (S1); held-out 4 s prediction error (S2; initially 14.3); update cost in ms on one CPU core, excluding stack selection. All rows except online Adam are sampled implementations of continuous-time laws covered by the analysis under its stated assumptions and gain conditions.

	S1 (teacher, $p = 25$)		S2 (manipulator, $p = 86$)		cost (ms)	
Method	func. error [range]	$\|\tilde\theta(t_f)\| \,/\, \|\tilde\theta(0)\|$	func. error [range]	pred. (4 s)	S1	S2
single-step	6.8e−2 [3.1e−2,1.1e−1]	0.89	3.2e−1 [2.2e−1,5.1e−1]	0.293	0.03	0.04
drift window (Alg. 1)	9.3e−3 [3.7e−3,1.7e−2]	0.80	7.8e−2 [6.1e−2,1.1e−1]	0.119	0.15	0.81
adjoint window (Alg. 2)	2.9e−2 [6.8e−3,7.3e−2]	0.85	8.9e−2 [8.0e−2,9.9e−2]	0.122	1.4	2.5
online Adam (adjoint)	2.4e−2 [9.0e−3,7.7e−2]	1.31	1.2e−1 [9.4e−2,1.8e−1]	0.157	1.4	2.5
CL, $k_f = 200$	6.9e−3 [2.6e−3,1.2e−2]	0.83	2.3e−2 [2.1e−2,2.6e−2]	0.0375	0.06	0.14
CL, $k_f = 800$	5.9e−3 [2.1e−3,9.9e−3]	0.83	1.4e−2 [1.4e−2,1.5e−2]	0.0399	0.06	0.14
CL-LS (adapted), $k_f = 200$	3.7e−3 [1.6e−3,6.0e−3]	0.93	2.0e−2 [1.8e−2,2.1e−2]	0.0232	0.14	0.78
CL-LS (adapted), $k_f = 800$	**2.8e−3** [1.2e−3,6.2e−3]	0.92	**7.1e−3** [6.3e−3,7.7e−3]	0.0249	0.14	0.78
NODE-CL (proposed)	3.4e−3 [9.8e−4,6.6e−3]	0.80	1.0e−2 [9.9e−3,1.2e−2]	0.0374	2.3	15
NODE-CL, equal memory	4.8e−3 [1.7e−3,8.5e−3]	**0.80**	1.7e−2 [1.4e−2,2.1e−2]	0.0279	0.98	5.1

in both studies, with the gradient held constant between updates. The forward and backward solves of the adjoint use second-order integration on the 10 ms buffer grid.

Baselines (S1 and S2). The single-step baseline uses $\Xi = T\,\Phi'(X(t),\hat\theta)^\top E(t;t)$ so that its gradient magnitude matches that of the windowed laws. CL stores up to 100 pairs $(X_j, \hat f(\tau_j))$; a new pair is admitted when the relative change of the Jacobian $\Phi'(X,\hat\theta)$ from the last stored point exceeds 0.1, and once the stack is full it replaces the stored point whose replacement most increases the minimum singular value of the stacked regressor, evaluated every 50 ms in S1 and 200 ms in S2 with the current $\hat\theta$; its gradient is $\Xi = (T/N)\sum_{j=1}^{N}\Phi'(X_j,\hat\theta)^\top(\hat f(\tau_j) - \Phi(X_j,\hat\theta))$ with a constant gain Γ_0, and the stack-selection check, which requires a singular value decomposition per stored point, is not included in its reported update cost. Online Adam uses $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rates 10^{-2} (S1) and 3×10^{-3} (S2), selected from $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$ on seed 0; the single-step gain $\alpha = 2$ was selected on the pendulum tuning seed.

Evaluation (S1 and S2). The evaluation states are 400 samples, equally spaced in time, from the trajectory after $t = 20$ s (120 s runs) or 15 s (60 s runs). The excited subspace in Figure 8(b) is a diagnostic computed from the teacher Jacobian: the span of eigenvectors of the average of $\Phi'(X,\theta_{\mathrm{tr}})^\top\Phi'(X,\theta_{\mathrm{tr}})$ over the evaluation states with eigenvalues above 10^{-3} of the largest; its invariance under the adaptation is not established, and the distinction drawn is between excited and weakly excited directions rather than a strict rank deficiency.

Benchmark suite details. Domains and tasks are pendulum (swingup), cartpole (swingup), acrobot (swingup), and reacher (easy) from Tassa et al. (2018), integrated by MuJoCo at 1 ms with the suite's models unchanged; all are contact free. State dimensions $2n$ are 2, 4, 4, and 4, with 1, 1, 1, and 2 actuators; the network input has dimension 4, 6, 7, and 7 after the $(\cos, \sin)$ features (only unlimited

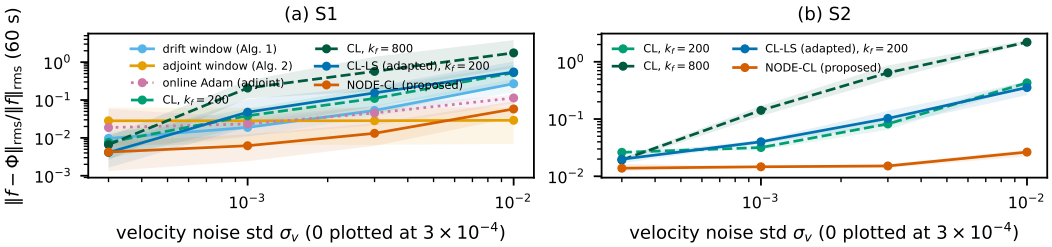

Figure 4: Terminal normalized functional error after 60 s against the standard deviation of white velocity measurement noise (three seeds, geometric mean with seed range) for (a) S1 and (b) S2.

Table 3: Benchmark suite, full results (five evaluation seeds, medians, 60 s): terminal normalized functional error on visited states, held-out 1 s error with seed range, held-out 4 s error, for clean (c) and noisy (n, $\sigma_v = 3 \times 10^{-3}$) measurements; mean cost per update and amortized stack-selection cost per simulated second (ms, one CPU core, NumPy). The 10^{-3} and 10^{-2} levels are shown in Figure 2; the 10^{-2} level was not run for the acrobot and reacher.

Domain	Method	func (c)	1 s (c)	4 s (c)	func (n)	1 s (n)	4 s (n)	upd ms	sel ms/s
pendulum	single-step	$1.0e+00$	1.643 [1.00, 2.05]	1.27	$8.7e-01$	1.650	1.24	0.03	0
pendulum	CL	$2.5e-02$	0.315 [0.17, 0.73]	0.29	$1.3e-01$	0.609	0.70	0.09	44
pendulum	CL-LS	$2.3e-02$	0.195 [0.03, 0.25]	0.11	$2.0e-01$	1.642	0.90	0.49	40
pendulum	NODE-replay	$5.0e-03$	0.183 [0.06, 0.26]	0.12	$1.5e-02$	0.141	0.11	4.77	188
pendulum	NODE-CL	$2.7e-03$	0.035 [0.01, 0.04]	0.02	$1.2e-02$	0.038	0.03	5.32	186
cartpole	single-step	$4.6e-01$	3.287 [1.65, 4.44]	7.50	$4.5e-01$	3.145	8.28	0.04	0
cartpole	CL	$7.0e-02$	0.623 [0.42, 0.95]	2.71	$1.2e-01$	1.363	7.96	0.16	196
cartpole	CL-LS	$2.2e-02$	0.251 [0.21, 0.95]	1.11	$1.1e-01$	1.984	28.61	1.91	186
cartpole	NODE-replay	$1.6e-02$	0.311 [0.25, 0.49]	1.81	$2.8e-02$	0.741	2.57	12.29	495
cartpole	NODE-CL	$2.1e-02$	0.381 [0.29, 0.57]	1.95	$1.7e-02$	0.302	2.09	13.54	489
acrobot	single-step	$5.5e+00$	1.621 [0.56, 9.48]	4.79	$3.9e+00$	1.232	2.72	0.04	0
acrobot	CL	$6.4e-01$	3.615 [0.92, 5.63]	16.24	$5.4e-01$	4.385	19.50	0.16	224
acrobot	CL-LS	$5.9e-02$	1.409 [1.21, 4.33]	7.04	$2.4e-01$	2.015	7.64	2.34	225
acrobot	NODE-replay	$3.3e-01$	1.002 [0.75, 2.55]	3.79	$3.3e-01$	2.129	6.15	14.45	598
acrobot	NODE-CL	$1.5e-01$	1.545 [0.31, 1.79]	5.09	$2.0e-01$	1.279	3.30	15.95	596
reacher	single-step	$1.3e+00$	1.509 [0.89, 4.02]	1.78	$1.3e+00$	1.516	1.78	0.04	0
reacher	CL	$6.5e-01$	1.007 [0.83, 1.63]	0.90	$6.4e-01$	0.976	0.98	0.16	236
reacher	CL-LS	$1.6e-01$	0.064 [0.05, 0.12]	0.24	$3.6e-01$	0.215	0.60	2.36	230
reacher	NODE-replay	$2.4e-01$	0.077 [0.03, 0.12]	0.11	$2.3e-01$	0.088	0.18	14.43	604
reacher	NODE-CL	$2.4e-01$	0.069 [0.04, 0.08]	0.12	$2.4e-01$	0.075	0.12	15.94	589

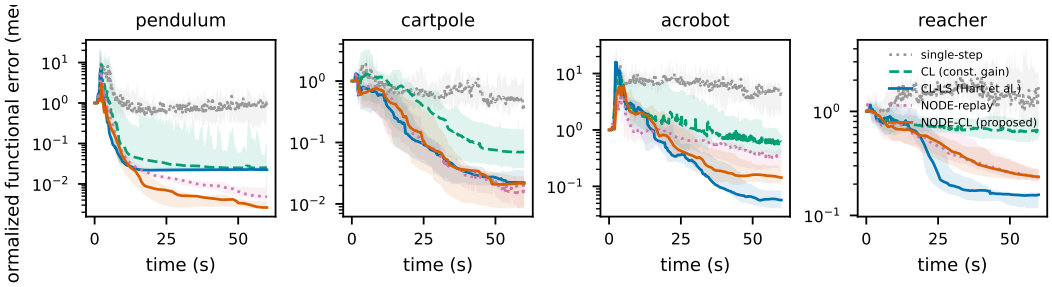

Figure 5: Benchmark suite learning curves: normalized functional error on the visited states against streaming time (five evaluation seeds, median with seed range, clean measurements).

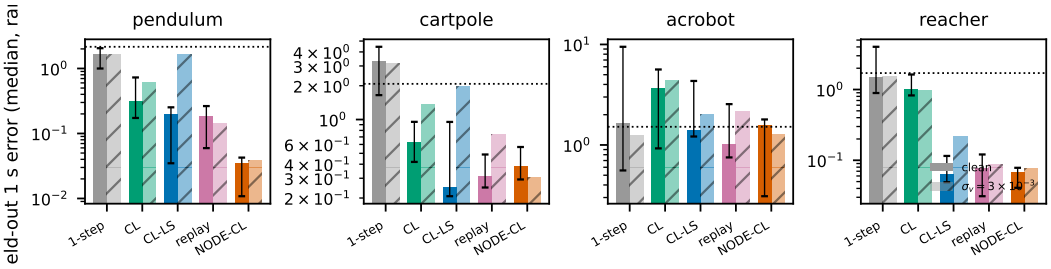

Figure 6: Held-out 1 s prediction error per domain (five evaluation seeds, median with seed range), clean (solid) and under velocity noise 3×10^{-3} (hatched); the dotted line is the initial model. Learning curves are given in Figure 5.

hinges receive the feature), the hidden layer has 12, 16, 16, and 16 $\tanh$ neurons, and $p = 73$, 146, 162, and 162. The excitation is a five-component multisine per actuator with frequencies uniform in $[0.1, 1.5]$ Hz, random phases, and amplitudes of 0.45, 0.5, 0.25, and 0.6 of the actuator range, plus a velocity (and, for the cartpole slider, position) feedback baseline that keeps the trajectory bounded; the held-out record uses a separate excitation seed and 12 s, from which eight initial conditions spaced 1.2 s apart are used. The network is initialized from $\mathcal{N}(0, 0.0025)$ with $\bar{\theta} = 60$, $\epsilon_\theta = 3$.

Benchmark suite gains. Observer gains are $\alpha_o = k_f = 200$ with the predictive hold. Single-step uses $\alpha = 0.1$ at a 1 ms update (selected on the pendulum from $\{0.5, 0.1, 0.02\}$; larger gains oscillate); CL uses $\alpha = 2$, $\Gamma = 5I_p$, a stack of 100 pairs admitted at 0.2 s checks with relative-Jacobian tolerance 0.1, at a 5 ms update; CL-LS adds Equation (10) in information form with $\beta_0 = 0.3$, $\bar{\lambda} = 10^3$, $\underline{\lambda} = 10^{-3}$, $k_\sigma = 10^{-6}$; NODE-CL uses segments of 0.05 s (200 on the pendulum, 100 elsewhere), admission checks every 0.1 s, $\alpha = 2$, $\beta_0 = 0.3$, $\bar{\lambda} = 10^6$, $\underline{\lambda} = 10^{-5}$, $k_\sigma = 10^{-9}$, $\Gamma_0 = 10^4 I_p$, at a 20 ms update; NODE-replay uses the same stack and update interval with learning rate 3×10^{-2} (selected on S2 from $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}\}$). The forgetting rate $\beta_0 = 0.3$ for both least-squares laws was selected on cartpole seed 0 from $\{0.1, 0.3, 1\}$; with the scalar forgetting of Slotine & Li (1989) the NODE-CL error on that seed was 0.109 against 0.037 with the direction-wise forgetting of Equation (10).

Benchmark suite memory and the swimmer. CL and CL-LS store 100 input-output pairs ($100(2n + m + n)$ scalars) plus a $p \times p$ gain for CL-LS; NODE-CL stores N_s segments of six samples ($6N_s(2n + m)$ scalars), a cache of N_s terminal sensitivities ($2nN_sp$ scalars), and the $p \times p$ gain. The swimmer (six links, $2n = 16$, $m = 5$, $p = 356$ with 12 hidden neurons, translation coordinates excluded from the input) was run with 30 segments, 0.1 s updates, and 0.5 s admission checks; after 60 s every method, including NODE-CL, has normalized functional error above 1 and held-out error above the initial model, so the domain is not discriminating at this data budget and is reported only here.

Relation to the concurrent-learning law of Hart et al. (2025a). CL-LS is not claimed as a new algorithm; it is an adapted baseline. The prior law updates all DNN layers by $\dot{\hat{\theta}} = \mathrm{proj}(\Gamma(\Phi'(X, \hat{\theta})^\top r - \gamma_1 \sum_{i=1}^N \Phi'(X_i, \hat{\theta})^\top (u_i - \hat{u}_i) - \gamma_2 \hat{\theta}))$ with $\frac{d}{dt}\Gamma^{-1} = -\beta\Gamma^{-1} + \gamma_1 \sum_{i=1}^N \Phi'(X_i, \hat{\theta})^\top \Phi'(X_i, \hat{\theta})$ frozen outside an eigenvalue band, where $u_i - \hat{u}_i$ is the stored residual between the applied input and its reconstruction from a state-derivative observer, and the stack is admitted after the observer settling time; its Lyapunov derivative bounds the gain-dynamics term by half the gradient term, as in the proof of Proposition 2. CL-LS in this paper restricts that mechanism to identification (no tracking term), with the windowed or stored residual $\hat{f}(\tau_j) - \Phi(X_j, \hat{\theta})$ normalized by T/N, the direction-wise forgetting and Lipschitz gate of Equation (10) in place of the frozen band, which yields the certified bounds of Lemma 4, and the projection in the Γ metric with the inequality of Lemma 5; the constant-gain CL row is the same DNN stack with Γ fixed, i.e., the point-stack rule of Chowdhary & Johnson (2011) applied to the DNN Jacobian, which is a simplification of the law of Hart et al. (2025a) rather than a separate prior method. Neither law is new in its residual; the contributions of this paper on the drift side are the certified gain and pro-

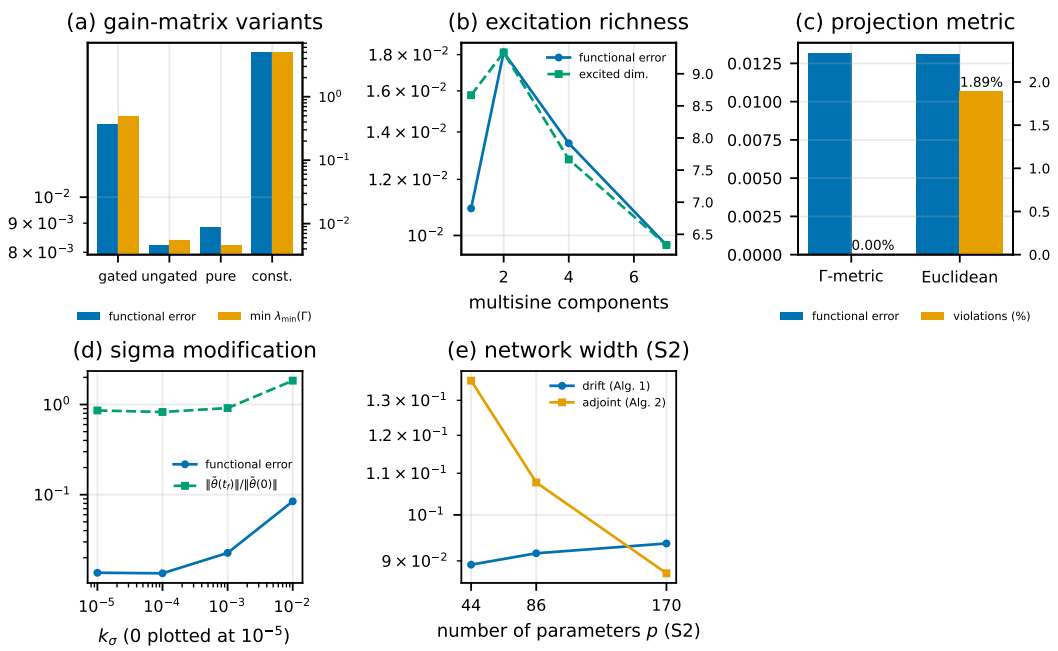

Figure 7: Additional ablations (three seeds, 60 s; geometric means; violation rates and excited dimensions are arithmetic means, and zero violations are drawn as zero). (a) Gain-matrix variants at $\alpha = 2$: terminal functional error and the smallest eigenvalue of Γ attained during the run. (b) Excitation richness: terminal functional error and the dimension of the excited subspace against the number of multisine components at equal total power. (c) Projection metric on a tight parameter set ($\bar{\theta} = \|\theta_{\mathrm{tr}}\|$, $\epsilon_\theta = 0.3$): terminal functional error and the fraction of updates that violate Equation (11). (d) Sigma-modification gain: functional and parameter error. (e) Network width in S2 for the drift and adjoint laws.

jection properties and the excitation-free guarantee, and on the trajectory side the decomposition of Theorem 1 and the law of Theorem 2.

Stored-set methods (S1 and S2). The following describes the diagnostic studies; the benchmark-suite settings are given above. CL-LS uses the CL stack (the same admission rule, tolerance, and check interval) with the scalar bounded-gain forgetting variant of the least-squares update, $\frac{d}{dt}\Gamma^{-1} = \varsigma_\Gamma \Psi_{\mathcal{D}} - \beta\Gamma^{-1}$ with $\beta = \beta_0(1 - \lambda_{\max}(\Gamma)/\bar{\lambda})$ (Slotine & Li, 1989), implemented by forward Euler at the update interval, with $\beta_0 = 1$, $\bar{\lambda} = 10^3$, $\underline{\lambda} = 10^{-3}$, $k_\sigma = 10^{-6}$, and $\Gamma_0 = 5I_p$. The stored-segment law uses $N_s = 200$ segments of $T_s = 0.05$ s (six samples on the 10 ms buffer grid), or 100 segments of 0.03 s (four samples) in the equal-memory variant, rolled out in a batch with second-order integration of $\hat{X}_{p,j}$ and S_j and trapezoidal quadrature of Equation (24); candidates are the most recent T_s of the buffer, tested every 0.1 s (0.2 s for the equal-memory variant) by the singular-value maximizing rule on the stacked terminal sensitivities, evaluated on the Gram matrix with rank-$2n$ updates; the gain uses $\alpha = 2$, $\beta_0 = 1$, $\bar{\lambda} = 10^6$, $\underline{\lambda} = 10^{-5}$, $k_\sigma = 10^{-9}$, $\Gamma_0 = 10^4 I_p$, and the update interval is 20 ms. The large $\bar{\lambda}$ and small k_σ reflect the T_s^3 scale of $G_{\mathcal{S}}$, on which $\Gamma \approx \beta G_{\mathcal{S}}^{-1}$ is correspondingly large; the leakage rate $k_\sigma \lambda_{\max}(\Gamma)$ is at most $10^{-3}\,\mathrm{s}^{-1}$ in all runs. The forward-sensitivity gradient agrees with the adjoint gradient of Theorem 8 to a relative error of 1.7×10^{-3} on a 0.25 s segment, which is the discretization level. The observer gain $k_f = 800$ runs use $\alpha_o = k_f = 800$ with the predictive hold. Velocity measurement noise is white Gaussian noise added to each 1 ms velocity sample used by the algorithms; the plant itself is noise free. The excitation-richness ablation keeps the first K multisine components at equal total power, and the width ablation uses 6, 12, and 24 hidden neurons in S2.

Ablations of the windowed laws. Figure 11 reports the ablations that bear on specific claims of the analysis. (a) The drift law improves slowly with longer windows, consistent with the linear

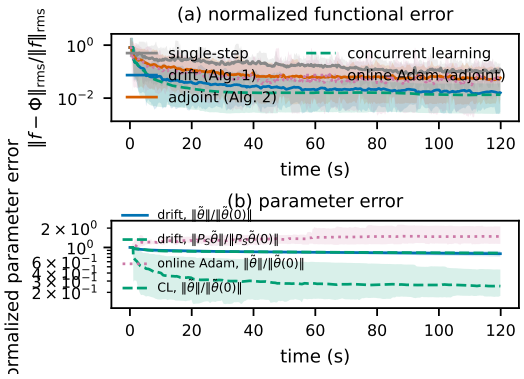

Figure 8: S1, five seeds, windowed laws and baselines. (a) Normalized functional error. (b) Parameter error: drift law, total and projected onto the dominant excited subspace S; totals for CL and online Adam. The projected component contracts while the complementary component changes little; the total error of the unconstrained Adam flow grows.

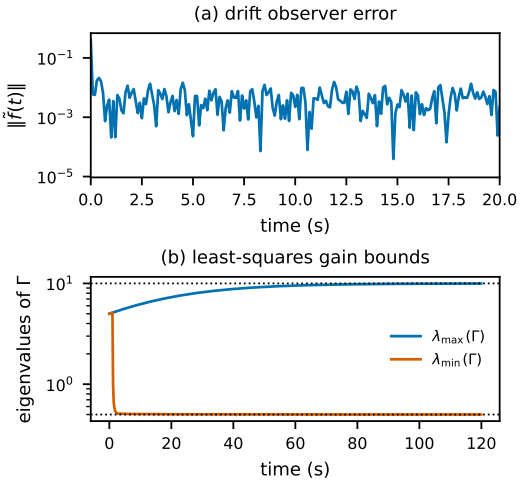

Figure 9: S1, representative seed, sampled predictive-hold observer. (a) Drift observer error $\left\|\tilde{f}(t)\right\|$. (b) Extreme eigenvalues of $\Gamma(t)$, which remain within the certified interval $[\underline{\lambda}, \bar{\lambda}]$ of Lemma 4 in every seed of both studies.

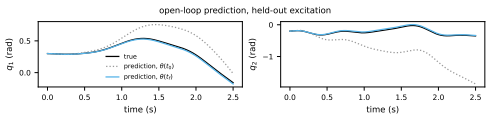

Figure 10: S2, 2.5 s open-loop prediction on the held-out excitation under the initial and final drift-law parameters of seed 0.

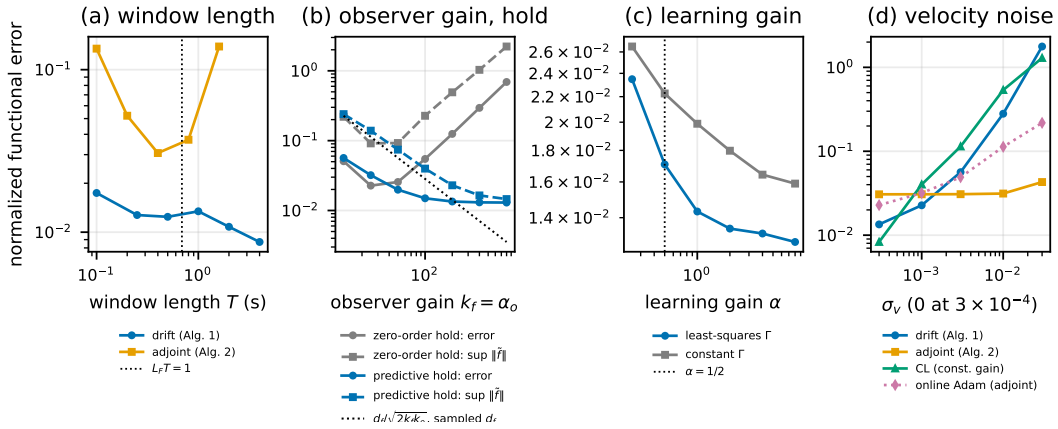

Figure 11: Ablations of the windowed laws (S1; three seeds, 60 s; geometric means). (a) Window length, with $L_F T = 1$ at the measured L_F. (b) Observer gain under a zero-order and a predictive hold: terminal error (solid), post-transient supremum of $\left\|\tilde{f}\right\|$ (dashed), and the residual formula $d_f/\sqrt{2k_f k_o}$ evaluated with the sampled d_f (dotted). (c) Learning gain with least-squares and constant gain matrices. (d) Velocity noise for the windowed laws, CL, and online Adam.

growth of its perturbation constants in T, whereas the adjoint law improves up to $T = 0.4$ s and then increases from 3.7×10^{-2} to 1.4×10^{-1} between 0.8 s and 1.6 s; the measured $L_F = 1.45$ places $L_F T = 1$ at $T = 0.69$ s, where the exponential growth of the constants in Equation (21) predicts the degradation to begin, and the same scaling is why the stored-segment law uses segments of 0.05 s with a Gauss–Newton gain rather than long windows with a constant one. (b) Under a zero-order hold the observer gain traces a U-shaped tradeoff, since the residual formula $\bar{z}_o$ decreases as $1/k_f$ while the hold error, amplified by k_f, grows as $k_f \Delta t$; the predictive hold removes the first-order hold error, and the observed floor follows the $1/k_f$ scaling up to $k_f \approx 100$ before saturating at the second-order residual that the continuous-time bound does not model. (c) The functional error decreases monotonically in α with no instability below $\alpha = \frac{1}{2}$, so the condition is sufficient rather than necessary, and the least-squares gain improves on a constant gain at every α. Removing the gate lowers the error by a further 35 percent over 60 s but lets $\lambda_{\min}(\Gamma)$ collapse to 5×10^{-3}, so the certified rate is lost; a Euclidean projection composed with the matrix gain violates Equation (11) in 1.9 percent of updates on a tight parameter set with no change in functional error here; and leakage above $k_\sigma = 10^{-3}$ degrades both errors, the tradeoff that $k_\sigma \bar{\theta}^2$ in Equation (14) expresses (Appendix K).

Additional ablations. Figure 7(a) compares the gain-matrix variants. The ungated and pure least-squares updates attain terminal errors of 8.2×10^{-3} and 8.9×10^{-3} against 1.3×10^{-2} for the gated update and 1.8×10^{-2} for a constant gain, but their smallest eigenvalue falls to 5×10^{-3} within 60 s and continues to decrease, so the certified rate in Proposition 2 degrades without bound; the gate trades a modest loss in short-horizon accuracy for a uniform lower bound. Figure 7(b) shows that the terminal error and the excited dimension vary weakly and non-monotonically with the number of multisine components at equal total power (between 9.7×10^{-3} and 1.8×10^{-2}, and between 6 and 9 dimensions), because fewer components at equal power produce larger excursions, which vary the Jacobian at least as much as additional frequencies do; the excited dimension is governed by the spatial extent of the trajectory rather than by the number of frequencies. Figure 7(c) reports the projection comparison discussed in Section 4.

Verification of the certified inequalities on simulated data. The following checks were performed on the code and the simulated signals. (i) The projection inequality in Equation (11) and the invariance of Θ_ϵ were verified on 20,000 random draws of $(\hat{\theta}, \theta^*, y, \Gamma)$, including 1,982 draws in the active region and 5,000 draws on the boundary, with no violations; on the simulated runs, the Γ-metric projection produced no violation in any update of any run. (ii) The gate ς implemented in code equals 0 at $\lambda_{\min}(\Gamma) = \underline{\lambda}$, 1 at $2\underline{\lambda}$, and is linear in between. (iii) The analytic Jacobians Φ' and $\partial\Phi/\partial X$ agree

with central finite differences to 10^{-9}, and the adjoint gradient agrees with central finite differences of the discretized Equation (16) with relative error 3×10^{-4} (S1) and 5×10^{-4} (S2) and cosine similarity 1.000000. (iv) In every seed of both studies, $\lambda_{\min}(\Gamma(t)) \geq \underline{\lambda}$ and $\lambda_{\max}(\Gamma(t)) \leq \bar{\lambda}$ hold at every logged instant. (v) On the logged S1 seed-0 trajectory after $t = 5\,\mathrm{s}$, the maximum sampled value of $\left\|\dot{f}\right\|$ is 2.83. Substitution into the continuous-time residual formula gives $\bar{z}_o \approx 1.4 \times 10^{-2}$ at $k_f = 200$, while the sampled predictive-hold observer has a post-transient supremum of $\left\|\tilde{f}\right\|$ of 2.3×10^{-2}. The sampled maximum is a diagnostic estimate rather than a certified bound over the full trajectory, and the sampled observer includes a hold residual not covered by the continuous-time result, as Figure 11(b) shows across gains. (vi) At $t = 10\,\mathrm{s}$ in S1, the identity $\Xi_f = \Psi\tilde{\theta} + \eta_f$ was evaluated with $\tilde{\theta}$ known, and $\|\eta_f\| = 0.42$ lies within the bound of Equation (9) evaluated with the measured $\bar{v}$, $c_{\theta\theta}$, and $\bar{f}_T$, which equals 5.2; the factor of twelve is the expected looseness of a Lagrange remainder bound. (vii) Two runs with the same seed produce bitwise identical parameter trajectories, and the plant trajectory is identical across the five methods for a given seed.