TOWARDS STABILISING HYBRID FILTERING VIA BOUNDED RESIDUAL CORRECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Filtering of dynamical systems remains a central concern in science and engineering. Classical model-based algorithms such as the Extended Kalman Filter offer interpretability and stability guarantees under regularity conditions, but can perform poorly under model misspecification. To compensate, hybrid methods combine model-based approaches with neural corrections, preserving some interpretability inherent in classical filters. However, naïve embedding of neural mappings risks driving hybrid filters to instability, a severe vulnerability in safety-critical applications. Aiming to resolve this instability without severely impairing expressivity or interpretability, we propose Bounded Residual Correction (BRC), an operationally filter- and architecture-agnostic framework that augments model-based representations with norm-bounded learned residual corrections. We then prove that to preserve Reif et al. (1999)'s probability-one error boundedness when correcting local linearisation-based estimates with learned residuals, bounding them is *necessary* and constraining their norms below a threshold is *sufficient* alongside nominal regularity conditions. Empirically, across extensive evaluations, self-supervised BRC-augmented filters outperform prior works in accuracy and probabilistic calibration, confirming that despite constraining residuals' expressivity, norm-bounding can enhance rather than hinder performance.

1 INTRODUCTION

In discrete-time dynamical systems, filtering algorithms estimate the distribution (or moments) of the system's latent state x_t^* causally, i.e., from its noisy measurements $y_{1:t}$ up to time t (Anderson & Moore, 1979). This requires an explicit *nominal* specification or an implicit specification of (1) a latent state process $x_t \in \mathbb{R}^s$ with transition dynamics $x_t \to x_{t+1}$, and (2) a measurement model relating observation $y_t \in \mathbb{R}^d$ to x_t. However, this specification may not faithfully represent physical reality; for instance, the true system might inhabit a state space of a different dimensionality $(\dim(x_t^*) \neq s)$, or the modelled mechanisms may differ systematically from the true ones.

Explicating a nominal specification ('nominal model') is the subject of System Identification (Ljung, 1998), achieved prior to filtering via expert knowledge, collected data or both. *Model-based* filters (e.g. the Extended Kalman Filter (Smith et al., 1962)) operate on such a model, without learnable neural parameterisation. In contrast, *data-driven* filters operate without prior system identification of the state transition dynamics. In this work, we restrict this category to neural networks (e.g. Recurrent Kalman Networks (Becker et al., 2019)) that implicitly model transition dynamics, while permitting known measurement models. Finally, *hybrid* filters (e.g. KalmanNet (Revach et al., 2022)) embed neural components into explicit nominal-model-based filtering recursions in closed loop.

In several real-world applications, e.g. autonomous transportation and robotics (Thrun et al., 2005), biomedical tracking (Sameni et al., 2007), and smart power grids (Dehghanpour et al., 2019), performant filtering is *desirable*, while safety (i.e. stability or non-divergence) is *essential*. Classical model-based filters like the Kalman Filter have stability, interpretability and optimal performance guarantees, but only under nominal model fidelity (Fitzgerald, 1971). Addressing misspecification through the covariance inflation heuristic (Anderson & Anderson, 1999) remains an area of active research (Piché et al., 2012; Duran-Martin et al., 2024) in model-based filtering. However, these methods cannot distinguish a true outlier from a mispredicted measurement (both produce large *innovations*), and can sub-optimally downweight the very measurements needed to correct the latter.

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.

Data-driven filters can outperform classical filters under misspecification (Ghosh et al., 2024; Norén et al., 2026), although their estimates completely lack the real-world semantics of knowledge-based nominal models. When explicit nominal models and data are sufficiently informative, hybrid filters address misspecification (see the survey by Shlezinger et al. (2025)) and approximation errors (e.g., via learned innovations in RBF-ESKF (Shaukat et al., 2021)) while still retaining some interpretability. However, to our knowledge, existing data-driven and hybrid filters lack formal stability guarantees. Unconstrained neural networks in recurrent estimation loops can be highly responsive even to transient measurement shocks and drive the filter to divergence; filter divergence has been observed empirically by Song et al. (2024).

These considerations motivate our research question: *what parsimonious intervention can preserve the stability of model-based filters, the performance potential of data-driven filters, and the interpretability of hybrid filters?* We explore whether bounding the norms of residual corrections to model-based state representations is such an intervention. Concretely, our contributions are:

1. **Bounded Residual Correction (BRC).** We present a distinct formulation for state estimation under misspecification: a gated addition of model-based state representations and bounded residual corrections in recurrent closed loop. Operationally, our modular framework is agnostic to both the filter producing the model-based representations and the mechanism, neural or otherwise, generating the residual corrections. However it requires that these residuals be norm-bounded and parameterised so that the augmented representations satisfy their model-based counterparts' structural constraints. For instance, augmented covariances must be positive definite.
2. **Provable Stability Guarantees.** Using stochastic Lyapunov theory, we prove that augmenting the canonical family of local linearisation-based filters (specifically the EKF and its covariance inflating extensions) with learned residual corrections via BRC preserves the stochastic stability of Reif et al. (1999), provided the norms of learned residuals remain below a threshold and nominal regularity conditions hold. This guarantees exponential mean-square and probability-one infinite-horizon, pathwise error bounds. To our knowledge, this is the **first** formal stability guarantee for hybrid filters.
3. **Structural Interpretability.** By applying learned corrections directly to the state representations, BRC explicitly quantifies the filter's per-time-step deviations from its nominal representations. This enables direct inspection of the neural network's exact influence, with residuals mapping directly to physical units when the nominal model is grounded in domain knowledge.
4. **Empirical Efficacy**. Across diverse settings involving model misspecification—dimensional mismatch in a linear Gaussian system, functional mismatch in a chaotic, nonlinear and non-Gaussian system, and data degradation in an uncharacterised real-world system—BRC achieves best-in-class accuracy and probabilistic calibration. Importantly, these results show that BRC's constrained expressivity can *improve* rather than *hinder* filtering performance.

2 A Flexible Hybrid Filtering Framework

We consider the special but canonical case of model-based filters such as the EKF that operate by propagating estimates of the first two conditional state moments (hereinafter 'Mean-Covariance Propagating' filters). Here, BRC can operate at the *a priori* (time-update) and/or the *a posteriori* (measurement-update) stages[1], respectively denoted by the time-step index $i = t-1$ and $i = t$. Specifically, BRC augments nominal, model-based (MB) state estimates $\mu_{\text{MB},\,t|i}$ and $P_{\text{MB},\,t|i}$ with learned residuals $\mu_{\text{BRC},\,t|i}$ and $P_{\text{BRC},\,t|i}$ of bounded norms as shown below:

$$\begin{aligned} \mu_{t|i} &= \mu_{\text{MB},\,t|i} + \mu_{\text{BRC},\,t|i}, \qquad P_{t|i} = P_{\text{MB},\,t|i} + P_{\text{BRC},\,t|i}, \\ \text{s.t.} \qquad & \|\mu_{\text{BRC},\,t|i}\|_2 \le c_{\max},\ P_{\text{BRC},\,t|i} \succeq 0 \text{ and } \|P_{\text{BRC},\,t|i}\|_F \le c_{\max} \end{aligned} \tag{1}$$

where $\|\cdot\|_2$ and $\|\cdot\|_F$ are respectively the Euclidean and Frobenius norms. $c_{\max}$ is a non-negative scalar hyperparameter, the norm-bound that is the cornerstone of BRC.[2] For our formal guarantee in §3, BRC does not require the learned residuals to have been neurally generated. Non-neural learning of the residual corrections via parametric approaches such as stochastic variational inference

[1] We do not intend for our use of the terms *a priori* and *a posteriori* to take on a Bayesian interpretation in general, although such an interpretation could be meaningful if the specified model were probabilistic.

[2] We rely on a single $c_{\max}$ for both the mean and covariance norms, across both stages for ease of exposition.

(Course & Nair (2023) explore this for non-causal state estimation) or nonparametric approaches such as Gaussian processes (Ko et al., 2007) may be viable. However, in differentiable filtering routines, neural networks can be learned conveniently via backpropagation. They are also the standard vehicles for addressing model misspecification in more recent works (see the survey Shlezinger et al. (2025)). In this paper, therefore, we proceed by considering the neural case alone.

We denote the BRC-augmented filter's *a priori* and *a posteriori* neural backbones by ψ_1 and ψ_2, respectively, with corresponding inputs $\mathcal{I}_{t|t-1}$ and $\mathcal{I}_{t|t}$. These inputs may include any causally available statistics (e.g. the model-based statistics themselves, previous innovations, or raw measurements). Algorithm 1 is an *example* of how the learned residuals $\mu_{\text{BRC},\,t|i}$ and $P_{\text{BRC},\,t|i}$ may be parameterised. In any case, however, the learned residuals *must* then augment state estimates per Eq. (1) and be piped through the nominal filter's recursive loop as detailed in Algorithm 2. BRC thus adds a constant computational overhead per time-step to the nominal filter's own routine. Intuitively, the BRC neural backbone's expressivity helps extract residual corrections (akin to missing empirical closure terms) that compensate for unmodelled system transition and/or observation mechanisms, continually mitigating error accumulation, while the nominal model acts as a structural anchor. Moreover, the learned residuals function as deterministic offsets to the nominal filter's state estimates; thus, whenever the nominal estimates are assumed to follow a location-scale probability distribution family, BRC-augmented estimates remain in the same family.

Algorithm 1 Example Procedure to Paremeterise BRC Neural Backbone Outputs

Require: Neural network (ψ, could be ψ_1 or ψ_2), corresponding input ($\mathcal{I}_{t|i}$), norm-bound ($c_{\max}$), activation function ($\sigma : \mathbb{R} \mapsto [0, 1]$), stabiliser ($\epsilon$, a small scalar).

$z \leftarrow \textsc{ForwardPass}(\psi,\ \mathcal{I}_{t|i})$ ▷ In this example, $z \in \mathbb{R}^{2+2s}$

$(a,\ b,\ z_\mu,\ z_P) \leftarrow \textsc{Partition}(z)$ ▷ $a,\ b \in \mathbb{R}$ and $z_\mu,\ z_P \in \mathbb{R}^s$

$(\tilde{a},\ \tilde{b}) \leftarrow c_{\max} \cdot \sigma([a, b])$

$\mu_{\text{BRC},\,t|i} \leftarrow \tilde{a} \cdot (z_\mu)/(\|z_\mu\|_2 + \epsilon)$

$P_{\text{BRC},\,t|i} \leftarrow \tilde{b} \cdot (z_P z_P^\top + \epsilon \cdot \mathbb{I})/(\|z_P z_P^\top + \epsilon \cdot \mathbb{I}\|_F)$ ▷ $\mathbb{I}$ is the identity matrix of dim. s

return $\mu_{\text{BRC},\,t|i}$, $P_{\text{BRC},\,t|i}$

Algorithm 2 BRC-Augmented Filtering Routine for Mean-Covariance Propagating Nominal Filters

Require: Data ($y_{1:T}$), model ($\mathcal{M}$), neural networks (ψ_1 and/or ψ_2).

$t \leftarrow 0$

while $t < T$ **do**

  $t \leftarrow t + 1$

  # 1. Nominal Time Update (see Appendix A.1.1 for examples)

  $(\mu_{\text{MB},\,t|t-1},\ P_{\text{MB},\,t|t-1}) \leftarrow \textsc{TimeUpdate}(\mu_{t-1|t-1}, P_{t-1|t-1}, \mathcal{M})$

  # 2. Optional Neural Correction (A priori); set $\psi_1 \equiv \mathbf{0}\ \forall\ t$ if no correction required

  $(\mu_{\text{BRC},\,t|t-1},\ P_{\text{BRC},\,t|t-1}) \leftarrow \psi_1(\mathcal{I}_{t|t-1})$

  $\mu_{t|t-1} \leftarrow \mu_{\text{MB},\,t|t-1} + \mu_{\text{BRC},\,t|t-1}$

  $P_{t|t-1} \leftarrow P_{\text{MB},\,t|t-1} + P_{\text{BRC},\,t|t-1}$

  # 3. Nominal Measurement Update (see Appendix A.1.1 for examples)

  $\mu_{\text{MB},\,t|t}, P_{\text{MB},\,t|t} \leftarrow \textsc{MeasUpdate}(\mu_{t|t-1}, P_{t|t-1}, y_t, \mathcal{M})$

  # 4. Optional Neural Correction (A posteriori); set $\psi_2 \equiv \mathbf{0}\ \forall\ t$ if no correction required

  $(\mu_{\text{BRC},\,t|t},\ P_{\text{BRC},\,t|t}) \leftarrow \psi_2(\mathcal{I}_{t|t})$

  $\mu_{t|t} \leftarrow \mu_{\text{MB},\,t|t} + \mu_{\text{BRC},\,t|t}$

  $P_{t|t} \leftarrow P_{\text{MB},\,t|t} + P_{\text{BRC},\,t|t}$

end while

return $(\mu_{t|t},\ P_{t|t})\ \forall\ t \in 1, \ldots, T$

BRC does not preclude supervised training of its neural backbones. However, since it may be prohibitively expensive to obtain sufficient labelled data in real-world applications, in this paper, we focus on self-supervision. Suitable self-supervision objective functions may include assumed-density log-likelihoods of measurements (see Minka (2001) for a discussion on Assumed-Density Filtering) or variational evidence lower bounds. We provide concrete examples of both in §4. However, hybrid and data-driven filters implicitly model the dynamical system nonlinearly because of

their reliance on neural networks that are inherently nonlinear. The fundamental parameter non-identifiability inherent in nonlinear systems (Ljung & Glad, 1994) may therefore surface during model (hyperparameter) selection; i.e. training may yield an equivalence class of neural network weights that optimise self-supervision objective functions but imply physically degenerate latent state processes, making model selection based on measurement-space metrics alone ill-posed. We present empirical evidence of this issue in Appendix B. To address this issue, we propose a model selection (validation) procedure applicable to data-driven and hybrid filters in Appendix C that makes use of a small set of ground truth trajectories when available. The training procedure can remain fully self-supervised on noisy measurements, with no pathway for validation labels to influence parameter updates. Such a setup, which we implement for §4, would mirror established protocols in self-supervised representation learning (Alain & Bengio, 2017; Chen et al., 2020; He et al., 2020).

This concludes our exposition of how BRC can augment the family of Mean-Covariance Propagating filters. Algorithm 2 thus covers a large swathe of filters, including covariance-inflating extensions of the EKF, such as AREKF (Xiong et al., 2009), which inflates *a priori* covariance estimates, cubature filters such as the Cubature Kalman Filter (Arasaratnam & Haykin, 2009) etc. In further support of our claim that BRC is operationally filter- and architecture-agnostic, we additionally present in Appendix A algorithms showing how BRC can operate on two other families of filters—Mixture-of-Density Filters (e.g. the Gaussian Sum Filter (Alspach & Sorenson, 1972)) and Sequential Monte Carlo-based filters (Naesseth et al., 2019)—using three types of neural networks: a stateless MLP, a stateful RNN and a history-dependent Causal Transformer.

3 Stability Guarantee

We consider the local linearisation-based family of filters, specifically the EKF and its covariance inflating extensions. They require specification of a model of the following form:

$$x_t = f(x_{t-1}) + v_t, \qquad y_t = h(x_t) + w_t \tag{2}$$

where $x_t \in \mathbb{R}^s$ and $y_t \in \mathbb{R}^d$ denote the state and measurement at time $t \in \mathbb{Z}_{\geq 0}$, $f : \mathbb{R}^s \to \mathbb{R}^s$ and $h : \mathbb{R}^s \to \mathbb{R}^d$ are Markovian transition and measurement functions and v_t and w_t are noise terms with positive definite covariances Q_t and Σ_t, respectively. In order for BRC to preserve their native stability, BRC-augmented filters inherit EKF's own requirement that this model meets the regularity conditions of Reif et al. (1999); these conditions are listed in Appendix G. The BRC-augmented EKF hybrid filter's formal stability guarantee is stated in Theorem 1. We defer its proof and extensions to Appendices I and J, respectively. Note that the theorem's result is with respect to the specified model, Eq. (2), the only available frame of reference to the true dynamical system.

Theorem 1. *Consider a BRC-augmented EKF whose specified model satisfies the regularity conditions listed in Appendix G. Let $e_t := x_t - \mu_{t|t}$ denote the filter's estimation error, where $\mu_{t|t}$ is the filter's* a posteriori *estimate. Then, there exist strictly positive scalars $\bar{\epsilon}$, $\bar{\epsilon}_v$, $\bar{\sigma}$, and $\bar{c}$, and functions β, γ and δ such that if the following conditions hold: (1) the initial error satisfies $\|e_0\|_2 \leq \epsilon$, for some $0 < \epsilon < \bar{\epsilon}$, (2) the filter's norm-bound satisfies $c_{\max} < \bar{c}$, (3) the trajectory drift bound (see Definition 3) satisfies $\epsilon_v < \bar{\epsilon}_v$, and (4) the measurement noise upper eigenvalue bound (see Assumption 4) satisfies $\bar{s} \leq \bar{\sigma}$, then the error e_t remains bounded with probability one for all $t \geq 0$ and exponentially bounded in the mean-square, i.e.:*

$$\mathbb{E}[\|e_t\|_2^2] \leq \beta(\|e_0\|_2^2, t) + \gamma(c_{\max}) + \delta(c_{\max}) \quad \forall\, t \geq 0 \tag{3}$$

where $\beta(\|e_0\|_2^2, t) \to 0$ exponentially as $t \to \infty$.

4 Experiments

4.1 Purpose and Scope of Empirical Evaluation

The hybrid filtering paradigm—where neural networks compensate for diverse misspecifications across transition and measurement models—has already been established (see survey by Shlezinger et al. (2025)). Our empirical evaluations focus on whether BRC's characteristic norm-bounding trades off performance for stability. Our evaluation is restricted to datasets previously used in hybrid

state estimation literature (Garcia Satorras et al., 2019; Revach et al., 2022). We evaluate BRC performance with diverse neural architectures; in the results presented in §4.2–4.3, BRC-augmented filters jointly learn non-neural model parameters such as the initial state and/or noise covariances alongside their neural backbones. In §4.2–4.3, we therefore also **ablate** the neural backbone from the BRC-augmented filter to isolate the backbone's contribution to incremental performance, if any.

Baselines. We benchmark BRC performance against several prior works across the model-based, hybrid and data-driven filtering paradigms. In addition to the EKF, which reduces to a Kalman Filter (KF) under linear models, we evaluate WoLF (Duran-Martin et al., 2024) and RORF (Piché et al., 2012), two model-based filters that gain outlier robustness by inflating measurement noise covariance in response to large innovations; both were implemented with linearisation. KNet (Revach et al., 2022) and RKNet (Mortada et al., 2025) are 'gain-learning' hybrid filters; they replace the EKF's analytical Kalman gain step with unbounded learned gains. KNet is a deterministic model producing point estimates of the conditional state mean; RKNet additionally learns a conditional state covariance estimate. APBM (Straka et al., 2025) is a hybrid filter that augments the filter state with neural parameters that it learns jointly alongside the nominal state estimate via Cubature Kalman filtering (CKF) (Arasaratnam & Haykin, 2009). Since neural parameters are not directly observable, however, APBM risks violating covariance boundedness conditions necessary for CKF's stability guarantee (Xu et al., 2016). SIN (Krishnan et al., 2017) is a fully data-driven deep generative state space model (see §5 for a description of its mechanics). DANSE (Ghosh et al., 2024) is a data-driven filter that replaces analytical time updates with RNN recursions learned via measurements likelihood maximisation and uses a known observation matrix for linear measurement updates. Building on this, VSE (Norén et al., 2026) replaces analytical measurement updates as well, with variationally learned RNN recursions. Additionally, we adapt the deep structured state space model Mamba-3 (Lahoti et al., 2026) and the transformer-based time-series forecasting model PatchTST (Nie et al., 2023) with linear projection heads for filtering.

The main performance metrics we report are the Mean Square Error (MSE) between filtered means and ground truth trajectories and the Gaussian state negative log-likelihood (NLL), which respectively reflect filtering accuracy and probabilistic calibration. To ensure statistical consistency, we also report |ANEES Dev.| (see Appendix D for definition) whenever we have $N > 1$ ground truth trajectories (i.e. in §4.2–4.3). In tables, we report the median, 25th and 75th percentiles of these metrics across 10 random seeds for trained filters; the variance in metrics, in these cases, stems exclusively from learnable parameter initialisation. **Importantly**, we report the number of random seeds that result in filter instability (divergence to ∞) in parentheses. Tables report the best filter metrics in **bold** and the second best underlined. Appendix E contains details pertaining to training, datasets, model constructions and hyperparameters.

4.2 Linear Benchmark with Dimensional Misspecification

Setup. We consider the linear benchmark of Garcia Satorras et al. (2019), where the system to track is a particle undergoing damped acceleration in 2D, described by a 6-dimensional state $(p_x, \dot{p}_x, \ddot{p}_x, p_y, \dot{p}_y, \ddot{p}_y)^\top$. The nominal model suffers a dimensional mismatch, as it assumes a 4-dimensional constant-velocity approximation $(p_x, \dot{p}_x, p_y, \dot{p}_y)^\top$ and observes only noisy 2D positions. The nominal and true models are described in detail in Appendix E.2. To address the model mismatch, we augment the nominal 4D KF with a simple BRC backbone: a stateless MLP with two 32-unit hidden layers, applied exclusively at the *a priori* stage; we call the resulting filter 'ResEKF'. This MLP was learned by minimising assumed-Gaussian measurements NLL. Given the full 6D linear model, including the true initial state distribution, however, the KF ('Oracle KF') produces the theoretically optimal MSE, thus serving as the gold standard.

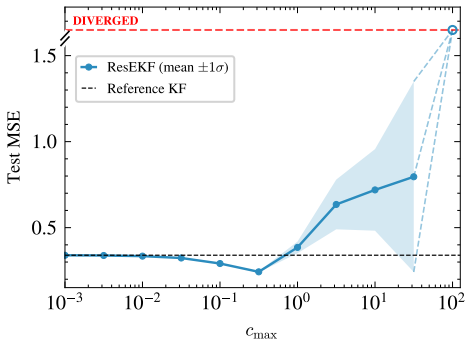

Figure 1: **ResEKF Test MSE vs.** $c_{\max}$. The curve validates Theorem 1's implication that unbounded residuals can destabilise ResEKF.

Results. Table 1 reports evaluation metrics on the linear benchmark. ResEKF substantially narrows the performance gap between the Biased and Oracle KFs as compared to other filters, while attaining best-in-class probabilistic calibration, maintaining statistical consistency and remaining stable, unlike competitors DANSE and APBM. This demonstrates that BRC's characteristic norm-bounding can preserve stability, while improving rather than hindering performance. Ablating ResEKF's MLP shows that learning the initial state and noise covariances alone accounts for the initial improvement over the Biased KF (MSE $0.35 \rightarrow 0.29$), while the further reduction in error ($0.29 \rightarrow 0.21$) isolates the contribution of the backbone. The high-capacity data-driven sequence models (Mamba-3 and PatchTST) achieve moderate to high MSE, but with extreme NLL and ANEES Dev. ($> 10^3$), demonstrating their inability to produce calibrated uncertainty without principled probabilistic update rules.

Table 1: **Results on the linear benchmark.** ResEKF successfully compensates for unmodelled dynamics, nearly recovering oracle performance, while other baselines underperform or diverge.

Filter	MSE ↓	NLL ↓	|ANEES Dev.| ↓
Oracle KF	0.18	1.67	0.00
Biased KF	0.35	3.85	1.87
PatchTST	$34.74_{(33.87,\ 36.10)}$	$> 10^3$	$> 10^3$
Mamba-3	$1.48_{(1.47,\ 1.50)}$	$> 10^3$	$> 10^3$
VSE	$0.95_{(0.95,\ 0.97)}$	$5.15_{(5.01,\ 5.19)}$	$0.47_{(0.37,\ 0.56)}$
SIN	$0.92_{(0.92,\ 0.93)}$	$4.86_{(4.82,\ 4.89)}$	$\underline{0.36}_{(0.18,\ 0.81)}$
RKNet	$0.74_{(0.44,\ 0.88)}$	$4.91_{(4.25,\ 4.96)}$	$2.05_{(1.04,\ 2.56)}$
KNet	$0.31_{(0.31,\ 0.32)}$	n.a.	n.a.
ResEKF (Ours)	$\mathbf{0.21}_{(0.21,\ 0.22)}$	$\mathbf{2.26}_{(2.21,\ 2.38)}$	$\mathbf{0.18}_{(0.09,\ 0.27)}$
Ablation			
ResEKF minus NN	$\underline{0.29}_{(0.29,\ 0.29)}$	$\underline{2.72}_{(2.71,\ 2.74)}$	$0.40_{(0.30,\ 0.56)}$
Diverged Models			
DANSE (2)	$0.90_{(0.90,\ 0.90)}$	$5.03_{(5.02,\ 5.05)}$	$2.42_{(2.31,\ 2.73)}$
APBM (3)	$0.29_{(0.29,\ 0.29)}$	$2.73_{(2.73,\ 2.74)}$	$0.32_{(0.23,\ 0.47)}$

Necessity for Norm-Bounding. Figure 1 provides empirical evidence of Theorem 1's implication: unbounded residual corrections can destabilise ResEKF. The figure shows test fold MSE across a grid of $c_{\max}$ values. ResEKF's initial state and noise covariances were fixed to isolate the impact of $c_{\max}$. Plotted as a horizontal line is ResEKF's nominal KF. In the low-$c_{\max}$ regime, ResEKF has little room to deviate from the nominal KF; it therefore operates well within its stability margin, but lacks the expressivity required to correct for dimensional mismatch. At the optimal $c_{\max}$, the BRC backbone is expressive enough to capture the unspecified physics while maintaining bounded stability. However, as $c_{\max}$ increases beyond this threshold, neural corrections consume an ever-greater portion of the available stability margin, while also raising the theoretical error upper-bound (Eq. (3)) via the terms $\gamma(c_{\max}) + \delta(c_{\max})$. Intuitively, the nominal dynamics amplify neural feedback faster than subsequent filter updates can sufficiently correct. At $c_{\max} = 100$, the margin is evidently entirely exhausted, pushing the filter outside its region of attraction ($c_{\max} > \bar{c}$) until it diverges. Practically, this calls for $c_{\max}$ to be tuned as a hyperparameter.

4.3 Lorenz Attractor with Functional Misspecification

Setup. We evaluate BRC on the well-known Lorenz attractor, a nonlinear chaotic dynamical system. The true system evolves in continuous time, while the nominal model is a discrete-time second order Taylor approximation with sampling period $\Delta t = 0.02$. This is a transition function mismatch studied in (Revach et al., 2022). Measurements are corrupted by a heavy-tailed Gaussian mixture:

$$y_t = x_t + w_t, \quad w_t = z_t\,\epsilon_1 + (1 - z_t)\,\epsilon_2, \quad z_t \sim \mathrm{Ber}(0.95), \quad \epsilon_1 \sim \mathcal{N}(0, I), \quad \epsilon_2 \sim \mathcal{N}(0, r^2 I)$$

with outlier scale $r = 16$ during training. We therefore include the model-based baselines RORF and WoLF that are specifically designed for outlier robustness in our evaluation. The BRC-augmented

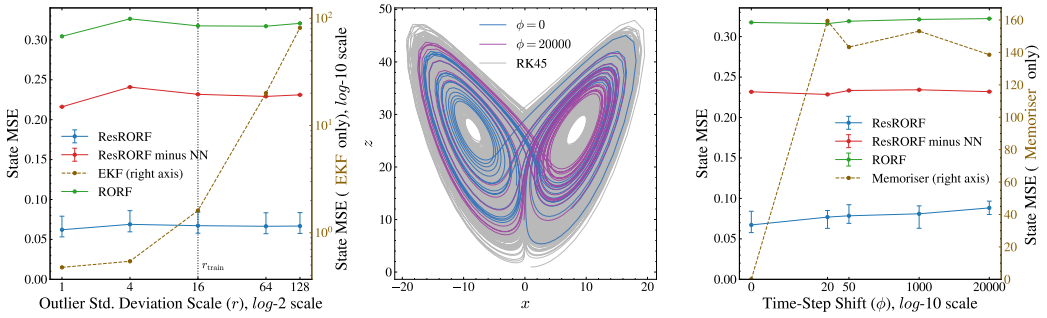

Figure 2: **Zero-shot OOD evaluation on the Lorenz attractor**. *Left*: Test MSE across unseen outlier measurement scales (r). *Centre*: Visualisation of ground truth trajectories shifted by ϕ time-steps. *Right*: Test MSE across ground truth trajectory shifts (ϕ) at $r = 16$. The catastrophic error of the 'Memoriser' baseline (plotted on the right axis) establishes that ResRORF learns the underlying dynamics rather than overfitting to the training sequence.

filter in this study is ResRORF, which uses RORF as its nominal anchor, and applies residual corrections with an MLP backbone at the *a priori* stage and a Causal Transformer backbone at the *a posteriori* stage. The BRC backbones were learned by minimising the variational evidence lower bound (ELBO) that RORF natively computes. All model-based and hybrid filters were supplied with a fixed initial state. More details are specified in Appendix E.3.

In-Distribution ($r = 16$) Test Results. Following Revach et al. (2022), we performed an in-distribution evaluation, where all filters were tasked with denoising held-out test measurements drawn from the same distribution as the training fold. In particular, the test ground truth trajectory was identical to its training counterpart, except that no filter had access to it—they only had access to noisy measurements. Yet, while DANSE achieves moderate performance, other data-driven filters (SIN and VSE) fail to outperform the EKF baseline (see Table 2 for results in this in-distribution evaluation). Conversely, WoLF and RORF demonstrated significant performance improvements, validating their native outlier robustness. Whereas existing hybrid filters (APBM, KNet and RKNet) are restricted to specific nominal anchors (CKF or EKF), BRC could be successfully employed to augment the outlier-robust RORF, allowing ResRORF to achieve best-in-class accuracy, probabilistic calibration and consistency. Ablating ResRORF's neural backbones confirms this: learning just noise covariances provided a baseline improvement over the nominal RORF, but the full neural architecture achieved further error reduction. Finally, although the unconstrained expressivity of KNet and RKNet's neural measurement update steps help them overcome their nominal EKF's outlier susceptibility on surviving seeds (9 and 8/10, respectively), the same can drive the filter to catastrophic divergence noted as in other seeds. Equipping KNet and RKNet with transformer backbones revealed exacerbated instability (see Appendix F).

Zero-shot OOD Test Results. We perform two zero-shot out-of-distribution evaluations. First, we held the in-distribution ground truth trajectory unchanged but realised unseen measurements, with outlier scale $r \in \{1, 4, 64, 128\}$. ResRORF continued to outperform its nominal RORF, both with and without its neural backbones (see Figure 2, left panel) across all evaluated scales; in contrast, EKF's errors scaled with r. Second, we shifted forward the in-distribution ground truth trajectory and corresponding measurement inputs by $\phi \in \{20, 50, 1000, 20000\}$ time-steps, holding outlier scale unchanged at $r = 16$. The centre panel of Figure 2 shows substantial variation in the ground truth trajectories to be recovered; for scale, a hypothetical estimator ('Memoriser') that outputs $\phi = 0$ ground truth trajectories irrespective of measurements shows MSE over 100 units (see Figure 2, right panel). In this evaluation too, ResRORF outperformed the ablated ResRORF and RORF across ϕ values, showing that it does not trivially overfit to its training. These evaluations demonstrate that the BRC's neural backbones help close the transition function mismatch on this nonlinear chaotic system. Outlier-robust filters like RORF and WoLF downweight measurement updates when innovations are large, but they fail to distinguish between a biased mean prediction and a true outlier. With BRC's backbones correcting for unmodelled transition function bias, RORF's nominal measurement update is enabled to downweight true outliers.

Table 2: **In-distribution** ($r = 16$) **results on the Lorenz attractor**. ResRORF extends BRC to an outlier-robust nominal anchor, achieving best-in-class accuracy and calibration, while purely data-driven baselines underperform and anchor-restricted hybrid filters (KNet and RKNet) diverge.

Filter	MSE ↓	NLL ↓	| ANEES Dev.| ↓
EKF	1.61	17.16	32.91
WoLF	0.33	3.19	3.62
RORF	0.32	2.56	0.89
SIN	$2.57_{(2.44,\ 2.68)}$	$5.66_{(5.59,\ 5.73)}$	$0.64_{(0.55,\ 0.77)}$
VSE	$2.37_{(1.63,\ 2.83)}$	$6.07_{(5.94,\ 6.19)}$	$1.85_{(1.72,\ 2.24)}$
DANSE	$0.64_{(0.61,\ 0.70)}$	$3.99_{(3.95,\ 4.19)}$	$1.72_{(1.55,\ 1.77)}$
APBM	$1.21_{(1.14,\ 1.31)}$	$3.53_{(3.16,\ 3.65)}$	$1.14_{(1.13,\ 1.16)}$
ResRORF (Ours)	$\mathbf{0.07}_{(0.06,\ 0.08)}$	$\mathbf{-0.92}_{(-1.36,\ -0.22)}$	$\mathbf{0.41}_{(0.28,\ 0.73)}$
Ablation			
ResRORF minus NN	$0.23_{(0.23,\ 0.23)}$	$1.62_{(1.61,\ 1.63)}$	$\underline{0.58}_{(0.52,\ 0.61)}$
Diverged Models			
RKNet (2)	$0.25_{(0.24,\ 0.27)}$	$2.53_{(2.48,\ 2.68)}$	$1.13_{(0.50,\ 1.39)}$
KNet (1)	$\underline{0.20}_{(0.20,\ 0.20)}$	n.a.	n.a.

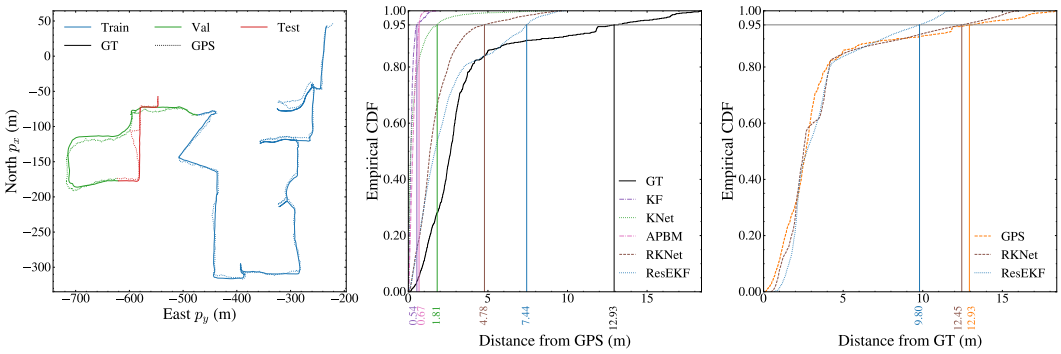

Figure 3: **NCLT dataset evaluation**. *Left*: Robot trajectory showing GPS readings corrupted by multipath noise vs. Ground Truth (GT). *Centre*: CDF of distance from GPS. KF, KNet and APBM nearly always collapse to the erroneous GPS readings, whereas ResEKF and RKNet deviate meaningfully. *Right*: CDF of state error (distance from GT). ResEKF yields the tightest error envelope.

4.4 Real-World Benchmark: the NCLT Dataset

Setup. The Michigan NCLT dataset (Carlevaris-Bianco et al., 2016), studied in prior hybrid state estimation work (Garcia Satorras et al., 2019; Revach et al., 2022), tracks a Segway robot navigating a university campus with several sensors. High quality sensor readings constitute the ground truth trajectory of the robot, while consumer-grade GPS readings, vulnerable to multipath noise from campus buildings, form the measurements. The nominal model dead reckons 2D positions using yaw (heading) and wheel speed sensor data as control inputs. While the robot's true dynamics are unknown, the nominal model is certainly misspecified: it discounts the nonlinear, geometric effect of road elevation, and suffers from control input biases. All filters were supplied with fixed initialisations and tasked with stateful filtering (terminal filter states are carried across training batches, validation, and test folds). The BRC-augmented filter, ResEKF, is otherwise identical to that in §4.2, except that its backbone is a specifically designed GRU (see Appendix E.4.2) and noise covariances are tuned as hyperparameters, not learned. Appendix E.4 presents more details.

Results. Table 3 presents metrics evaluated on the NCLT dataset's test fold. Evidently, the evaluated data-driven filters (SIN, VSE and DANSE) struggled to implicitly model the influence of control inputs on the robot's trajectory. The model-based KF and the hybrid filters (KNet, APBM,

RKNet and ResEKF), however, had access to at least a nominal model of these inputs. However, Figure 3 shows that the KF, KNet and APBM almost always followed (centre panel) the noisy GPS measurements (left panel). ResEKF outperformed RKNet on accuracy and probabilistic calibration. CDFs plotted in Figure 3 aggregates performance across all seeds.

Table 3: **Results on the NCLT dataset.** Data-driven filters fail to implictly model the influence of control inputs. ResEKF outperforms the KF and competitors.

Filter	MSE $[m^2] \downarrow$	NLL $\downarrow$
KF	13.17	<u>5.43</u>
SIN	$> 10^3$	$> 10^3$
VSE	$> 10^3$	$12.57_{(12.45,\ 12.61)}$
DANSE	$> 5 \cdot 10^2$	$12.56_{(8.25,\ 20.63)}$
KNet	$13.15_{(13.12,\ 13.16)}$	n.a.
APBM	$13.13_{(12.98,\ 13.43)}$	$5.63_{(5.62,\ 5.70)}$
RKNet	<u>12.16</u>$_{(12.06,\ 12.38)}$	$5.73_{(5.64,\ 5.88)}$
ResEKF	**9.40**$_{(8.95,\ 9.50)}$	**5.25**$_{(5.23,\ 5.26)}$

5 Related Work

The literature at the intersection of deep learning and classical state estimation is rich, but can be mapped to multiple distinct objectives. *Smoothing* or non-causal state estimation (Anderson & Moore, 1979) is an objective, pursued in works such as Garcia Satorras et al. (2019) and Revach et al. (2023). Data Assimilation with Machine Learning (DA-ML) (Bocquet et al., 2019; Brajard et al., 2020; Farchi et al., 2021) uses neural networks to augment or fully replace a nominal transition function, thus performing *System Identification*. A model-based filter operating on the resulting model remains vulnerable to approximation error and, since the model itself is unbounded, has no assurance of satisfying regularity conditions stability guarantees may require. Deep generative state space models (DGSSMs) (Krishnan et al., 2017; Fraccaro et al., 2016; Karl et al., 2017; Dowling et al., 2024) employ variational inference to jointly learn a transition and/or measurement model alongside an 'inference network' capable of data-driven *filtering*. Neural *filtering*, sometimes called 'differentiable filtering' (Kloss et al., 2021) can be purely data-driven or hybrid. We situate our work as in intervention in the hybrid filtering paradigm. Data-driven Recurrent Kalman Networks (RKN) (Becker et al., 2019) encode measurements into a feature space where linear Kalman filtering is performed using a learned transition matrix, before decoding the resulting posterior into the filter's state space. Backprop KF (BKF) (Haarnoja et al., 2016) can function as a hybrid filter, using nominal transition and pseudo-measurement models that map nominal states to pseudo-measurements, which a neural network extracts from the true measurements. However, to our knowledge, while data-driven and prior hybrid filters have been shown to be performant, they lack the formal stability guarantees of classical model-based filters.

6 Conclusion

Stability is a cornerstone of classical state estimation theory and essential for safety-critical applications. Yet, it remains understudied in hybrid filtering literature, where the neural expressivity that compensates for model misspecifications risks filter divergence. We propose the Bounded Residual Corrections (BRC) framework as a step towards resolving this fundamental tension between expressivity and divergence in hybrid filtering. For linearisation-based filters, we prove that bounding learned residuals is necessary to guarantee stability, and enforcing a sufficiently small bound is sufficient to prevent the destabilisation of inherently stable nominal models. Furthermore, across linear, nonlinear, and real-world evaluations, we demonstrate that despite restricting neural expressivity, norm-bounding can enhance rather than hinder estimation accuracy and probabilistic calibration. Given BRC's operational modularity, these results motivate future work into investigating whether norm-bounding can similarly correct misspecification while preserving native stability properties—such as particle weight non-degeneration—in other filter families like Sequential Monte Carlo.

AI USE STATEMENT

We obtained critical ingredients for proving our mathematical claims from GenAI. Specifically, it helped us discover prior work (particularly (Reif et al., 1999; Dieudonné, 1969)), which were critically important for our theoretical work. We did not use GenAI for the following purposes: generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformating of datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used GenAI assistance to create the figures (independently verified) and edit the paper for better readability. GenAI also assisted more broadly with discovery and retrieval of literature. We take responsibility for the final content of this work.

ETHICS STATEMENT

Like all advances in robust state estimation, BRC carries dual-use potential: analysis to filter stability under adversarial conditions benefit civilian safety-critical systems but may equally apply to autonomous systems in contested environments. We intend for our work to exclusively benefit civilian applications. If selected for publication, we intend to release our code publicly under the Hippocratic License 3.0 (HL3) Haque (2021).

REPRODUCIBILITY STATEMENT

Anonymised code has been attached as a supplement to the paper. The paper itself details all original algorithms we use or propose. Hyperparameter grids are supplied in Appendix E, where details pertaining to training, datasets, model constructions etc. are shared.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

D. Alspach and H. Sorenson. Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, August 1972. ISSN 1558-2523. doi: 10.1109/TAC.1972.1100034. URL https://ieeexplore.ieee.org/abstract/document/1100034.

Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice Hall, Englewood Cliffs, N.J, 1979. ISBN 978-0-13-638122-8.

Jeffrey L. Anderson and Stephen L. Anderson. A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Monthly Weather Review*, 127(12):2741–2758, December 1999. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1999)127⟨2741:AMCIOT⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/127/12/1520-0493_1999_127_2741_amciot_2.0.co_2.xml.

Ienkaran Arasaratnam and Simon Haykin. Cubature Kalman Filters. *IEEE Transactions on Automatic Control*, 54(6):1254–1269, June 2009. ISSN 1558-2523. doi: 10.1109/TAC.2009.2019800. URL https://ieeexplore.ieee.org/document/4982682.

Yaakov Bar-Shalom, Xiao-Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*. Wiley, New York, NY, 2001. ISBN 978-0-471-41655-5.

Axel Barrau and Silvère Bonnabel. The Invariant Extended Kalman Filter as a Stable Observer. *IEEE Transactions on Automatic Control*, 62(4):1797–1812, April 2017. ISSN 1558-2523. doi: 10.1109/TAC.2016.2594085. URL https://ieeexplore.ieee.org/document/7523335/authors.

Philipp Becker, Harit Pandya, Gregor Gebhardt, Cheng Zhao, C. James Taylor, and Gerhard Neumann. Recurrent Kalman Networks: Factorized Inference in High-Dimensional Deep Feature Spaces. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 544–552. PMLR, May 2019. URL https://proceedings.mlr.press/v97/becker19a.html.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Marc Bocquet, Julien Brajard, Alberto Carrassi, and Laurent Bertino. Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Processes in Geophysics*, 26(3):143–162, July 2019. ISSN 1023-5809. doi: 10.5194/npg-26-143-2019. URL https://npg.copernicus.org/articles/26/143/2019/.

Julien Brajard, Alberto Carrassi, Marc Bocquet, and Laurent Bertino. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *Journal of Computational Science*, 44:101171, July 2020. ISSN 1877-7503. doi: 10.1016/j.jocs.2020.101171. URL https://www.sciencedirect.com/science/article/pii/S1877750320304725.

Olivier Cappé and Eric Moulines. On-Line Expectation–Maximization Algorithm for latent Data Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(3):593–613, June 2009. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2009.00698.x. URL https://doi.org/10.1111/j.1467-9868.2009.00698.x.

Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. University of Michigan North Campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9):1023–1035, August 2016. ISSN 0278-3649. doi: 10.1177/0278364915614638. URL https://doi.org/10.1177/0278364915614638.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. Differentiable Particle Filtering via Entropy-Regularized Optimal Transport. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2100–2111. PMLR, July 2021. URL https://proceedings.mlr.press/v139/corenflos21a.html.

Kevin Course and Prasanth B. Nair. State estimation of a physical system with unknown governing equations. *Nature*, 622(7982):261–267, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06574-8. URL https://www.nature.com/articles/s41586-023-06574-8.

Kaveh Dehghanpour, Zhaoyu Wang, Jianhui Wang, Yuxuan Yuan, and Fankun Bu. A Survey on State Estimation Techniques and Challenges in Smart Distribution Systems. *IEEE Transactions on Smart Grid*, 10(2):2312–2322, March 2019. ISSN 1949-3061. doi: 10.1109/TSG.2018.2870600. URL https://ieeexplore.ieee.org/abstract/document/8466598.

Jean Dieudonné. *Foundations of Modern Analysis*, volume 10-I of *Pure and Applied Mathematics*. Academic Press, New York-London, Enlarged and Corrected Printing edition, 1969. ISBN 978-0-12-215550-5.

Matthew Dowling, Yuan Zhao, and Il Memming Park. eXponential FAmily Dynamical Systems (XFADS): Large-scale nonlinear Gaussian state-space modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 13458–13488. Curran Associates, Inc., 2024. doi: 10.52202/079017-0430. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/18595bc3e802a3b11035927fd928eb9c-Abstract-Conference.html.

Gerardo Duran-Martin, Matias Altamirano, Alex Shestopaloff, Leandro Sánchez-Betancourt, Jeremias Knoblauch, Matt Jones, Francois-Xavier Briol, and Kevin Patrick Murphy. Outlier-robust Kalman Filtering through Generalised Bayes. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12138–12171. PMLR, July 2024. URL https://proceedings.mlr.press/v235/duran-martin24a.html.

G.A. Einicke and L.B. White. Robust extended Kalman filtering. *IEEE Transactions on Signal Processing*, 47(9):2596–2599, September 1999. ISSN 1941-0476. doi: 10.1109/78.782219. URL https://ieeexplore.ieee.org/document/782219.

Huazhen Fang, Mulugeta A. Haile, and Yebin Wang. Robust Extended Kalman Filtering for Systems With Measurement Outliers. *IEEE Transactions on Control Systems Technology*, 30(2): 795–802, March 2022. ISSN 1558-0865. doi: 10.1109/TCST.2021.3077535. URL https://ieeexplore.ieee.org/document/9444610.

Alban Farchi, Patrick Laloyaux, Massimo Bonavita, and Marc Bocquet. Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, 147(739):3067–3084, 2021. ISSN 1477-870X. doi: 10.1002/qj.4116. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.4116.

R. Fitzgerald. Divergence of the Kalman filter. *IEEE Transactions on Automatic Control*, 16(6): 736–747, December 1971. ISSN 1558-2523. doi: 10.1109/TAC.1971.1099836. URL https://ieeexplore.ieee.org/document/1099836.

Marco Fraccaro, Sø ren Kaae Sø nderby, Ulrich Paquet, and Ole Winther. Sequential Neural Models with Stochastic Layers. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/208e43f0e45c4c78cafadb83d2888cb6-Abstract.html.

Victor Garcia Satorras, Zeynep Akata, and Max Welling. Combining Generative and Discriminative Models for Hybrid Inference. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a660d4563b8f62dd5282319cc643d950-Abstract.html.

T. C. Gard. *Introduction to Stochastic Differential Equations*. Marcel Dekker, New York, NY, 1988. ISBN 978-0-8247-7776-0.

Anubhab Ghosh, Antoine Honoré, and Saikat Chatterjee. DANSE: Data-Driven Non-Linear State Estimation of Model-Free Process in Unsupervised Learning Setup. *IEEE Transactions on Signal Processing*, 72:1824–1838, 2024. ISSN 1941-0476. doi: 10.1109/TSP.2024.3383277. URL https://ieeexplore.ieee.org/document/10485649.

N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F Radar and Signal Processing*, 140(2):107, 1993. ISSN 0956375X. doi: 10.1049/ip-f-2.1993.0015. URL https://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0015.

Ido Greenberg, Netanel Yannay, and Shie Mannor. Optimization or Architecture: How to Hack Kalman Filtering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50482–50505, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9dfcc83c01e94d02c751c47517855c9f-Abstract-Conference.html.

Tuomas Haarnoja, Anurag Ajay, Sergey Levine, and Pieter Abbeel. Backprop KF: Learning Discriminative Deterministic State Estimators. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/697e382cfd25b07a3e62275d3ee132b3-Abstract.html.

Sameeul Haque. Hippocratic License 3.0. https://firstdonoharm.dev, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear Estimation*. Prentice Hall, Upper Saddle River, NJ, 2000. ISBN 0-13-022464-2.

Nikolas Kantas, Arnaud Doucet, Sumeetpal S. Singh, Jan Maciejowski, and Nicolas Chopin. On Particle Methods for Parameter Estimation in State-Space Models. *Statistical Science*, 30(3):328–351, August 2015. ISSN 0883-4237, 2168-8745. doi: 10.1214/14-STS511. URL https://projecteuclid.org/journals/statistical-science/volume-30/issue-3/On-Particle-Methods-for-Parameter-Estimation-in-State-Space-Models/10.1214/14-STS511.full.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data. In *5th International Conference on Learning Representations, ICLR 2017*, February 2017. URL https://openreview.net/forum?id=HyTqHL5xg.

Alina Kloss, Georg Martius, and Jeannette Bohg. How to train your differentiable filter. *Autonomous Robots*, 45(4):561–578, May 2021. ISSN 1573-7527. doi: 10.1007/s10514-021-09990-9. URL https://doi.org/10.1007/s10514-021-09990-9.

Jonathan Ko, Daniel J. Klein, Dieter Fox, and Dirk Haehnel. GP-UKF: Unscented kalman filters with Gaussian process prediction and observation models. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1901–1907, October 2007. doi: 10.1109/IROS.2007.4399284. URL https://ieeexplore.ieee.org/document/4399284. ISSN: 2153-0866.

Rahul Krishnan, Uri Shalit, and David Sontag. Structured Inference Networks for Nonlinear State Space Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), February 2017. ISSN 2374-3468. doi: 10.1609/aaai.v31i1.10779. URL https://ojs.aaai.org/index.php/AAAI/article/view/10779.

Aakash Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved Sequence Modeling using State Space Principles. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=HwCvaJOiCj.

Lennart Ljung. *System Identification: Theory for the User*. Prentice-Hall information and system sciences series. Prentice Hall, Upper Saddle River, NJ, 2. ed., 14. printing edition, December 1998. ISBN 978-0-13-656695-3.

Lennart Ljung and Torkel Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265–276, February 1994. ISSN 0005-1098. doi: 10.1016/0005-1098(94)90029-9. URL https://www.sciencedirect.com/science/article/pii/0005109894900299.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, UAI'01, pp. 362–369, San Francisco, CA, USA, August 2001. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-800-9. URL https://dl.acm.org/doi/10.5555/2074022.2074067.

Hassan Mortada, Cyril Falcon, Yanis Kahil, Mathéo Clavaud, and Jean-Philippe Michel. Recursive KalmanNet: Deep Learning-Augmented Kalman Filtering for State Estimation with Consistent Uncertainty Quantification. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pp. 885–889, September 2025. doi: 10.23919/EUSIPCO63237.2025.11226444. URL https://ieeexplore.ieee.org/document/11226444.

Christian A. Naesseth, Fredrik Lindsten, and Thomas B. Schön. Elements of Sequential Monte Carlo. *Foundations and Trends in Machine Learning*, 12(3):187–306, November 2019. ISSN 1935-8237. doi: 10.1561/2200000074. URL https://doi.org/10.1561/2200000074.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Gustav Norén, Anubhab Ghosh, Fredrik Cumlin, and Saikat Chatterjee. VSE: Variational State Estimation of Complex Model-Free Process. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4921–4925, May 2026. doi: 10.1109/ICASSP55912.2026.11463548. URL https://ieeexplore.ieee.org/document/11463548. ISSN: 2379-190X.

Robert Piché, Simo Särkkä, and Jouni Hartikainen. Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate student-t distribution. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, September 2012. doi: 10.1109/MLSP.2012.6349794. URL https://ieeexplore.ieee.org/document/6349794. ISSN: 2378-928X.

K. Reif, S. Gunther, E. Yaz, and R. Unbehauen. Stochastic stability of the discrete-time extended Kalman filter. *IEEE Transactions on Automatic Control*, 44(4):714–728, April 1999. ISSN 1558-2523. doi: 10.1109/9.754809. URL https://ieeexplore.ieee.org/abstract/document/754809.

Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adrià López Escoriza, Ruud J. G. van Sloun, and Yonina C. Eldar. KalmanNet: Neural Network Aided Kalman Filtering for Partially Known Dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022. ISSN 1941-0476. doi: 10.1109/TSP.2022.3158588. URL https://ieeexplore.ieee.org/abstract/document/9733186. Conference Name: IEEE Transactions on Signal Processing.

Guy Revach, Xiaoyong Ni, Nir Shlezinger, Ruud J. G. van Sloun, and Yonina C. Eldar. RTSNet: Learning to Smooth in Partially Known State-Space Models. *IEEE Transactions on Signal Processing*, 71:4441–4456, 2023. ISSN 1941-0476. doi: 10.1109/TSP.2023.3329964. URL https://ieeexplore.ieee.org/abstract/document/10322579.

Reza Sameni, Mohammad B. Shamsollahi, Christian Jutten, and Gari D. Clifford. A Nonlinear Bayesian Filtering Framework for ECG Denoising. *IEEE Transactions on Biomedical Engineering*, 54(12), December 2007. ISSN 1558-2531. doi: 10.1109/TBME.2007.897817. URL https://ieeexplore.ieee.org/document/4360034.

Nabil Shaukat, Ahmed Ali, Muhammad Javed Iqbal, Muhammad Moinuddin, and Pablo Otero. Multi-Sensor Fusion for Underwater Vehicle Localization by Augmentation of RBF Neural Network and Error-State Kalman Filter. *Sensors*, 21(4):1149, January 2021. ISSN 1424-8220. doi: 10.3390/s21041149. URL https://www.mdpi.com/1424-8220/21/4/1149.

Siyuan Shen, Jichen Chen, Guanfeng Yu, Zhengjun Zhai, and Pujie Han. KalmanFormer: using transformer to model the Kalman Gain in Kalman Filters. *Frontiers in Neurorobotics*, 18, January 2025. ISSN 1662-5218. doi: 10.3389/fnbot.2024.1460255. URL https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2024.1460255/full.

Nir Shlezinger, Guy Revach, Anubhab Ghosh, Saikat Chatterjee, Shuo Tang, Tales Imbiriba, Jindrich Dunik, Ondrej Straka, Pau Closas, and Yonina C. Eldar. Artificial Intelligence-Aided Kalman Filters: AI-Augmented Designs for Kalman-Type Algorithms. *IEEE Signal Processing Magazine*, 42(3):52–76, May 2025. ISSN 1558-0792. doi: 10.1109/MSP.2025.3569395. URL https://ieeexplore.ieee.org/document/11016798.

Gerald L. Smith, Stanley F. Schmidt, and Leonard A. McGee. Application of Statistical Filter Theory to the Optimal Estimation of Position and Velocity on Board a Circumlunar Vehicle. Technical Report NASA-TR-R-135, NASA Ames Research Center, Moffett Field, CA, United States, January 1962. URL https://ntrs.nasa.gov/citations/19620006857.

Jian Song, Wei Mei, Yunfeng Xu, Qiang Fu, and Lina Bu. Practical Implementation of KalmanNet for Accurate Data Fusion in Integrated Navigation. *IEEE Signal Processing Letters*, 31: 1890–1894, 2024. ISSN 1558-2361. doi: 10.1109/LSP.2024.3431443. URL https://ieeexplore.ieee.org/document/10605082.

Eduardo D. Sontag. Input to State Stability: Basic Concepts and Results. In Andrei A. Agrachev, A. Stephen Morse, Eduardo D. Sontag, Héctor J. Sussmann, Vadim I. Utkin, Paolo Nistri, and Gianna Stefani (eds.), *Nonlinear and Optimal Control Theory: Lectures given at the C.I.M.E. Summer School held in Cetraro, Italy June 19–29, 2004*, pp. 163–220. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-77653-6. doi: 10.1007/978-3-540-77653-6_3. URL https://doi.org/10.1007/978-3-540-77653-6_3.

Ondřej Straka, Jindřich Duník, Pau Closas, and Tales Imbiriba. Interpretable Augmented Physics-Based Model for Estimation and Tracking. In *2025 28th International Conference on Information Fusion (FUSION)*, pp. 1–8, July 2025. doi: 10.23919/FUSION65864.2025.11124036. URL https://ieeexplore.ieee.org/document/11124036.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. Intelligent Robotics and Autonomous Agents series. MIT Press, Cambridge, MA, USA, August 2005. ISBN 978-0-262-20162-9.

Kai Xiong, Hongyue Zhang, and Liangdong Liu. Adaptive Robust Extended Kalman Filter. In *Kalman Filter Recent Advances and Applications*. IntechOpen, April 2009. ISBN 978-953-307-000-1 978-953-51-5843-1. doi: 10.5772/6803. URL https://www.intechopen.com/chapters/6327.

Bo Xu, Peng Zhang, Hongzhi Wen, and Xu Wu. Stochastic stability and performance analysis of Cubature Kalman Filter. *Neurocomputing*, 186:218–227, April 2016. ISSN 0925-2312. doi: 10.1016/j.neucom.2015.12.087. URL https://www.sciencedirect.com/science/article/pii/S0925231215020639.

A GENERALISING BRC OPERATIONS TO DIVERSE FILTERS AND NEURAL ARCHITECTURES

In §2, we establish how the BRC framework applies to nominal filters that operate by propagating estimates of the first two state moments. However, the framework is operationally agnostic to both the nominal filter and the neural networks outputting residual correction terms, provided these terms are appropriately parameterised and norm-bounded. In support of this claim, §A.1 concretises how BRC can operate with three distinct families of filters, covering the nominal model required by the filters, their own update algorithms and how BRC can augment these routines. One of these families is sampling-based, where differentiability of operations may be a concern; we therefore also address this briefly. Further, we focus on the autonomous systems case, but as we show in §4.4, BRC can also operate with systems that use control inputs. §A.2 concretises how BRC's corrective residuals may be parameterised with three distinct types of neural networks.

A.1 REPRESENTATIVE EXAMPLES OF FILTER FAMILIES

A.1.1 MEAN-COVARIANCE PROPAGATING FILTERS

We focus here on two sub-families:

- Local linearisation-based filters, among which the Extended Kalman Filter (EKF) (Smith et al., 1962) is canonical.
- Cubature filters, among which the Cubature Kalman Filter (CKF) (Arasaratnam & Haykin, 2009) is canonical.

Nominal Model. The EKF and CKF both require a model as specified in Eq. (2). The CKF assumes that the zero-mean noise terms v_t and w_t are Gaussian, whereas the EKF only requires that they have valid covariances Q_t and Σ_t, respectively. On the other hand, the EKF requires that the transition and measurement functions, f and h, be differentiable.

Nominal Filtering Algorithms. The EKF's time- and measurement-update routines are presented in Algorithms 3 and 4, respectively, where we denote the Jacobian matrix of a vector-valued function ς evaluated at state x as $\mathbf{J}_\varsigma(x) := \frac{\partial \varsigma}{\partial x}\big|_x$. When f and h are linear, the EKF's updates simplify to the Kalman Filter's updates.

Algorithm 3 EKF Time Update

function TIMEUPDATE($\mu_{t-1|t-1}, P_{t-1|t-1}, f, Q_t$)
  $F_t \leftarrow \mathbf{J}_f(\mu_{t-1|t-1})$
  $\mu_{\text{EKF},\ t|t-1} \leftarrow f(\mu_{t-1|t-1})$
  $P_{\text{EKF},\ t|t-1} \leftarrow F_t P_{t-1|t-1} F_t^\top + Q_t$
  return $\mu_{\text{EKF},\ t|t-1}, P_{\text{EKF},\ t|t-1}$
end function

Algorithm 4 EKF Measurement Update

function MEASUPDATE($\mu_{t|t-1}, P_{t|t-1}, y_t, h, \Sigma_t$)
  $\hat{y}_t \leftarrow h(\mu_{t|t-1})$
  $H_t \leftarrow \mathbf{J}_h(\mu_{t|t-1})$
  $S_t \leftarrow H_t P_{t|t-1} H_t^\top + \Sigma_t$
  $K_t \leftarrow P_{t|t-1} H_t^\top S_t^{-1}$
  $\mu_{\text{EKF},\ t|t} \leftarrow \mu_{t|t-1} + K_t(y_t - \hat{y}_t)$
  $P_{\text{EKF},\ t|t} \leftarrow (\mathbb{I} - K_t H_t) P_{t|t-1}$ ▷ Joseph form could also be used
  return $\mu_{\text{EKF},\ t|t}, P_{\text{EKF},\ t|t}$
end function

The CKF's time- and measurement-update routines are presented in Algorithms 5 and 6, respectively, where we denote ξ as the block matrix $\sqrt{s} \cdot [\mathbb{I} \mid -\mathbb{I}]$ of shape $2s \times s$, where $\mathbb{I}$ is the identity matrix of dimension s.

Algorithm 5 CKF Time Update

function TIMEUPDATE($\mu_{t-1|t-1}, P_{t-1|t-1}, f, Q_t$)
  $L_{t-1|t-1} \leftarrow \text{CHOLESKYFACTOR}(P_{t-1|t-1})$ ▷ Lower Triangular Factor: $s \times s$
  $X_{t-1|t-1} \leftarrow L_{t-1|t-1}\xi^\top + \mu_{t-1|t-1}\mathbf{1}_{1\times 2s}$ ▷ Cubature Points Matrix: $s \times 2s$
  $X^*_{t|t-1} \leftarrow f(X_{t-1|t-1})$ ▷ Propagated Points Matrix: $s \times 2s$
  $\mu_{\text{CKF},\, t|t-1} \leftarrow \frac{1}{2s} X^*_{t|t-1}\mathbf{1}_{2s\times 1}$
  $\mathcal{X}_{t|t-1} \leftarrow \frac{1}{\sqrt{2s}}\left(X^*_{t|t-1} - \mu_{\text{CKF},\, t|t-1}\mathbf{1}_{1\times 2s}\right)$ ▷ Weighted Centred Error Matrix: $s \times 2s$
  $P_{\text{CKF},\, t|t-1} \leftarrow \mathcal{X}_{t|t-1}\left(\mathcal{X}_{t|t-1}\right)^\top + Q_t$
  return $\mu_{\text{CKF},\, t|t-1}, P_{\text{CKF},\, t|t-1}$
end function

Algorithm 6 CKF Measurement Update

function MEASUPDATE($\mu_{t|t-1}, P_{t|t-1}, y_t, h, \Sigma_t$)
  $L_{t|t-1} \leftarrow \text{CHOLESKYFACTOR}(P_{t|t-1})$ ▷ Lower Triangular Factor: $s \times s$
  $X_{t|t-1} \leftarrow L_{t|t-1}\xi^\top + \mu_{t|t-1}\mathbf{1}_{1\times 2s}$ ▷ Cubature Points Matrix: $s \times 2s$
  $Y_{t|t-1} \leftarrow h(X_{t|t-1})$ ▷ Predicted Measurement Points Matrix: $d \times 2s$
  $\hat{y}_{t|t-1} \leftarrow \frac{1}{2s} Y_{t|t-1}\mathbf{1}_{2s\times 1}$ ▷ Predicted Measurement Mean: $d \times 1$
  $\tilde{Y}_{t|t-1} \leftarrow Y_{t|t-1} - \hat{y}_{t|t-1}\mathbf{1}_{1\times 2s}$ ▷ Unweighted Centred Measurement Error Matrix: $d \times 2s$
  $S_{t|t-1} \leftarrow \frac{1}{2s}\tilde{Y}_{t|t-1}\tilde{Y}^\top_{t|t-1} + \Sigma_t$ ▷ Innovation Covariance: $d \times d$
  $\tilde{X}_{t|t-1} \leftarrow X_{t|t-1} - \mu_{t|t-1}\mathbf{1}_{1\times 2s}$ ▷ Unweighted Centred State Error Matrix: $s \times 2s$
  $P_{xy,\, t|t-1} \leftarrow \frac{1}{2s}\tilde{X}_{t|t-1}\tilde{Y}^\top_{t|t-1}$ ▷ Cross-covariance Matrix: $s \times d$
  $K_t \leftarrow P_{xy,\, t|t-1}S^{-1}_{t|t-1}$ ▷ Gain Matrix: $s \times d$
  $\mu_{\text{CKF},\, t|t} \leftarrow \mu_{t|t-1} + K_t\left(y_t - \hat{y}_{t|t-1}\right)$
  $P_{\text{CKF},\, t|t} \leftarrow P_{t|t-1} - K_t S_{t|t-1}K_t^\top$
  return $\mu_{\text{CKF},\, t|t}, P_{\text{CKF},\, t|t}$
end function

BRC-augmented Filtering Algorithm. Algorithm 2 presented in §2 is precisely the procedure of BRC-augmentation for Mean-Covariance Propagation filters including EKF and CKF.

A.1.2 MIXTURE-OF-DENSITY FILTERS

Nominal Model. Mixture-of-Density filters operate under a model of the following form:

$$\begin{aligned} x_0 &\sim p_0(\theta_0,\ \alpha_0) \\ x_t|x_{t-1} &\sim p_x(\theta_x|x_{t-1}) \\ y_t|x_t &\sim p_y(\theta_y|x_t) \end{aligned} \tag{4}$$

where p_0 is the mixture density, with component distribution parameters θ_0 and mixture weights α_0, p_x is a conditional density, with parameters θ_x and p_y is a conditional density, with parameters θ_y. The Gaussian Sum Filter (GSF) (Alspach & Sorenson, 1972), in particular, assumes a mixture-of-Gaussians density for the initial state x_0 and Gaussian transition and emission (measurement model) densities. Precisely:

$$
\begin{aligned}
p(x_0) &= \sum_{i=1}^{M} \alpha_0^{(0)} \mathcal{N}(x_0 | \mu_0^{(i)},\ P_0^{(i)}) \\
p(x_t | x_{t-1}) &= \mathcal{N}(x_t | f(x_{t-1}),\ Q_t) \\
p(y_t | x_t) &= \mathcal{N}(y_t | h(x_t),\ \Sigma_t)
\end{aligned}
\tag{5}
$$

where M is the number of mixture components, with $\alpha_0^{(i)} \in [0, 1]$, $\mu_0^{(i)} \in \mathbb{R}^s$ and $P_0^{(i)} \in \mathbb{R}^{s \times s}$ being the weight, mean and covariance of component i and $\sum_{i=1}^{M} \alpha_0^{(i)} = 1$. The functions f and h are differentiable transition and measurement functions, and Q_t and Σ_t are covariance matrices of dimensions s and d, respectively.

Nominal Filtering Algorithm. Like the EKF and CKF, the GSF also operates by performing time- and measurement-updates, as presented in Algorithms 7 and 8, respectively, where, again, we denote the Jacobian matrix of a vector-valued function ς evaluated at state x as $\mathbf{J}_\varsigma(x) := \frac{\partial \varsigma}{\partial x}\Big|_x$. Note that the filtered state $x_t | y_{1:t}$ is itself assumed to be distributed as a mixture of M Gaussians, with weights $\{\alpha_t^{(i)}\}_{i=1}^M$ and component parameters $\{\mu^{(i)}_{\text{GSF},\ t|t}, P^{(i)}_{\text{GSF},\ t|t}\}_{i=1}^M$; the GSF propagates not just the mixture component means and coavariances but also the mixture weights.

Algorithm 7 GSF Time Update

function TIMEUPDATE($\{\mu^{(i)}_{t-1|t-1}, P^{(i)}_{t-1|t-1}\}_{i=1}^M, f, Q_t$)
  for $i = 1, \ldots, M$ **do**
    $\mu^{(i)}_{\text{GSF},\ t|t-1}, P^{(i)}_{\text{GSF},\ t|t-1} \leftarrow$ EKF-TIMEUPDATE($\mu^{(i)}_{t-1|t-1}, P^{(i)}_{t-1|t-1}, f, Q_t$)
  end for
  return $\{\mu^{(i)}_{\text{GSF},\ t|t-1}, P^{(i)}_{\text{GSF},\ t|t-1}\}_{i=1}^M$
end function

Algorithm 8 GSF Measurement Update

function MEASUPDATE($\{\alpha^{(i)}_{t-1}, \mu^{(i)}_{t|t-1}, P^{(i)}_{t|t-1}\}_{i=1}^M, y_t, h, \Sigma_t$)
  for $i = 1, \ldots, M$ **do**
    # We assume that EKF-MEASUPDATE *also returns* $\hat{y}_t^{(i)}$ *and* $S_t^{(i)}$*, which it does compute*
    $\mu^{(i)}_{\text{GSF},\ t|t}, P^{(i)}_{\text{GSF},\ t|t}, \hat{y}_t^{(i)}, S_t^{(i)} \leftarrow$ EKF-MEASUPDATE($\mu^{(i)}_{t|t-1}, P^{(i)}_{t|t-1}, y_t, h, \Sigma_t$)
    $w_t^{(i)} \leftarrow \alpha_{t-1}^{(i)} \cdot \mathcal{N}\left(y_t \mid \hat{y}_t^{(i)}, S_t^{(i)}\right)$
  end for
  for $i = 1, \ldots, M$ **do**
    $\alpha_t^{(i)} \leftarrow \frac{w_t^{(i)}}{\sum_{j=1}^M w_t^{(j)}}$ ▷ Normalize mixture weights
  end for
  return $\{\alpha_t^{(i)}, \mu^{(i)}_{\text{GSF},\ t|t}, P^{(i)}_{\text{GSF},\ t|t}\}_{i=1}^M$
end function

BRC-augmented Filtering Algorithm. In general, BRC could use $2M$ neural backbones, one for each of the M GSF components' EKF time- and measurement-update stages. It is a design decision whether to share parameters across these backbones or whether to apply BRC at both stages. We proceed with this general case, denoting the *a priori* and *a posteriori* neural backbones corresponding to the i-th component, respectively, as $\psi_1^{(i)}$ and $\psi_2^{(i)}$, with corresponding inputs $\mathcal{I}^{(i)}_{t|t-1}$ and $\mathcal{I}^{(i)}_{t|t}$. The BRC-augmented time- and measurement-updates for GSF then follow Algorithms 9 and 10. Each output of neural network $\psi_\cdot^{(i)}$ *may* be parameterised according to Algorithm 1. Notice that in GSF's native measurement-update (Algorithm 8), the mixture weights $\alpha_t^{(i)}$ do **not** depend on the filtered (*a posteriori*) state estimates; BRC preserves this property.

Algorithm 9 BRC-augmented GSF Time Update

function TIMEUPDATE($\{\mu^{(i)}_{t-1|t-1}, P^{(i)}_{t-1|t-1}, \psi^{(i)}_1\}_{i=1}^M, f, Q_t$)
  for $i = 1, \ldots, M$ **do**
    $\mu^{(i)}_{\text{GSF},\ t|t-1}, P^{(i)}_{\text{GSF},\ t|t-1} \leftarrow$ EKF-TIMEUPDATE$(\mu^{(i)}_{t-1|t-1}, P^{(i)}_{t-1|t-1}, f, Q_t)$
    $(\mu^{(i)}_{\text{BRC},\ t|t-1}, P^{(i)}_{\text{BRC},\ t|t-1}) \leftarrow \psi^{(i)}_1(\mathcal{I}^{(i)}_{t|t-1})$
    $\mu^{(i)}_{t|t-1} \leftarrow \mu^{(i)}_{\text{GSF},\ t|t-1} + \mu^{(i)}_{\text{BRC},\ t|t-1}$
    $P^{(i)}_{t|t-1} \leftarrow P^{(i)}_{\text{GSF},\ t|t-1} + P^{(i)}_{\text{BRC},\ t|t-1}$
  end for
  return $\{\mu^{(i)}_{t|t-1}, P^{(i)}_{t|t-1}\}_{i=1}^M$
end function

Algorithm 10 BRC-augmented GSF Measurement Update

function MEASUPDATE($\{\alpha^{(i)}_{t-1}, \mu^{(i)}_{t|t-1}, P^{(i)}_{t|t-1}, \psi^{(i)}_2\}_{i=1}^M, y_t, h, \Sigma_t$)
  for $i = 1, \ldots, M$ **do**
    $\mu^{(i)}_{\text{GSF},\ t|t}, P^{(i)}_{\text{GSF},\ t|t}, \hat{y}^{(i)}_t, S^{(i)}_t \leftarrow$ EKF-MEASUPDATE$(\mu^{(i)}_{t|t-1}, P^{(i)}_{t|t-1}, y_t, h, \Sigma_t)$
    $(\mu^{(i)}_{\text{BRC},\ t|t}, P^{(i)}_{\text{BRC},\ t|t}) \leftarrow \psi^{(i)}_2(\mathcal{I}^{(i)}_{t|t})$
    $\mu^{(i)}_{t|t} \leftarrow \mu^{(i)}_{\text{GSF},\ t|t} + \mu^{(i)}_{\text{BRC},\ t|t}$
    $P^{(i)}_{t|t} \leftarrow P^{(i)}_{\text{GSF},\ t|t} + P^{(i)}_{\text{BRC},\ t|t}$
    $w^{(i)}_t \leftarrow \alpha^{(i)}_{t-1} \cdot \mathcal{N}\left(y_t \mid \hat{y}^{(i)}_t, S^{(i)}_t\right)$
  end for
  for $i = 1, \ldots, M$ **do**
    $\alpha^{(i)}_t \leftarrow \frac{w^{(i)}_t}{\sum_{j=1}^M w^{(j)}_t}$ ▷ Normalize mixture weights
  end for
  return $\{\alpha^{(i)}_t, \mu^{(i)}_{t|t}, P^{(i)}_{t|t}\}_{i=1}^M$
end function

A.1.3 SAMPLING-BASED (SEQUENTIAL MONTE CARLO) FILTERS

We focus on the Bootstrap Filter (BF) (Gordon et al., 1993), a representative example of Sequential Monte Carlo (SMC) (Naesseth et al., 2019) filters, which represent the conditional distribution of $x_t|y_{1:t}$ as a set of particles (samples) with weights.

Nominal Model and Nominal Filtering Algorithm. The BF requires specification of:

- A transition distribution with pdf $p_{\theta_x}(x_t|x_{t-1})$, from which one may sample realisations.
- An emission pdf $p_{\theta_y}(y_t|x_t)$, specified at least up to its normalising constant.
- The number of particles (samples), N.

The BF iterates Algorithm 11 through time.

BRC-augmented Filtering Algorithm. We present one possible approach for BRC-augmented bootstrap filtering, where the prior particles (i.e., particles sampled from the nominal transition distribution (Step 3 of Algorithm 11)) are augmented with norm-bounded residuals. Let BRC's neural backbone be denoted by ψ and its input by $\mathcal{I}_t$ at time-step t. This input may include the current time step's prior particles, previous un-normalised weights, and any other causal information. This BRC-augmented Bootstrap Filter iteration is presented in Algorithm 13.

Learning BRC's neural backbones. BRC's neural backbones are learnable via backpropagation in differentiable filtering routines. However, the sampling and resampling steps in the SMC

Algorithm 11 Bootstrap Filter Iteration

function BF-ITERATION($\{x_{t-1}^{(i)}\}_{i=1}^N, y_t, p_{\theta_x}, p_{\theta_y}$)
  for $i = 1, \dots, N$ **do**
    Sample $\tilde{x}_{\text{BF},\,t}^{(i)} \sim p_{\theta_x}(\cdot|x_{t-1}^{(i)})$ ▷ Sample from Transition Distribution
    $\tilde{w}_t^{(i)} \leftarrow p_{\theta_y}(y_t \mid \tilde{x}_{\text{BF},\,t}^{(i)})$ ▷ Un-normalised Weights (Emission Likelihood)
  end for
  for $i = 1, \dots, N$ **do**
    $w_t^{(i)} \leftarrow \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^N \tilde{w}_t^{(j)}}$ ▷ Normalise Weights
  end for
  $\{x_t^{(i)}\}_{i=1}^N \leftarrow$ RESAMPLE($\{\tilde{x}_{\text{BF},\,t}^{(i)}, w_t^{(i)}\}_{i=1}^N$) ▷ See Algorithm 12
  return $\{x_t^{(i)}\}_{i=1}^N$
end function

Algorithm 12 Particle Resampling

function RESAMPLE($\{\tilde{x}^{(i)}, w^{(i)}\}_{i=1}^N$)
  for $i = 1, \dots, N$ **do**
    Sample index $a_i \sim$ Categorical($\{w^{(1)}, \dots, w^{(N)}\}$)
    $x^{(i)} \leftarrow \tilde{x}^{(a_i)}$
  end for
  return $\{x^{(i)}\}_{i=1}^N$
end function

filtering routines may not be. We therefore suggest using an Online Expectation–Maximisation Algorithm (Cappé & Moulines, 2009; Kantas et al., 2015) to update the neural parameters, without requiring backpropagation through the discrete (re)sampling steps. Recent advances in Differentiable Particle Filtering (Corenflos et al., 2021) may also be relevant to this issue, as they replace discrete sampling with differentiable alternatives.

A.2 Representative Examples of Neural Backbones

Consider the following mappings from the causally available information set $\mathcal{I}_k$ to a latent embedding z, using three representative neural backbones; k is $t|t$ or $t|t-1$ for Mean-Covariance Propagating filters and the GSF, and t for the Bootstrap Filter:

$$\textbf{MLP:} \quad z = \text{MLP}(\mathcal{I}_k) \tag{6a}$$

$$\textbf{RNN:} \quad z = \text{Dec}(h), \quad h = \text{RNN}(\mathcal{I}_k, h_{\text{prev}}) \tag{6b}$$

$$\textbf{Transformer:} \quad z = \text{CausalTransformer}(\mathcal{I}_k, \mathcal{I}_{k-1}, \dots) \tag{6c}$$

where Dec is a decoding network, h represents the recursive RNN hidden state, and Eq. (6c) utilises a causal transformer (e.g., self-attention and feed-forward non-linear layers) over a historical context window, with the final time-step's embedding considered as z. Eqs. (6a)–(6c) represent possible implementations of the FORWARDPASS function in Algorithm 1. The latent embedding z can thus function as a precursor to BRC's norm-bounded residuals. The residual covariance parameterisation step (step 5) of Algorithm 1 is presented as a rank-1 formulation, but can alternatively be a rank-n or a full Cholesky factor formulation; the dimensionality of z must, in such a case, be increased accordingly. Additionally, in §4.4, we use an alternative to Algorithm 1. In general, the practitioner may need to choose an architecture that works best for their application.

B Latent State Non-Identifiability

Figures 4–7 illustrate the non-identifiability vulnerability described in §2 across four representative filters spanning hybrid and purely data-driven approaches in the Linear Benchmark experiment

Algorithm 13 BRC-Augmented Bootstrap Filter Iteration

function BF-ITERATION($\{x_{t-1}^{(i)}\}_{i=1}^{N}, y_t, p_{\theta_x}, p_{\theta_y}, \psi$)
  for $i = 1, \ldots, N$ **do**
    Sample $\tilde{x}_{\text{BF},\,t}^{(i)} \sim p_{\theta_x}(\cdot | x_{t-1}^{(i)})$ ▷ Sample from Transition Distribution
  end for
  $\{\tilde{x}_{\text{BRC},\,t}^{(i)}\}_{i=1}^{N} \leftarrow \psi(\mathcal{I}_t)$
  for $i = 1, \ldots, N$ **do**
    $\tilde{x}_t^{(i)} \leftarrow \tilde{x}_{\text{BF},\,t}^{(i)} + \tilde{x}_{\text{BRC},\,t}^{(i)}$
    $\tilde{w}_t^{(i)} \leftarrow p_{\theta_y}(y_t \mid \tilde{x}_t^{(i)})$ ▷ Un-normalised Weights (Emission Likelihood)
  end for
  for $i = 1, \ldots, N$ **do**
    $w_t^{(i)} \leftarrow \frac{\tilde{w}_t^{(i)}}{\sum_{j=1}^{N} \tilde{w}_t^{(j)}}$ ▷ Normalise Weights
  end for
  $\{x_t^{(i)}\}_{i=1}^{N} \leftarrow \text{RESAMPLE}(\{\tilde{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^{N})$ ▷ See Algorithm 12
  return $\{x_t^{(i)}\}_{i=1}^{N}$
end function

(§4.2); blue and green curves represent state and measurement metrics (NLL or MSE) respectively. For each filter, we contrast a *reference configuration* (dashed curve) against the *selected configuration* (solid curve) chosen by our proposed validation procedure (Appendix C). The reference configurations demonstrate that continued reduction in measurement NLL, or in the model KNet's case, measurement MSE, does not guarantee latent state fidelity. Reference and selected configuration hyperparameters are provided below for reproducibility.

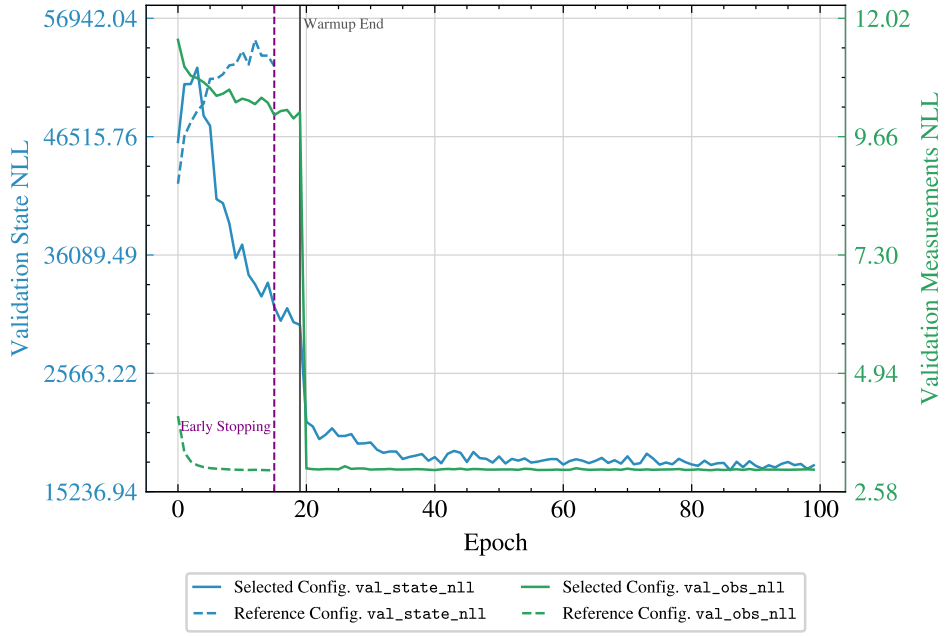

Figure 4: **GRU.** The reference configuration's measurement NLL quickly approaches that of the selected configuration; its state NLL, however, anti-correlates and exceeds 50k, illustrating that continued reduction in the former does not guarantee latent state fidelity. The reference configuration is halted by early stopping on the state NLL metric, while the selected configuration employs warmup loss and stabilises state NLL near 15k—avoiding the anti-correlation, but remaining poorly calibrated.

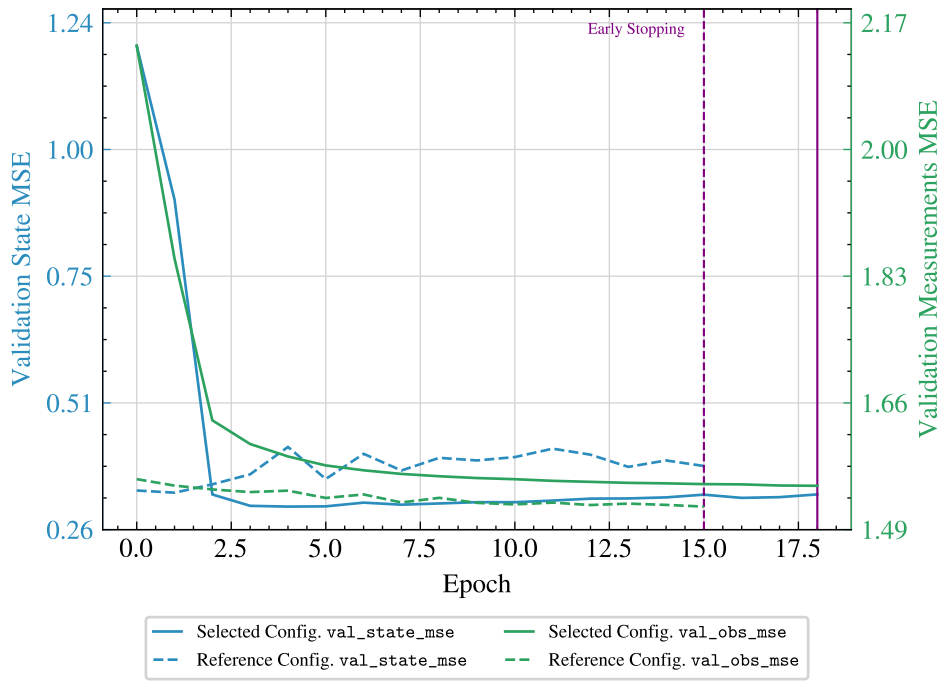

Figure 5: **KNet.** A deterministic model that does not produce calibrated uncertainty estimates; MSE replaces NLL and ANEES for validation (Appendix C). The reference configuration shows state and measurement MSE decoupling after epoch 2, with the selected configuration achieving meaningfully better state MSE.

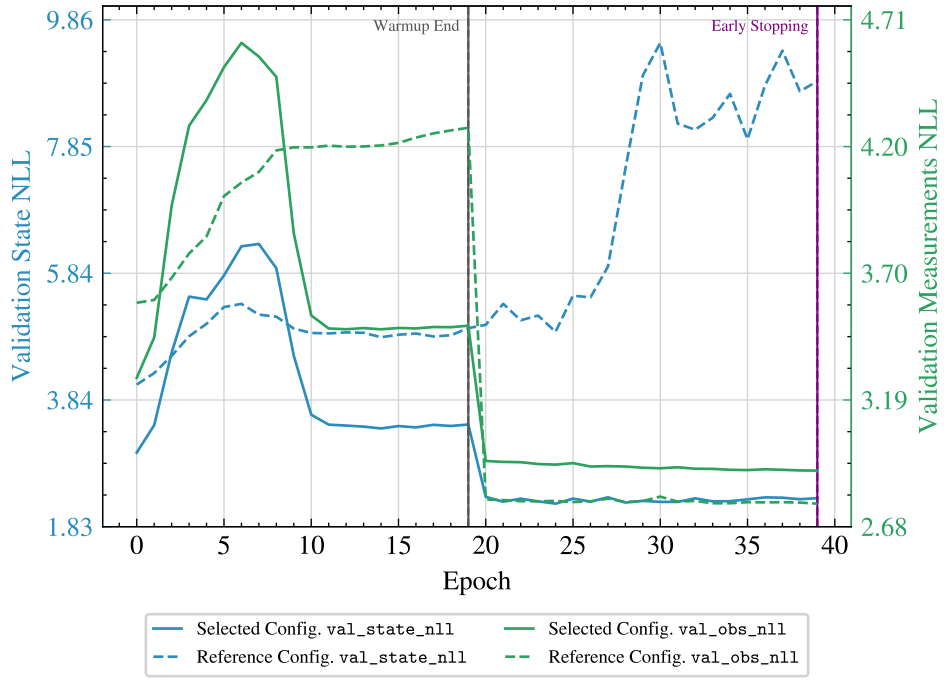

Figure 6: **ResEKF.** The non-identifiability vulnerability affects ResEKF as well, even if estimation errors remain bounded. The reference configuration exhibits anti-correlation in the post-warmup phase, while the selected configuration maintains closer alignment between state and measurement NLL, illustrating the effectiveness of the proposed validation procedure.

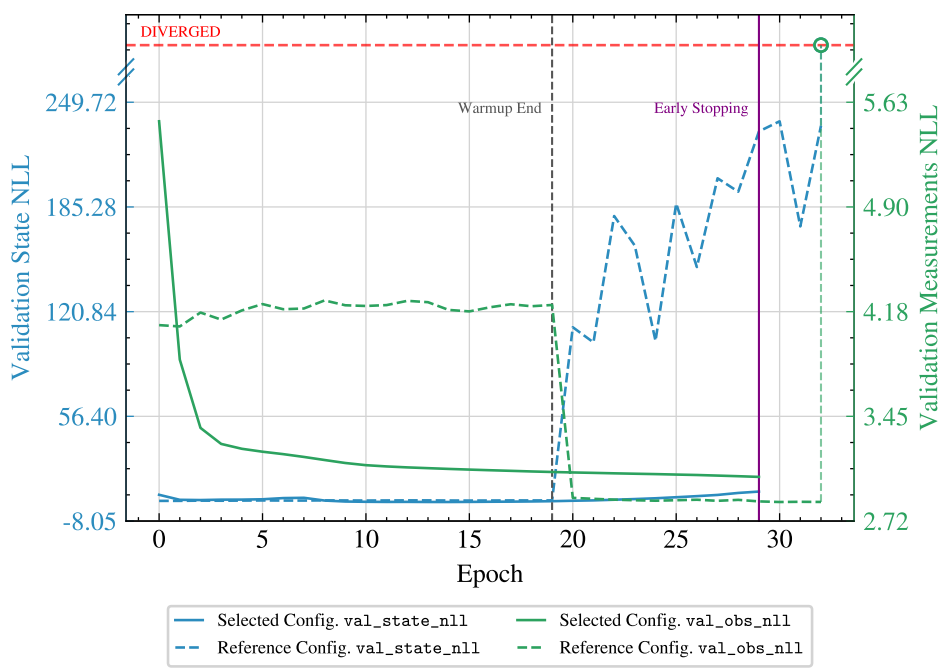

Figure 7: **RKNet.** The starkest case: the reference configuration's state NLL diverges to NaN after warmup while measurement NLL continues to improve, demonstrating that unconstrained gain-learning methods can catastrophically decouple latent state estimate quality from measurement fit. The proposed validation procedure identifies and avoids this degenerate region.

Reference configuration hyperparameters were obtained from the random search within the search space specified in Appendix E.2:

- **GRU**: `seed`: 42, `lr`: 8.71e-4, `weight_decay`: 6.26e-5, `warmup_loss`: N, `hidden_dim`: 26
- **KNet**: `seed`: 42, `lr`: 1.88e-3, `weight_decay`: 6.80e-5, `in_dim_mult`: 36, `out_dim_mult`: 21
- **ResEKF**: `seed`: 42, `lr`: 1.47e-3, `weight_decay`: 1.99e-4, `warmup_loss`: Y, `c_max`: 1.79
- **RKNet**: `seed`: 42, `lr`: 1.02e-3, `weight_decay`: 3.73e-3, `warmup_loss`: Y, `in_dim_mult`: 28, `out_dim_mult`: 26

Selected configuration hyperparameters, identified by the proposed validation procedure (Appendix C):

- **GRU**: `seed`: 42, `lr`: 1.21e-2, `weight_decay`: 2.47e-5, `warmup_loss`: Y, `hidden_dim`: 25
- **KNet**: `seed`: 42, `lr`: 1.14e-4, `weight_decay`: 4.63e-5, `in_dim_mult`: 39, `out_dim_mult`: 13
- **ResEKF**: `seed`: 42, `lr`: 4.09e-4, `weight_decay`: 3.63e-3, `warmup_loss`: Y, `c_max`: 0.49
- **RKNet**: `seed`: 42, `lr`: 3.74e-5, `weight_decay`: 3.03e-4, `warmup_loss`: N, `in_dim_mult`: 37, `out_dim_mult`: 1

The GRU filter was constructed in the same manner as the Mamba-3 filter, as described in Appendix E.1.4.

C Validation (Model Selection) Procedure

Checkpoint Selection. Let $\mathbb{H} := \{\mathbb{H}_1, \ldots, \mathbb{H}_M\}$ denote the ordered set of M hyperparameter configurations used to train a given filter. During training, each configuration is evaluated on a held-out validation dataset at the end of every training epoch. For each configuration, we identify its

optimal checkpoint, i.e., the set of learned parameters from the epoch that minimises a designated validation criterion. If the model is probabilistic, this criterion is the Gaussian state negative log-likelihood (NLL); otherwise, the criterion is the state mean squared error (MSE). Let the optimal checkpoint state MSE of a given hyperparameter configuration $\mathbb{H}_i$ be denoted by m_i and $\mathbb{M}$ be the ordered set $\{m_1, m_2, \ldots, m_M\}$. Further, if the model is probabilistic, let the optimal checkpoint Gaussian state NLL and absolute ANEES[3] deviation from the state dimension of $\mathbb{H}_i$ be denoted by n_i and a_i respectively. Let $\mathbb{N}$ and $\mathbb{A}$ be the ordered sets $\mathbb{N} := \{n_1, n_2, \ldots, n_M\}$ and $\mathbb{A} := \{a_1, a_2, \ldots, a_M\}$ if the model is probabilistic and empty sets otherwise.

Model Selection (Hyperparameter Optimisation). Using the optimal checkpoint metric sets $\mathbb{M}$, $\mathbb{N}$ and $\mathbb{A}$, and an instability threshold of $\mathcal{T} = 10^6$, we conduct model selection as presented in Algorithm 14. This procedure makes use of a metrics normalising function. Two possible implementations of this normalising function are shown in Algorithms 15 and 16. The latter is more robust to skewed distributions of metrics.

Algorithm 14 Model Selection (Hyperparameter Optimisation)

Require: Hyperparameter Configurations, $\mathbb{H}$
Require: Metric Sets, $\mathbb{M}$, $\mathbb{N}$ and $\mathbb{A}$
Require: Instability threshold, $\mathcal{T}$ (default: 10^6)

▷ **Stage 1: Candidate Filtering**

1: $\mathcal{V} \leftarrow \emptyset$
2: **for** $i = 1, \ldots, M$ **do**
3: **if** $\mathbb{H}_i$ results in filter divergence **then**
4: **continue**
5: **end if**
6: **if** $\mathbb{N} \neq \emptyset$ **then** ▷ Probabilistic Model
7: **if** $\max(m_i, n_i, a_i) \leq \mathcal{T}$ **then**
8: $\mathcal{V} \leftarrow \mathcal{V} \cup \{i\}$
9: **end if**
10: **else if** $m_i \leq \mathcal{T}$ **then** ▷ Deterministic Model
11: $\mathcal{V} \leftarrow \mathcal{V} \cup \{i\}$
12: **end if**
13: **end for**
14: **if** $\mathcal{V} = \emptyset$ **then**
15: **return** $\emptyset$ ▷ No Valid Configurations
16: **end if**

▷ **Stage 2: Metric Normalisation**

17: $\{\bar{m}_i\}_{i \in \mathcal{V}} \leftarrow \textsc{Normalise}(\{m_i\}_{i \in \mathcal{V}})$
18: **if** $\mathbb{N} \neq \emptyset$ **then**
19: $\{\bar{n}_i\}_{i \in \mathcal{V}} \leftarrow \textsc{Normalise}(\{n_i\}_{i \in \mathcal{V}})$
20: $\{\bar{a}_i\}_{i \in \mathcal{V}} \leftarrow \textsc{Normalise}(\{a_i\}_{i \in \mathcal{V}})$
21: **end if**

▷ **Stage 3: Configuration Scoring**

22: **for** $i \in \mathcal{V}$ **do**
23: **if** $\mathbb{N} \neq \emptyset$ **then**
24: $C_i \leftarrow \frac{1}{3}(\bar{m}_i + \bar{n}_i + \bar{a}_i)$ ▷ Probabilistic Model Score
25: **else**
26: $C_i \leftarrow \bar{m}_i$ ▷ Deterministic Model Score
27: **end if**
28: **end for**
29: $i^* \leftarrow \arg\min_{i \in \mathcal{V}} C_i$
30: **return** $\mathbb{H}_{i^*}$

In our empirical evaluations we used one of the following functions (Algorithms 15 or 16) to implement the NORMALISE function in steps 17-20; see Appendix E for more details:

[3] See Appendix D for definition.

Algorithm 15 Min-Max Normalisation

function MINMAXSCALE($\{s_i\}_{i\in\mathcal{V}}$)
  $s_{\min} \leftarrow \min_{j\in\mathcal{V}} s_j$
  $s_{\max} \leftarrow \max_{j\in\mathcal{V}} s_j$
  for $i \in \mathcal{V}$ **do**
    if $s_{\max} = s_{\min}$ **then**
      $\bar{s}_i \leftarrow 0.0$ ▷ Prevent division by zero
    else
      $\bar{s}_i \leftarrow \frac{s_i - s_{\min}}{s_{\max} - s_{\min}}$ ▷ Map to $[0, 1]$
    end if
  end for
  return $\{\bar{s}_i\}_{i\in\mathcal{V}}$
end function

Algorithm 16 Percentile Rank Normalisation

function RANKSCALE($\{s_i\}_{i\in\mathcal{V}}$)
  $N_{\mathcal{V}} \leftarrow |\mathcal{V}|$ ▷ Number of valid configurations
  for $i \in \mathcal{V}$ **do**
    $B_i \leftarrow |\{j \in \mathcal{V} \mid s_j < s_i\}|$ ▷ Count of strictly smaller elements
    $E_i \leftarrow |\{j \in \mathcal{V} \mid s_j = s_i\}|$ ▷ Count of equal elements (including s_i itself)
    $r_i \leftarrow B_i + \frac{E_i+1}{2}$ ▷ Compute average rank for ties
    $\bar{s}_i \leftarrow \frac{r_i}{N_{\mathcal{V}}}$ ▷ Map to $(0, 1]$
  end for
  return $\{\bar{s}_i\}_{i\in\mathcal{V}}$
end function

D DEFINITION OF ANEES

ANEES (Average Normalised Estimation Error Squared) measures the statistical consistency (Bar-Shalom et al., 2001, Chapter 5.4.) of a filter at a given time step t. Given N independent Monte-Carlo realisations of the state process (or N ground truth states) at time t, it can be empirically computed as:

$$\begin{aligned} \text{ANEES}(t) &:= \frac{1}{N}\sum_{n=1}^{N} \text{NEES}(t,n), \text{ where} \\ \text{NEES}(t,n) &:= (x_t^{(n)} - \mu_{t|t}^{(n)})^\top P_{t|t}^{(n)\,-1}(x_t^{(n)} - \mu_{t|t}^{(n)}) \end{aligned} \tag{7}$$

where $x_t^{(n)}$ is the state, and $\mu_{t|t}^{(n)}$ and $P_{t|t}^{(n)}$ are the filter's *a posteriori* mean and covariance, respectively, for the n-th realisation at time t.

A consistent filter yields ANEES $\approx s$, the state dimension. To see this, consider a filter for which the estimation error $e_t^{(n)} := x_t^{(n)} - \mu_{t|t}^{(n)}$ is zero-mean with covariance $P_{t|t}^{(n)}$, i.e., $\mathbb{E}[e_t^{(n)}] = 0$ and $\mathbb{E}[e_t^{(n)}(e_t^{(n)})^\top] = P_{t|t}^{(n)}$. The expected NEES is then:

$$\begin{aligned} \mathbb{E}\left[\text{NEES}(t,n)\right] &= \mathbb{E}\left[(e_t^{(n)})^\top P_{t|t}^{(n)\,-1} e_t^{(n)}\right] \\ &= \mathbb{E}\left[\text{Tr}\left(P_{t|t}^{(n)\,-1} e_t^{(n)}(e_t^{(n)})^\top\right)\right] \\ &= \text{Tr}\left(P_{t|t}^{(n)\,-1}\,\mathbb{E}[e_t^{(n)}(e_t^{(n)})^\top]\right) \\ &= \text{Tr}\left(P_{t|t}^{(n)\,-1} P_{t|t}^{(n)}\right) = \text{Tr}(I_s) = s \end{aligned} \tag{8}$$

Cross-Seed Aggregation. For the sake of presenting a statistical consistency metric alongside the time- and trajectory-averaged MSE and NLL metrics, we define the time-averaged ANEES for a filter trained under a specific random initialisation (seed), r. Because our evaluation trajectories span a large number of time steps T, the time-average serves as a reliable estimator for that specific trained filter; we note that in our empirical evaluations $T = 100$ in §4.2, 1000 in §4.3 and 800 in §4.4 in the test folds:

$$\text{ANEES}^{(r)} := \frac{1}{T}\sum_{t=1}^{T}\text{ANEES}(t)^{(r)} \tag{9}$$

By the linearity of expectation, this time-averaged metric is an unbiased estimator of the theoretical consistency target s:

$$\mathbb{E}\left[\text{ANEES}^{(r)}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\text{ANEES}(t)^{(r)}\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\text{ANEES}(t)^{(r)}] = \frac{1}{T}\sum_{t=1}^{T}s = s \tag{10}$$

Thus, ANEES deviation ('ANEES Dev.'), which we define as the deviation of $\text{ANEES}^{(r)}$ from the state dimension s, is a measure of statistical consistency for a filter with initialisation from a given seed r. However, considering computational costs, we evaluate filter performance over $R = 10$ seeds, with $R \ll T$. Hence, reporting the median and interquartile range of raw ANEES deviations could present apparently false consistency: specifically seeds with negative ANEES Dev. could compensate for those with positive ANEES Dev., resulting in ANEES Dev. ≈ 0. We therefore report the median and interquartile range over $R = 10$ seeds of the absolute value |ANEES Dev.| for trained filters.

E Experimental Details

E.1 Common Settings

E.1.1 Training Setup

Across experiments, all trainable models were trained using the AdamW optimiser. The following mechanics were applied uniformly across all models and experiments:

- **Differential Learning Rates:** Neural backbone parameters were optimized using a base learning rate (lr) and weight decay (wd). All other non-neural (structural) parameters (e.g., initial states, noise covariances) utilized a learning rate of $10 \times lr$ and zero weight decay to account for differing gradient dynamics.
- **Optimiser Hyperparameter Search Space:** Configurations were sampled via random search from log-uniform distributions: $lr \in [10^{-5}, 5 \times 10^{-2}]$ and $wd \in [10^{-5}, 10^{-1}]$.
- **Gradient Clipping:** A maximum L_2 norm of 1.0 was enforced to prevent gradient explosion.
- **Optional Warmup:** Probabilistic models optionally included a 20-epoch warmup phase (included in the maximum training epochs budget), training exclusively on (self-supervised) measurement MSE before transitioning to the negative log-likelihood (NLL) objective in Eq. (11).
- **Maximum Training Epochs Budget:** 100 epochs. *Exception:* APBM in Experiment 1 was truncated to 20 epochs due to severe computational bottlenecks caused by its prior-mean bisection algorithm (12–17 hours for 20 APBM epochs vs. ≤ 6.5 hours for 100 ResEKF epochs).
- **Early Stopping:** Training was halted if the monitored validation metric failed to improve by $\Delta_{\min}$ over a 15-epoch patience window. Early stopping was strictly disabled during the warmup phase.
- **Checkpoint Monitoring:** Following our validation procedure (Appendix C), probabilistic models monitored validation state NLL, while deterministic models (e.g., KalmanNet) monitored validation state MSE. The improvement threshold $\Delta_{\min}$ was set to 5×10^{-3}.

The (post-warmup) training objective for all probabilistic models except ResRORF was:

$$\mathcal{L}(\theta) = \frac{1}{2}\sum_{t=1}^{T}\Big(\log\det(S_t) + (y_t - \hat{y}_t)^\top S_t^{-1}(y_t - \hat{y}_t)\Big) + \text{const.} \tag{11}$$

where $y_t, \hat{y}_t$ and S_t are respectively the measurement, predicted measurement and (predicted) innovation covariance at time t. ResRORF was trained on variational evidence lower bound maximisation computed natively by RORF (Piché et al., 2012).

E.1.2 Validation Procedure

Across experiments, the validation procedure for all trainable models followed Appendix C. Hyperparameter search strategy was random search, conducted using Weights and Biases (Biewald, 2020) sweeps; across experiments and models, the run budget N_{trials} for the afore-mentioned random search was given by:

$$N_{\text{trials}} = \begin{cases} \max(30, \max(n_{cc}, 1) \cdot k \cdot 10) & \text{if } n_{cc} \leq 4 \\ \max(30, (n_{cat} + k) \cdot 20) & \text{if } n_{cc} > 4 \end{cases}$$

where n_{cat} denotes the number of categorical hyperparameters, n_{cc} denotes the total number of valid combinatorial configurations for those categorical variables (i.e., the product of their cardinalities), and k the number of continuous hyperparameters. All models had a minimum budget of 30 runs. In the linear benchmark (§4.2) and Lorenz attractor (§4.3) studies the normalisation function used alongside Appendix C was the MinMaxScale (Algorithm 15); in the NCLT experiment where validation metrics were more skewed (for all models), the RankScale function was used (Algorithm 16). In both cases, all models evaluated used the same validation procedure for model selection.

E.1.3 Square-Root Matrix Operations

All covariance updates across all models and experiments are performed in square-root (Cholesky) form (Kailath et al., 2000, Chapter 12.3), ensuring positive definiteness and improved numerical precision — benefiting all baselines regardless of whether they natively employ such parameterisations.

E.1.4 Baseline Model Constructions

To ensure a rigorous and fair comparison, we adapted the original authors' implementations of all baseline models to support the autonomous, continuous filtering environments required by our benchmarks.

Universal Adaptations. The following modifications were applied across the baseline suite where applicable:

- **Optional Statefulness:** All models were modified to preserve hidden states, recurrent contexts, or sliding windows across batch and train/val/test fold boundaries. Applies only to the NCLT experiment (§4.4).
- **Parameter Learning:** All relevant baselines were augmented to jointly learn initial states (μ_0, P_0), initial RNN hidden states, and static noise covariances (Q, Σ).
- **Strict Causality:** Sequence models (Mamba, PatchTST) were explicitly padded and shifted to guarantee that predictions at time t rely strictly on observations $y_{1:t-1}$.

Model-Specific Details.

- **APBM:** Follows the official implementation, utilising 10 iterations of the bisection algorithm to enforce the prior-mean epsilon constraint.

- **DANSE Ghosh et al. (2024) & VSE Norén et al. (2026):** GRU variants used, since the authors themselves use GRU in the original papers.
- **KalmanNet Revach et al. (2022) (KNet), RecursiveKalmanNet Mortada et al. (2025) (RKNet), & KalmanFormer (Shen et al., 2025):** To mitigate immediate divergence in chaotic regimes (e.g., the Lorenz attractor), thse gain-learning methods were patched with a "safe start" protocol, forcing the neural network outputs to exactly 0 at $t = 0$.
- **Mamba-3 Lahoti et al. (2026):** Setup using a SISO configuration with `chunk_size=32`, `d_state=16` and `headdim=64`. The network yields a $2s$-dimensional output, which is explicitly decomposed into the filtered state mean (dimension s) and a state covariance factor (dimension s). These latent state statistics are subsequently mapped to the observation space (yielding predicted observation means and innovation covariances) using the nominal measurement matrix and a custom linear layer utilising differentiable QR decomposition.
- **PatchTST Nie et al. (2023):** Configured with a causal sliding window. To bypass Flash Attention limits and prevent out-of-memory (OOM) errors on long continuous sequences, we implemented a dynamic binary-search automated chunking mechanism (max chunk ≤ 65535) alongside manual gradient checkpointing. Its $2s$-dimensional sequence output is processed identically to Mamba-3, being decomposed into latent state moments before being analytically mapped to the observation space.
- **AREKF (Xiong et al., 2009):** Adaptive Robust EKF; a recursively forgetting-factor-weighted estimate of the innovation covariance is used to threshold and downweight outlying measurements.
- **IS-EKF (Fang et al., 2022):** Innovation-Saturated EKF; maintains two coupled recursive trackers, a saturation bound and an innovation-energy estimate, that jointly bound each innovation's influence on the measurement update.
- **RORF (Piché et al., 2012):** RORF's adaptive measurement-noise inflation factor is capped at a worst-case value of 400 to guarantee stability (see Appendix J.1).
- **SIN (Krishnan et al., 2017):** No special adaptations made; the RNN-type we used for evaluations was GRU.
- **WoLF (Duran-Martin et al., 2024):** WoLF's per-step weight was floored at $w_{\min} = 0.05$ to guarantee stability (see Appendix J.1).

E.1.5 Parameter Initialisation

Across all experiments, all hybrid models received initial noise parameters (Q and Σ) based on pretuning a nominal (Extended) Kalman Filter to a stable configuration, pursuant to Remark 7. Furthermore, across all experiments and for all models, all remaining non-neural learnable parameters were initialised from $\mathcal{N}(0, I)$, with one exception—the initial covariance corresponding to APBM's neural parameters were drawn from $\mathrm{Wishart}\Big(\mathrm{dof} = n_\theta,\ \Sigma_0 = \frac{1}{n_\theta}\mathbf{I}\Big)$, where n_θ is the number of neural parameters. This exception was made so that in expectation, the resulting covariance would have identity covariance, matching other the spirit of other initialisations.

E.1.6 Random Seeds for Training and Evaluation

Seeds 37-46 were used for evaluation, with 42 being the seed used for model selection (Appendix C).

E.1.7 Time-Invariant Noise Covariances

In all experiments noise covariances Q_t and Σ_t were constant in time; to be clear, where they could be learned, their values would change between training batches, but remain constant in time.

E.2 Linear Benchmark Experiment

E.2.1 Dataset

- **True Dynamics:** The true system tracks a 6-dimensional state comprising 2D position, velocity, and damped acceleration: $x_t = [p_x, \dot{p}_x, \ddot{p}_x, p_y, \dot{p}_y, \ddot{p}_y]^\top$. The discrete-time true transition matrix

F_{true} and measurement matrix H_{true} operating at $\Delta t = 1.0$ are defined as:

$$F_{\text{true}} = \begin{bmatrix} C & 0_{3\times3} \\ 0_{3\times3} & C \end{bmatrix}, \quad C = \begin{bmatrix} 1 & \Delta t - c\frac{\Delta t^2}{2} & \frac{\Delta t^2}{2} \\ 0 & 1 - c\Delta t + (c^2 - \tau)\frac{\Delta t^2}{2} & \Delta t - c\frac{\Delta t^2}{2} \\ 0 & -\tau\Delta t + \tau c\frac{\Delta t^2}{2} & 1 - \tau\frac{\Delta t^2}{2} \end{bmatrix}$$

$$H_{\text{true}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

where $c = 0.06$ and $\tau = 0.17$.

- **Noise Profile:** The true process and measurement noises are zero-mean Gaussian, drawn from true covariances Q_{true} and Σ_{true}:

$$Q_{\text{true}} = 0.1^2 \cdot \text{diag}\left(\frac{\Delta t}{3}, \Delta t, 3\Delta t, \frac{\Delta t}{3}, \Delta t, 3\Delta t\right), \quad \Sigma_{\text{true}} = 0.5^2 I_2$$

- **Nominal Misspecification:** The nominal model assumes a 4-dimensional constant-velocity state $x_t = [p_x, \dot{p}_x, p_y, \dot{p}_y]^\top$, resulting in a severe dimensional mismatch. Its transition and measurement matrices are:

$$F_{\text{nom}} = \begin{bmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad H_{\text{nom}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

- **Generation & Splits:** 32,000 independent and identically distributed (I.I.D.) sequences were simulated from an initial state $x_0 = [0, 0.1, 0, 0, 0.1, 0]^\top$. Each sequence comprises 100 timesteps (after discarding a 50-step burn-in). The dataset was partitioned into 22,400 training, 6,400 validation, and 3,200 test sequences (70-20-10 split).

E.2.2 BRC Backbone Architecture

Backbone inputs at each time step t were the current model-based state estimates $\mathcal{I}_{t|t-1} = [\mu_{\text{KF},t|t-1}^\top \mid \text{vech}(P_{\text{KF},t|t-1})^\top]^\top$, where $\text{vech}(P)$ is the vectorised lower-triangle of P. Backbone output parameterisation follows Algorithm 1.

E.2.3 Parameter Initialisation

Hybrid models were initialised with $Q = 0.54I$, $\Sigma = 0.08I$, $P_{0|0} = 0.03I$ and $\mu_{0|0} = [\text{meas}_x, 0, \text{meas}_y, 0]$ (first measured position with 0 velocities); Biased KF alone received oracle $\mu_{0|0}$ as mentioned in §4.2.

E.2.4 Hyperparameters

Biased KF.

Hyperparameter	Search space	Selected value
Process Noise Scale	$[10^{-3},\ 2]$ (log-uniform)	0.91674
Measurement Noise Scale	$[10^{-3},\ 2]$ (log-uniform)	0.29225
Initial State Noise Scale	$[10^{-3},\ 2]$ (log-uniform)	0.00246

Ablation: ResEKF minus NN. (Note: In the code repository, this model must be invoked as 'okf', since a KF with learnable covariances was presented as Optimised Kalman Filter (OKF) (Greenberg et al., 2023). Our ablation differs from OKF in that it admits learnable initial states as well.)

Hyperparameter	Search space	Selected value
Learning Rate	$\{1, 5\} \times 10^{\{-5,\ldots,-2\}} \cup \{10^{-1}\}$ (deterministic grid search)	0.1
Loss warmup	$\{\text{yes},\ \text{no}\}$	no

ResEKF.

Hyperparameter	Search space	Selected value
$c_{\max}$	$[10^{-5},\ 3]$ (log-uniform)	0.48653
Loss warmup	{yes, no}	yes
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$4.09{\times}10^{-4}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	$3.63{\times}10^{-3}$

DANSE.

Hyperparameter	Search space	Selected value
Dense layer width	$\{10, \ldots, 40\}$ (integer uniform)	17
Loss warmup	{yes, no}	yes
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	0.04121
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	0.06360
Hidden dimension (h)	16 (fixed)	16

VSE.

Hyperparameter	Search space	Selected value
Dense layer width	$\{10, \ldots, 40\}$ (integer uniform)	35
Loss warmup	{yes, no}	yes
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$2.88{\times}10^{-3}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	$2.11{\times}10^{-3}$
Hidden dimension (h)	16 (fixed)	16

Mamba-3.

Hyperparameter	Search space	Selected value
Hidden dimension (h)	$\{32,\ 64\}$	64
Loss warmup	{yes, no}	no
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$2.92{\times}10^{-3}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	0.02621

PatchTST.

Hyperparameter	Search space	Selected value
Patch length (P)	$\{4,\ 8\}$	4
Model dimension (d)	$\{16,\ 32\}$	32
Loss warmup	{yes, no}	yes
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$2.23{\times}10^{-3}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	0.03066

RKNet.

Hyperparameter	Search space	Selected value
Input dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	37
Output dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	1
Loss warmup	$\{\text{yes}, \text{no}\}$	no
Learning rate	$[10^{-5}, 5{\times}10^{-2}]$ (log-uniform)	$3.74{\times}10^{-5}$
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	$3.03{\times}10^{-4}$

KNet.

Hyperparameter	Search space	Selected value
Input dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	39
Output dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	13
Learning rate	$[10^{-5}, 5{\times}10^{-2}]$ (log-uniform)	$1.14{\times}10^{-4}$
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	$4.63{\times}10^{-5}$

APBM

Hyperparameter	Search space	Selected value
Constraint mode	$\{\text{SSA}, \text{SSR}\}$	SSA
Threshold	$[10^{-3}, 20]$ (log-uniform)	0.02163
Loss warmup	$\{\text{yes}, \text{no}\}$	no
Learning rate	$[10^{-5}, 5{\times}10^{-2}]$ (log-uniform)	0.02411

SIN.

Hyperparameter	Search space	Selected value
Inference mode	$\{\text{mean} - \text{field}, \text{structured}\}$	mean-field
Emission mode	$\{\text{known}, \text{learned}(\text{mlp})\}$	known
Emission hidden dim.	$\{10, \ldots, 40\}$ (integer uniform)	25
Transition dim.	$\{10, \ldots, 40\}$ (integer uniform)	35
Loss warmup	$\{\text{yes}, \text{no}\}$	yes
Learning rate	$[10^{-5}, 5{\times}10^{-2}]$ (log-uniform)	0.01559
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	$1.20{\times}10^{-5}$

E.2.5 Diverged Seeds

- **APBM**: $39, 44, 46$
- **DANSE**: $37, 44$

E.3 Lorenz Attractor Experiment

E.3.1 Dataset

- **Continuous Dynamics:** The true latent trajectory governs a chaotic 3-dimensional system. A base trajectory was generated by numerically integrating the continuous-time Lorenz equations using the `solve_ivp` solver:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = \sigma(x_2 - x_1)$$

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = x_1(\rho - x_3) - x_2$$

$$\frac{\mathrm{d}x_3}{\mathrm{d}t} = x_1 x_2 - \beta x_3$$

where parameters were set to the standard values, $\sigma = 10.0, \rho = 28.0$, and $\beta = 8/3$. The continuous trajectory was sampled at a frequency yielding $\Delta t = 0.02$.

- **In-Distribution Dataset Construction:** A single 1,000-step window (extracted after a 1,000-step burn-in from an initial state $x_0 = [1.0, 1.0, 1.0]^\top$) was cloned to create $N = 300$ identical base sequences.
- **Heavy-Tailed Noise Application:** I.I.D. Gaussian Mixture Model (GMM) measurement noise, as described in §4.3. This creates 300 I.I.D. measurement sequences mapping to identical latent physical trajectories.
- **Splits:** The 300 sequences were split using a 70/20/10 ratio, yielding 210 training, 60 validation, and 30 test sequences.

E.3.2 BRC Backbone Architecture

Backbone inputs were the current model-based state estimates $\mathcal{I}_{t|\cdot} = [\mu_{\text{KF},t|\cdot}^\top \mid \text{vech}(P_{\text{KF},t|\cdot})^\top]^\top$, where $\text{vech}(P)$ is the vectorised lower-triangle of P. Backbone output parameterisation follows Algorithm 1.

E.3.3 Parameter Initialisation

Hybrid models were initialised with $Q = 0.1I$, $\Sigma = 8.8I$, $P_{0|0} = 0$ (perfect initial knowledge) and $\mu_{0|0}$ = true first state; note that initialising $P_{0|0} = 0$ technically violates Assumption 5 required by Theorem 1. By choosing $P_{0|0} = 0$ and demonstrating stable performance, we empirically validate Remark 4 as well.

E.3.4 Hyperparameters

EKF.

Hyperparameter	Search space	Selected value
Process Noise Scale	$[10^{-3},\ 10]$ (log-uniform)	0.02284
Measurement Noise Scale	$[10^{-3},\ 10]$ (log-uniform)	0.69888

DANSE.

Hyperparameter	Search space	Selected value
Dense layer width	$\{10, \ldots, 40\}$ (integer uniform)	25
Loss warmup	{yes, no}	yes
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$2.52{\times}10^{-3}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	$5.67{\times}10^{-3}$
Hidden dimension (h)	16 (fixed)	16

VSE.

Hyperparameter	Search space	Selected value
Dense layer width	$\{10, \ldots, 40\}$ (integer uniform)	40
Loss warmup	{yes, no}	yes
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$1.18{\times}10^{-3}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	0.0204
Hidden dimension (h)	16 (fixed)	16

Mamba-3.

Hyperparameter	Search space	Selected value
Hidden dimension (h)	$\{32,\ 64\}$	64
Loss warmup	$\{\text{yes},\ \text{no}\}$	yes
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$1.58{\times}10^{-3}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	0.0468

PatchTST.

Hyperparameter	Search space	Selected value
Patch length (P)	$\{4,\ 8\}$	4
Model dimension (d)	$\{16,\ 32\}$	32
Loss warmup	$\{\text{yes},\ \text{no}\}$	no
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	0.0238
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	$1.20{\times}10^{-5}$

RKNet.

Hyperparameter	Search space	Selected value
Input dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	9
Output dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	13
Loss warmup	$\{\text{yes},\ \text{no}\}$	no
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	0.0117
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	$1.01{\times}10^{-4}$

KNet.

Hyperparameter	Search space	Selected value
Input dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	18
Output dim. multiplier	$\{1, \ldots, 40\}$ (integer uniform)	16
Learning rate	$[10^{-5},\ 5{\times}10^{-2}]$ (log-uniform)	$8.62{\times}10^{-3}$
Weight decay	$[10^{-5},\ 10^{-1}]$ (log-uniform)	$1.90{\times}10^{-4}$

ResRORF.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	yes
$c_{\max}$	log-uniform $[0.1, 10]$	3.55
Number of heads	{2, 4, 8}	4
Model dimension	{16, 32, 64}	32
Learning rate	log-uniform $[10^{-5}, 5 \times 10^{-2}]$	5.17×10^{-3}
Weight decay	log-uniform $[10^{-5}, 10^{-1}]$	8.08×10^{-3}
ν	log-uniform $[1, 30]$	13.43

SIN.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	yes
Hidden dimension	fixed	16
Inference mode	{mean-field, structured}	mean-field
Emission mode	{known, learned (mlp)}	known
Emission hidden dim.	int-uniform $\{10, \ldots, 40\}$	12
Transition dimension	int-uniform $\{10, \ldots, 40\}$	24
Learning rate	log-uniform $[10^{-5}, 5 \times 10^{-2}]$	1.54×10^{-2}
Weight decay	log-uniform $[10^{-5}, 10^{-1}]$	5.97×10^{-3}

Ablation: ResRORF minus NN (RORF-OKF)

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	no
Learning rate	log-uniform $[10^{-5}, 5 \times 10^{-2}]$	3.32×10^{-2}
ν	log-uniform $[1, 30]$	8.71

KalmanFormer.

Hyperparameter	Search space	Selected value
Learning rate	log-uniform $[10^{-5}, 5 \times 10^{-2}]$	1.44×10^{-2}
Weight decay	log-uniform $[10^{-5}, 10^{-1}]$	9.03×10^{-3}

APBM.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	no
Constraint mode	{SSA, SSR}	SSA
Threshold	log-uniform $[10^{-3}, 20]$	0.173
Learning rate	log-uniform $[10^{-5}, 5 \times 10^{-2}]$	1.87×10^{-2}

RKNet-Former.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	No selectable model
Number of heads	{2, 4, 8}	No selectable model
Model dimension	{16, 32, 64}	No selectable model
Learning rate	log-uniform $[10^{-5}, 5 \times 10^{-2}]$	No selectable model
Weight decay	log-uniform $[10^{-5}, 10^{-1}]$	No selectable model

CKF.

Hyperparameter	Search space	Selected value
Process noise scale	log-uniform $[10^{-3}, 10]$	0.0645
Measurement noise scale	log-uniform $[10^{-3}, 10]$	2.77

AREKF.

Hyperparameter	Search space	Selected value
Process noise scale	log-uniform $[10^{-3}, 10]$	1.54
Measurement noise scale	log-uniform $[10^{-3}, 10]$	5.02
α	log-uniform $[0.01, 5]$	0.0141
ρ	uniform $[0.8, 0.999]$	0.977

IS-EKF.

Hyperparameter	Search space	Selected value
Process noise scale	log-uniform $[10^{-3}, 10]$	0.0168
Measurement noise scale	log-uniform $[10^{-3}, 10]$	0.0747
$\lambda_1^{(0,1,2)}$	uniform $[0.01, 0.99]^3$	(0.779, 0.710, 0.0781)
$\lambda_2^{(0,1,2)}$	uniform $[0.01, 0.99]^3$	(0.608, 0.387, 0.192)
$\gamma_1^{(0,1,2)}$	log-uniform $[10^{-3}, 100]^3$	(1.28, 36.1, 5.99)
$\gamma_2^{(0,1,2)}$	log-uniform $[0.1, 9.99]^3$	(0.959, 0.255, 3.35)

RORF.

Hyperparameter	Search space	Selected value
Process noise scale	log-uniform $[10^{-3}, 10]$	0.396
Measurement noise scale	log-uniform $[10^{-3}, 10]$	1.01
ν	log-uniform $[1, 30]$	3.05

WoLF.

Hyperparameter	Search space	Selected value
Process noise scale	log-uniform $[10^{-3}, 10]$	0.148
Measurement noise scale	log-uniform $[10^{-3}, 10]$	0.103
Weighting Function	{imq, md}	imq
c	log-uniform $[0.1, 20]$	1.31

E.3.5 Diverged Seeds and Performance on Non-Diverged Seeds

- **KNet**: 37
- **RKNet**: 40, 43
- **KFormer**: 37, 38, 40, 41, 43, 45, 46

E.4 NCLT Dataset Experiment

E.4.1 Dataset

- **Preprocessing:** A candidate measurements trajectory was selected from the data; it was the longest contiguous segment of GPS data that satisfied two constraints: no gap between consecutive fixes exceeded 5 seconds, and no continuous run of degraded measurements (mode $\neq 3$) lasted longer than 5 seconds. Ground truth and GPS fixes were interpolated onto a uniform 5 Hz grid, with ground-truth heading unwrapped before interpolation and re-wrapped to $[-\pi, \pi]$ afterwards to avoid the $\pm\pi$ discontinuity. Wheel-speed and gyro-heading control inputs were instead assigned via strictly-causal nearest-past-sample lookup rather than interpolation, so no control value reflects information from after its own time-step.
- **Validation Noise Calibration:** The raw validation-fold GPS was substantially cleaner than training in the x-dimension, which would distort model selection (Appendix C) for all models. We therefore fit an AR(1) model to the training x-residual (lag-1 autocorrelation $\phi = 0.999$) and

added synthetically generated noise, calibrated so the pooled (p_x, p_y)-position MSE on validation matches training exactly.

- **Splits:** The selected 8,041-step segment was chronologically split 70/20/10 after dropping the 41 leading steps needed for even chunking, yielding 8,000 usable steps: a 5,600-step training fold (chunked into 28 sequences of 200 steps for stateful processing), a 1,600-step validation fold, and an 800-step test fold.
- **Nominal Model:** The nominal dead-reckoning and control input function $f(x, u)$ for the state vector (2D positions) $x_t = [p_x, p_y]^\top$ and control inputs (yaw and wheel speed) $u_t = [\chi_t, v_t]^\top$, operated at $\Delta t = 0.2$, and the measurement function $h(x_t)$ are given by:

$$f(x_t, u_t) = x_{t-1} + v_{t-1}\Delta t[\cos(\chi_t), \sin(\chi_t)]^\top$$

$$h(x_t) = x_t$$

E.4.2 BRC Backbone Architecture

The BRC backbone for the NCLT experiment was designed around problem-specific prior knowledge from Carlevaris-Bianco et al. (2016): like the GPS measurements, the control inputs come from physical sensors (wheel encoders and a gyroscope), so we expected the biases in speed v_t and heading χ_t to be key sources of error. We also reasoned that these biases were unlikely to be constant, since conditions such as wheel slip change during the course of the robot's drive, but are also unlikely to change rapidly. Rather than predicting an unstructured position residual, the backbone therefore estimates these two biases.

Specifically, a GRU head outputs $z_t = [z_t^\chi, z_t^v] \in \mathbb{R}^2$, which drives bias correction terms:

$$b_t^\chi = \lambda b_{t-1}^\chi + r_\chi \tanh(z_t^\chi), \qquad b_t^v = \lambda b_{t-1}^v + r_v \tanh(z_t^v),$$

where $\lambda \in (0, 1)$ is a forgetting factor and r_χ, r_v are rates limiting the per-step change in each bias; we set $\lambda = 0.999$, $r_\chi = 0.1^\circ$, $r_v = 0.01\,\mathrm{m/s}$, encoding our belief that sensor biases evolve slowly.

The GRU takes as input the previous innovation, $(\cos\chi_t, \sin\chi_t, v_t)$, finite-difference rates of χ_t and v_t, the diagonal of the posterior covariance factor, and the previous biases b_{t-1}^χ, b_{t-1}^v.

The biases are mapped to a position residual through the nominal transition function. The nominal model changes the robot's position by $\Delta p_t = v_t \Delta t\,[\cos\chi_t,\ \sin\chi_t]^\top$. If the true controls are $v_t + b_t^v$ and $\chi_t + b_t^\chi$, a first-order expansion (assuming small biases) of the true step gives the displacement error:

$$\tilde{\mu}_t = \frac{\partial \Delta p_t}{\partial v_t}\, b_t^v + \frac{\partial \Delta p_t}{\partial \chi_t}\, b_t^\chi = \Delta t\, b_t^v \begin{bmatrix}\cos\chi_t \\ \sin\chi_t\end{bmatrix} + v_t \Delta t\, b_t^\chi \begin{bmatrix}-\sin\chi_t \\ \cos\chi_t\end{bmatrix}.$$

We therefore set:

$$\mu_{\text{BRC},\ t|t-1} = \tilde{\mu}_t \frac{c_{\max}}{\|\tilde{\mu}_t\|_2} \tanh\left(\frac{\|\tilde{\mu}_t\|_2}{c_{\max}}\right)$$

so that $\|\mu_{\text{BRC},\ t|t-1}\|_2 \le c_{\max}$ holds. Lacking a similar inductive bias for the covariance, we leave it uncorrected; i.e., we set $P_{\text{BRC},\ t|t-1} \equiv 0$. Notice that this output parameterisation differs from Algorithm 1.

E.4.3 Parameter Initialisation

The noise scales of models were tuned and fixed rather than initialised and learned. Initial states were also fixed at $P_{0|0} = 0$ (perfect initial knowledge) and $\mu_{0|0}$ = true first state; note that, again, initialising $P_{0|0} = 0$ technically violates Assumption 5 required by Theorem 1.

E.4.4 HYPERPARAMETERS

APBM.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	yes
Constraint mode	{SSA, SSR}	SSA
Threshold	$[10^{-3}, 20]$ (log-uniform)	0.0612
Learning rate	$[10^{-5}, 5 \times 10^{-2}]$ (log-uniform)	5.48×10^{-4}
Noise ratio (σ_r/σ_q)	[0.005, 0.5] (log-uniform)	0.389
Noise product ($\sigma_r\sigma_q$)	[15, 40] (uniform)	19.5

DANSE.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	no
Hidden dimension	fixed	16
Dense layer width	[10, 40] (int-uniform)	37
Learning rate	$[10^{-5}, 5 \times 10^{-2}]$ (log-uniform)	2.84×10^{-3}
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	2.12×10^{-3}

KF.

Hyperparameter	Search space	Selected value
Noise ratio (σ_r/σ_q)	[0.005, 0.5] (log-uniform)	0.41526
Noise product ($\sigma_r\sigma_q$)	[15, 40] (uniform)	18.9128

KNet.

Hyperparameter	Search space	Selected value
Input dim. multiplier	[1, 40] (int-uniform)	5
Output dim. multiplier	[1, 40] (int-uniform)	11
Learning rate	$[10^{-5}, 5 \times 10^{-2}]$ (log-uniform)	1.67×10^{-5}
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	5.77×10^{-5}
Noise ratio (σ_r/σ_q)	[0.005, 0.5] (log-uniform)	0.0693
Noise product ($\sigma_r\sigma_q$)	[15, 40] (uniform)	40.0

ResEKF.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	no
Hidden dimension	fixed	16
$c_{\max}$	[0.01, 1.0] (log-uniform)	0.0689
Learning rate	$[10^{-5}, 5 \times 10^{-2}]$ (log-uniform)	2.07×10^{-3}
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	0.0125
Noise ratio (σ_r/σ_q)	[0.005, 0.5] (log-uniform)	0.0146
Noise product ($\sigma_r\sigma_q$)	[15, 40] (uniform)	17.8

RKNet.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	yes
Hidden dimension	fixed	10
Input dim. multiplier	[1, 40] (int-uniform)	9
Output dim. multiplier	[1, 40] (int-uniform)	11
Learning rate	$[10^{-5}, 5 \times 10^{-2}]$ (log-uniform)	6.71×10^{-5}
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	4.32×10^{-5}
Noise ratio (σ_r/σ_q)	[0.005, 0.5] (log-uniform)	0.0442
Noise product ($\sigma_r\sigma_q$)	[15, 40] (uniform)	24.6

SIN.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	no
Hidden dimension	fixed	16
Inference mode	{mean-field, structured}	mean-field
Emission mode	{known, learned (mlp)}	known
Emission hidden dim.	[10, 40] (int-uniform)	15
Transition dimension	[10, 40] (int-uniform)	32
Learning rate	$[10^{-5}, 5 \times 10^{-2}]$ (log-uniform)	0.0367
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	1.64×10^{-5}

VSE.

Hyperparameter	Search space	Selected value
Loss warmup	{yes, no}	yes
Hidden dimension	fixed	16
Dense layer width	[10, 40] (int-uniform)	10
Learning rate	$[10^{-5}, 5 \times 10^{-2}]$ (log-uniform)	0.0434
Weight decay	$[10^{-5}, 10^{-1}]$ (log-uniform)	4.04×10^{-5}

F Lorenz Attractor Experiment: Additional Evaluations

Table 4 reports additional in-distribution ($r = 16$) filter performances on the test fold of the Lorenz attractor study (§4.3). KalmanFormer (Shen et al., 2025) ('KFormer') is an implementation of KalmanNet (Revach et al., 2022) which replaces the latter's GRUs with transformers. Divergence was observed in 7/10 random seeds. RKNet-Former is a de novo adaptation of Recursive KalmanNet (Mortada et al., 2025), where we replaced the native GRUs with ResRORF's (see §4.3) own causal transformer in a fully tunable setting. Out of a 100-run budget for hyperparameter tuning, none of the models survived divergence. Hyperparameter grids used for the training the filters reported in Table 4 are also listed in Appendix E.3.

G Regularity Conditions

Assumption 1 ($\mathcal{C}^2$ and Non-Singularity)**.** *The functions f and h (Eq. (2)) are twice continuously differentiable on $\mathbb{R}^s$ and the Jacobian matrix $\frac{\partial f}{\partial x}$ at any point along state trajectories is non-singular.*

Assumption 2 (Nonlinear Observability Rank Condition)**.** *The nonlinear observability matrix $U(x) \in \mathbb{R}^{ds \times s}$ (defined below, with $f^k := \underbrace{f \circ f \circ \ldots f}_{k \text{ times}}$) has full column rank s for all $x \in \mathbb{R}^s$:*

$$U(x)^\top := \left[\left(\frac{\partial h}{\partial x}(x) \right)^\top, \left(\frac{\partial h}{\partial x}(f(x)) \frac{\partial f}{\partial x}(x) \right)^\top, \ldots, \left(\frac{\partial h}{\partial x}(f^{s-1}(x)) \frac{\partial f}{\partial x}(f^{s-2}(x)) \cdots \frac{\partial f}{\partial x}(x) \right)^\top \right] \quad (12)$$

Table 4: Additional in-distribution ($r = 16$) filter performances on the test fold of the Lorenz attractor study.

Filter	MSE ↓	NLL ↓	| ANEES Dev.| ↓
AREKF (Xiong et al., 2009)	3.56	6.26	1.92
CKF (Arasaratnam & Haykin, 2009)	1.62	6.09	7.21
IS-EKF (Fang et al., 2022)	0.53	24.52	51.46
PatchTST	$10.31_{(9.50,\ 10.63)}$	$> 10^3$	$> 10^3$
Mamba-3	$2.56_{(2.25,\ 2.63)}$	$> 10^3$	$> 10^3$
Diverged Models			
KFormer (Shen et al., 2025) (7)	$0.30_{(0.28,\ 0.32)}$	n.a.	n.a.
RKNet-Former (n.a.)	Diverged	Diverged	Diverged

Definition 1 (Deterministic Process Model and Solution)**.** *The process (transition) model specified via Eq. 2, has a corresponding deterministic process model where $v_t \equiv 0,\ t \geq 0$, yielding deterministic dynamics $\tilde{x}_t = f(\tilde{x}_{t-1})$; we refer to $\tilde{x}_t$ as the* deterministic solution*.*

Definition 2 (Point to Path Distance)**.** *The distance between a point $\tilde{z} \in \mathbb{R}^n$ and a path (i.e., a set of points) $\mathcal{P} \subset \mathbb{R}^n$ is defined as $\inf_{z \in \mathcal{P}} \|z - \tilde{z}\|_2$ for all $n \geq 1$.*

Assumption 3 (Bounded State Assumption)**.** *There exists a compact set $\mathcal{X} \subset \mathbb{R}^s$ and finite $\epsilon_v, \epsilon_{\mathcal{X}} > 0$ (both independent of t) such that with probability one, the following hold for all $t \geq 0$:*

1. *The sample paths of deterministic solutions $\tilde{x}_t$ reside within $\mathcal{X}$, and are therefore bounded,*
2. *$\mathcal{X}$ contains all points within distance $\epsilon_{\mathcal{X}}$ of the sample paths of x_t, and*
3. *The unmodelled trajectory drift ($\tilde{x}_t - x_t$) satisfies $\|\tilde{x}_t - x_t\|_2 \leq \epsilon_v$.*

Remark 1. *Assumption 3 implies that the process noise v_t and the state x_t, due to the bound on $\tilde{x}_t$, are also bounded. In particular, the bound on v_t implies that it cannot be Gaussian. However, this does not preclude it from having a valid covariance matrix.*

Definition 3 (Trajectory Drift Bound)**.** *We interchangably refer to the constant ϵ_v as the Trajectory Drift Bound or Trajectory Deviation Constant.*

Assumption 4 (White Noise with Bounded, Positive Definite Covariances)**.** *The noise variables v_t and w_t are zero-mean and independent, with covariances Q_t and Σ_t respectively, such that:*

$$\underline{q} I \preceq Q_t \preceq \bar{q} I \quad \text{and} \quad \underline{s} I \preceq \Sigma_t \preceq \bar{s} I \quad \forall\, t \geq 0 \tag{13}$$

where I is the identity matrix. $\underline{q}, \underline{s}, \bar{q}$ and $\bar{s}$ are strictly positive, finite scalars bounding the eigenvalues of the covariance matrices.

Assumption 5 (Initial Covariance)**.** *The initial state covariance, $P_{0|0}$ is positive definite.*

Assumption 6 (Model Fidelity)**.** *We explicate what Reif et al. (1999) implicitly assume: the model specified via Eq. (2) must hold.*[4]

H Exact Error Dynamics of the BRC-Augmented EKF

We start with the Taylor series expansion of the transition and measurement functions, f and h defined in Eq. (2), to obtain:

[4]We remark on the implications of this assumption in Remark 5.

$$f(x_t) - f(\mu_{t|t}) = F_{t+1}e_t + \varphi(x_t, \mu_{t|t}); \quad h(x_t) - h(\mu_{t|t-1}) = H_t e_{t|t-1} + \chi(x_t, \mu_{t|t-1}) \tag{14}$$

where F_{t+1} and H_t are respectively the Jacobian matrices of f and h evaluated at $\mu_{t|t}$ and $\mu_{t|t-1}$. φ and χ are functions representing the sum of non-linear terms of the Taylor series expansions of f and h respectively.

Now consider the *a priori* error, $e_{t|t-1} := x_t - \mu_{t|t-1}$:

$$\begin{aligned}
e_{t|t-1} &:= x_t - \mu_{t|t-1} \\
&= f(x_{t-1}) - \mu_{\text{KF},\, t|t-1} + v_t - \mu_{\text{BRC},\, t|t-1} && \text{[using Eqs. (1) and (2)]} \\
&= f(x_{t-1}) - f(\mu_{t-1|t-1}) + v_t - \mu_{\text{BRC},\, t|t-1} && \text{[using Algorithm 3]} \\
&= F_t e_{t-1} + \varphi(x_{t-1}, \mu_{t-1|t-1}) + v_t - \mu_{\text{BRC},\, t|t-1} && \text{[using Eq. (14)]}
\end{aligned} \tag{15}$$

We then consider the estimation error, $e_t := x_t - \mu_{t|t}$:

$$\begin{aligned}
e_t &:= x_t - \mu_{t|t} \\
&= x_t - \mu_{\text{KF},\, t|t} - \mu_{\text{BRC},\, t|t} && \text{[using Eq. (1)]} \\
&= x_t - \mu_{t|t-1} - K_t(y_t - h(\mu_{t|t-1})) - \mu_{\text{BRC},\, t|t} && \text{[using Algorithm 4]} \\
&= x_t - \mu_{t|t-1} - K_t(h(x_t) - h(\mu_{t|t-1})) - K_t w_t - \mu_{\text{BRC},\, t|t} && \text{[using Eq. (2)]} \\
&= e_{t|t-1} - K_t(h(x_t) - h(\mu_{t|t-1})) - K_t w_t - \mu_{\text{BRC},\, t|t} && [e_{t|t-1} := x_t - \mu_{t|t-1}] \\
&= (I - K_t H_t)e_{t|t-1} - K_t\chi(x_t, \mu_{t|t-1}) - K_t w_t - \mu_{\text{BRC},\, t|t} && \text{[using Eq. (14)]}
\end{aligned} \tag{16}$$

where I is the identity matrix. Now, substituting Eq. (15) in (16), we have:

$$\begin{aligned}
e_t = \underbrace{(I - K_t H_t)F_t}_{:=E_t} e_{t-1} + \underbrace{(I - K_t H_t)\varphi(x_{t-1}, \mu_{t-1|t-1}) - K_t\chi(x_t, \mu_{t|t-1})}_{:=s_t} \\
+ \underbrace{(I - K_t H_t)v_t - K_t w_t}_{:=r_t} \\
\underbrace{-(I - K_t H_t)\mu_{\text{BRC},\, t|t-1} - \mu_{\text{BRC},\, t|t}}_{:=u_t}
\end{aligned} \tag{17}$$

We restate the above result for clearer reference within this paper as:

$$e_t = E_t e_{t-1} + s_t + r_t + u_t, \text{ where:} \tag{18}$$

$$E_t := (I - K_t H_t)F_t \qquad \text{[Error Dynamics Matrix]} \tag{19a}$$

$$s_t := (I - K_t H_t)\varphi(x_{t-1}, \mu_{t-1|t-1}) - K_t\chi(x_t, \mu_{t|t-1}) \qquad \text{[Linearisation Residue]} \tag{19b}$$

$$r_t := (I - K_t H_t)v_t - K_t w_t \qquad \text{[Noise Term]} \tag{19c}$$

$$u_t := -(I - K_t H_t)\mu_{\text{BRC},\, t|t-1} - \mu_{\text{BRC},\, t|t} \qquad \text{[Neural Feedback Term]} \tag{19d}$$

I Proof of Theorem 1

This section is organised into as follows:

- In §I.1 we present a sketch of the proof.

- We then present the intermediaries (lemmas, propositions, definitions etc.) required in §I.2.
- All arguments are collected and presented as the 'final' proof in §I.3.
- We present remarks about Theorem 1 in §I.4.

We note the following:

- The sketch of our proof, intermediary results and the final proof refer to Appendix H, where we derive the dynamics of the BRC-augmented EKF's estimation errors.
- Lemmas that depend on Assumption 3 are to be understood as valid with probability one under the stated conditions.
- We interchangably refer to the BRC-augmented EKF as 'ResEKF' and the model specified in Eq. (2) as the 'nominal model'. By extension, specifications of the nominal model (e.g. state trajectory, transition function, etc.) are also called nominal.
- References to Algorithm 2 must, in this section, be seen as applying the algorithm with EKF's Time and Measurement Updates (Algorithms 3 and 4).
- **Notations**: In this section, $\|\cdot\|_2$ represents the Euclidean norm if $\cdot$ is a vector and the induced operator norm (spectral norm for matrices) otherwise. I represents the identity matrix.

I.1 Sketch of Proof

Sketch of Proof. The proof proceeds by induction. Assuming that the filter's estimation error remains bounded for time indices $1 \leq t < \tau$, we first establish uniform bounds on the nominal model's Jacobian matrices (F_t, H_t) and the estimated state covariances $(P_{t|t})$. We then define the stochastic Lyapunov candidate $V_t(e_t) := e_t^\top P_{t|t}^{-1} e_t$. Evaluating the expected forward difference $\mathbb{E}[V_t|e_{t-1}] - V_{t-1}$ along the exact ResEKF error dynamics (Eq. (18)), we obtain multiple cross terms between unforced error $E_t e_{t-1}$ and the total disturbance term $(r_t + s_t + u_t)$, where r_t arises from stochastic noise, s_t from higher-order Taylor remainder terms after EKF linearisation, and u_t from ResEKF's neural residuals—the novel term relative to Reif et al. (1999). We apply Young's inequality to decouple these cross terms and formulate a generalised supermartingale descent condition, which guarantees a bound on the estimation error at time τ with probability one. To satisfy this condition, we construct the bounds $\bar{\epsilon}$, $\bar{\epsilon}_v$, $\bar{\sigma}$ and $\bar{c}$ on e_0, ϵ_v, Σ_t and $c_{\max}$ respectively, thus establishing a stable region of attraction for the filter and closing the induction with probability one. Finally, unrolling the expected Lyapunov candidate over time yields the theorem's upper-bound on mean-square errors, with specific forms for the transient function β, which exponentially decays to 0, and the asymptotic floor of this upper-bound, composed of the functions γ and δ. □

I.2 Intermediaries

Hypothesis 1 (Estimation Error Boundedness at time t)**.** *There exists a finite, strictly positive scalar ϵ_{sup} such that ResEKF's estimation error e_t at any time t remains bounded as:*

$$\|e_t\|_2 \leq \epsilon_{sup} \tag{20}$$

Definition 4 (Operating Domains)**.** *Let $\mathcal{E} := \{e \in \mathbb{R}^s : \|e\|_2 \leq \epsilon_{sup}\}$, where ϵ_{sup} is as defined in Hypothesis 1. Let $\mathcal{B}_{c_{\max}} := \{v \in \mathbb{R}^s : \|v\|_2 \leq c_{\max}\}$, where $c_{\max}$ is ResEKF's norm-bound. Given a compact set $\Xi \subset \mathbb{R}^s$ and a continuous mapping $\phi : \mathbb{R}^s \to \mathbb{R}^s$, we define the following sets:*

- *The a posteriori domain:* $\Omega(\Xi) := \Xi \oplus \mathcal{E}$
- *The a priori domain:* $\Gamma(\Xi, \phi) := \phi(\Omega(\Xi)) \oplus \mathcal{B}_{c_{\max}}$
- *The operating domain:* $D(\Xi, \phi) := \Omega(\Xi) \cup \Gamma(\Xi, \phi)$

where $\oplus$ denotes the Minkowski sum and $\cup$ denotes the set union operators.

Lemma L.1 (Compactness of Domains)**.** *Let $\mathcal{X} \subset \mathbb{R}^s$ be the compact set defined in Assumption 3 and f be the nominal transition function. Under Assumptions 1 and 3, the sets $\Omega(\mathcal{X})$, $\Gamma(\mathcal{X}, f)$, and $D(\mathcal{X}, f)$ defined in Definition 4 are compact subsets of $\mathbb{R}^s$, whenever they exist.*

Proof. Assume that the sets $\Omega(\mathcal{X})$, $\Gamma(\mathcal{X}, f)$, and $D(\mathcal{X}, f)$ exist, as predicated by the lemma.

Now, by Assumption 3 and Definition 4, $\mathcal{X}$ and $\mathcal{E}$ are compact. Because the Minkowski sum of two compact sets is compact, $\Omega(\mathcal{X})$ is strictly compact.

Further, by Assumption 1, the nominal transition function f is continuous. Because the continuous image of a compact set is compact, $f(\Omega(\mathcal{X}))$ is compact. Since the ball $\mathcal{B}_{c_{\max}}$ is also compact, their Minkowski sum $\Gamma(\mathcal{X}, f)$ is compact.

Finally, because the union of a finite number of compact sets is compact, the operating domain $D(\mathcal{X}, f) := \Omega(\mathcal{X}) \cup \Gamma(\mathcal{X}, f)$ is strictly compact. □

Lemma L.2 (Universal Jacobian Bounds)**.** *Under Assumptions 1–3, there exist finite, strictly positive scalars $\bar{F}$ and $\bar{H}$ satisfying:*

$$\bar{F} := \max_{z \in D(\mathcal{X}, f)} \left\| \frac{\partial f}{\partial x}(z) \right\|_2, \qquad \bar{H} := \max_{z \in D(\mathcal{X}, f)} \left\| \frac{\partial h}{\partial x}(z) \right\|_2 \tag{21}$$

where $\mathcal{X}$ is the domain of nominal states x_t for all $t \geq 0$ (Assumption 3) and f and h are the nominal transition and measurement functions, provided the operating domain $D(\mathcal{X}, f)$ exists.

Proof. Assume that the operating domain $D(\mathcal{X}, f)$ exists, as predicated by the lemma.

By Lemma L.1, $D(\mathcal{X}, f)$ is a strictly compact set in $\mathbb{R}^s$. By Assumption 1, the functions f and h are continuously differentiable, ensuring that their respective Jacobian mappings $x \to \frac{\partial f}{\partial x}(x)$ and $x \to \frac{\partial h}{\partial x}(x)$ are continuous. Because the spectral norm is a continuous function of matrix elements, the composite mappings from $x \in \mathbb{R}^s$ to the corresponding spectral norms are continuous real-valued functions. By the Extreme Value Theorem, a continuous function evaluated over a compact set must attain a finite maximum. Thus, the universal bounds $\bar{F}$ and $\bar{H}$ exist and are finite.

We further show that $\bar{H}, \bar{F} > 0$ by contradiction. Suppose $\bar{H} = 0$; then the measurement Jacobian $\frac{\partial h}{\partial x}(z) \equiv 0$ for all $z \in D(\mathcal{X}, f)$. This would imply nonlinear observability matrix $U(z) \equiv 0$ (Eq. (12)), contradicting the assumption that it has full column rank (Assumption 2). Thus, there exists at least one $z \in D(\mathcal{X}, f)$ where the spectral norm of the measurement Jacobian matrix is non-zero, ensuring $\bar{H} > 0$. As for $\bar{F}$, Assumption 1 requires that the transition Jacobian matrix is non-singular at any evaluation point in the set $\mathcal{X}$; thus, there exists at least one $z \in D(\mathcal{X}, f)$ where the spectral norm of the transition Jacobian matrix is non-zero, ensuring $\bar{F} > 0$. □

Lemma L.3 (Containment of State Estimates)**.** *Let Hypothesis 1 hold for all $1 \leq t < \tau$. Under Assumptions 1 and 3, the a posteriori estimates $\mu_{t|t}$ and the a priori estimates $\mu_{t+1|t}$ generated by Algorithm 2 are bounded such that $\mu_{t|t} \in \Omega(\mathcal{X})$ and $\mu_{t+1|t} \in \Gamma(\mathcal{X}, f)$, where $\mathcal{X} \subset \mathbb{R}^s$ is as defined in Assumption 3 and f is the nominal transition function, for all $1 \leq t < \tau$.*

Proof. Under Hypothesis 1, ϵ_{sup} exists, and therefore, the sets $\Omega(\mathcal{X})$, $\Gamma(\mathcal{X}, f)$ and $D(\mathcal{X}, f)$ exist for all $1 \leq t < \tau$.

From the definition of the estimation error, we can express the *a posteriori* estimate as $\mu_{t|t} = x_t - e_t$. By Assumption 3, the nominal state satisfies $x_t \in \mathcal{X}$. By Hypothesis 1, the estimation error is bounded such that $e_t \in \mathcal{E}$ for all $t < \tau$. The properties of the Minkowski sum, therefore, imply that $\mu_{t|t} \in \mathcal{X} \oplus \mathcal{E} := \Omega(\mathcal{X})$ (Definition 4) for all $t < \tau$.

Next, consider the *a priori* estimate $\mu_{t+1|t} = f(\mu_{t|t}) + \mu_{\text{BRC},\ t+1|t}$ (from Algorithm 2). By Assumption 1, f is continuous; therefore, since $\mu_{t|t} \in \Omega(\mathcal{X})$, we have $f(\mu_{t|t}) \in f(\Omega(\mathcal{X}))$ for all $t < \tau$. By construction (Eq. (1), $\|\mu_{\text{BRC},\ t+1|t}\|_2 \leq c_{\max}$, implying that $\mu_{\text{BRC},\ t+1|t} \in \mathcal{B}_{c_{\max}}$, and further, that $\mu_{t+1|t} \in f(\Omega(\mathcal{X})) \oplus \mathcal{B}_{c_{\max}} := \Gamma(\mathcal{X}, f)$ (Definition 4) for all $t < \tau$. □

Lemma L.4 (ResEKF Jacobian Bounds)**.** *Let F_t and H_t be the Jacobian matrices evaluated in Algorithm 2. Under the conditions of Lemmas L.2 and L.3, the spectral norms of these matrices satisfy:*

$$\|F_t\|_2 \leq \bar{F}, \qquad \|H_t\|_2 \leq \bar{H} \qquad \forall\, 1 \leq t \leq \tau \tag{22}$$

Proof. By Lemma L.3, the ResEKF estimates $\mu_{t|t}$ and $\mu_{t+1|t}$ are contained within the compact set $D(\mathcal{X}, f)$, with $\mathcal{X}$ as defined in Assumption 3 and f as the nominal transition function, for all $t < \tau$. Since Lemma L.2 establishes $\bar{F}$ and $\bar{H}$ as the respective maxima of the spectral norms of the transition and measurement Jacobian matrices over the entire set $D(\mathcal{X}, f)$, the bounds in Eq. (22) hold at ResEKF's Jacobian evaluation points $(\mu_{0|0}, \ldots, \mu_{\tau-1|\tau-1})$ and $(\mu_{1|0}, \ldots, \mu_{\tau|\tau-1})$, i.e., for all Jacobians F_t and H_t in time steps $1 \leq t \leq \tau$. □

Remark 2 (Temporal Propagation of Boundedness)**.** *Note that Lemma L.4 only assumes inductive hypothesis (Hypothesis 1) validity for time indices $1 \leq t < \tau$. Because of ResEKF's recursive filtering loop (see Algorithm 2), we obtain boundedness of F_t and H_t for time indices $1 \leq t \leq \tau$, as we show in Lemma L.3 and its proof. A number of other such bounds are derived in subsequent lemmas for the interval $1 \leq t \leq \tau$ without assuming inductive hypothesis validity at $t = \tau$.*

Definition 5 (Deterministic Trajectory Mappings)**.** *Let $\tilde{x}_t$ be the deterministic solution trajectory defined in Definition 1 and f and h be the nominal transition and measurement functions. We define the corresponding sequence of deterministic Jacobian matrices evaluated strictly along this trajectory as:*

$$\tilde{F}_t := \frac{\partial f}{\partial x}(\tilde{x}_{t-1}), \qquad \tilde{H}_t := \frac{\partial h}{\partial x}(\tilde{x}_t) \qquad \forall\, t \geq 1 \tag{23}$$

For any $i \geq t$, the deterministic trajectory matrix $\tilde{\Phi}_{i,t}$ is defined by the product:

$$\tilde{\Phi}_{i,t} := \begin{cases} \prod_{j=t}^{i-1} \tilde{F}_j := \tilde{F}_{i-1}\tilde{F}_{i-2}\cdots\tilde{F}_t, & \text{when } i > t \\ I, & \text{when } i = t \end{cases} \tag{24}$$

The deterministic observability Gramian over a window $k \geq 0$ is then given by:

$$\tilde{M}_{t+k,t} := \sum_{i=t}^{t+k} (\tilde{H}_i \tilde{\Phi}_{i,t})^\top \tilde{H}_i \tilde{\Phi}_{i,t} \tag{25}$$

Lemma L.5 (Deterministic Jacobian Bounds)**.** *Let $\tilde{F}_t, \tilde{H}_t$ be the deterministic Jacobian matrices from Definition 5. Under the conditions of Lemma L.2, the spectral norms of these matrices satisfy:*

$$\|\tilde{F}_t\|_2 \leq \bar{F}, \qquad \|\tilde{H}_t\|_2 \leq \bar{H} \qquad \forall\, t \geq 1 \tag{26}$$

Proof. By Assumption 3, the deterministic solution trajectory $\tilde{x}_t$ remains within the compact set $\mathcal{X} \subset D(\mathcal{X}, f)$, with $\mathcal{X}$ as defined in Assumption 3 and f as the nominal transition function, for all $t \geq 1$. Again, since Lemma L.2 establishes $\bar{F}$ and $\bar{H}$ as the respective maxima of the spectral norms of the transition and measurement Jacobian matrices over the entire set $D(\mathcal{X}, f)$, the bounds hold for all Jacobian evaluation points along the deterministic solution trajectory. □

Lemma L.6 (Deterministic Observability)**.** *Let $\tilde{M}_{t+k,t}$ be the deterministic observability Gramian as defined in Definition 5. Under Assumptions 1–3, there exists a strictly positive, finite scalar $m_{det}(k) > 0$ such that:*

$$\tilde{M}_{t+k,t} \succeq m_{det}(k) I, \qquad \forall\, t \geq 1 \tag{27}$$

over any window $k \geq s - 1$.

Proof. By Assumption 2, the nonlinear observability matrix $U(\tilde{x}_t)$ has full column rank s for any $\tilde{x} \in \mathcal{X}$, implying that the product $U(\tilde{x}_t)^\top U(\tilde{x}_t)$ is strictly positive definite. Notice that the deterministic observability Gramian $\tilde{M}_{t+s-1,t}$ over window $k = s-1$ is itself the product $U(\tilde{x}_t)^\top U(\tilde{x}_t)$. Thus, we can define $m_{\det}(s-1)$ for the window $k = s-1$ as:

$$m_{\det}(s-1) := \min_{x \in \mathcal{X}} \lambda_{\min}(U(x)^\top U(x)) \tag{28}$$

where $\lambda(\cdot)$ is the eigenvalue of $\cdot$, satisfying Eq. (27), provided $\infty > m_{\det}(s-1) > 0$.

We proceed to show that under Assumption 1, $\infty > m_{\det}(s-1) > 0$ does in fact hold, for the window $k = s-1$. By the afore-mentioned assumption, the mappings $x \to \frac{\partial f}{\partial x}$ and $x \to \frac{\partial h}{\partial x}$ are continuous. Since the minimum eigenvalue of a matrix is a continuous function of its entries, the mapping $x \to \lambda_{\min}(U(x)^\top U(x))$ is continuous. By the Extreme Value Theorem, this function evaluated over the compact set $\mathcal{X}$ (Assumption 3) must attain a finite maximum and minimum. Since, as argued earlier, $U(x)^\top U(x)$ is positive definite, the latter has strictly positive eigenvalues for all $x \in \mathcal{X}$, ensuring that $m_{\det}(s-1)$ is in fact strictly positive and finite.

For any $k > s-1$, $\tilde{M}_{t+k,t}$ is formed by adding further positive semi-definite terms $(\tilde{H}_i \tilde{\Phi}_{i,t})^\top \tilde{H}_i \tilde{\Phi}_{i,t}$ where $i \geq s$ to $\tilde{M}_{t+s-1,t}$. Consequently, the minimum eigenvalue can only increase or remain constant, ensuring the bound $\tilde{M}_{t+k,t} \succeq m_{\text{det}}(k) I$ holds with $\infty > m_{\det}(k) > 0$ for all $t \geq 0$ and window $k \geq s-1$. □

Lemma L.7 (Jacobian Perturbation Bounds)**.** *Let F_t, H_t be the Jacobian matrices evaluated in Algorithm 2 and $\tilde{F}_t, \tilde{H}_t$ be the deterministic Jacobian matrices from Definition 5. Under the conditions of Lemmas L.1 and L.3, there exist finite strictly positive scalars $L_{\nabla f}$ and $L_{\nabla h}$ such that the following spectral norm bounds hold:*

$$\|F_t - \tilde{F}_t\|_2 \leq L_{\nabla f}(\epsilon_{\sup} + \epsilon_v), \qquad \|H_t - \tilde{H}_t\|_2 \leq L_{\nabla h}(\bar{F}(\epsilon_{\sup} + \epsilon_v) + c_{\max}) \qquad \forall\, 1 \leq t \leq \tau \tag{29}$$

where $\epsilon_{\sup}$ is the estimation error bound from Hypothesis 1, ϵ_v is the trajectory deviation constant from Assumption 3, $\bar{F}$ is the universal transition function Jacobian bound (Lemma L.2) and $c_{\max}$ is the ResEKF norm-bound.

Proof. By Assumption 1, the functions f and h are twice continuously differentiable on $\mathbb{R}^s$. Since the operating domain $D(\mathcal{X}, f)$ is compact (Lemma L.1), the Jacobian mappings $x \to \frac{\partial f}{\partial x}$ and $x \to \frac{\partial h}{\partial x}$ are Lipschitz continuous on $D(\mathcal{X}, f)$ with finite, strictly positive Lipschitz constants $L_{\nabla f}$ and $L_{\nabla h}$, respectively for all $1 \leq t \leq \tau$.

Moreover, Lemma L.3 grants that the evaluation points $\mu_{t-1|t-1}$ and $\tilde{x}_{t-1}$ both reside within $D(\mathcal{X}, f)$. Therefore, the Lipschitz continuity property of the continuously differentiable transition Jacobian matrix over $D(\mathcal{X}, f)$ yields:

$$\|F_t - \tilde{F}_t\|_2 = \left\| \frac{\partial f}{\partial x}(\mu_{t-1|t-1}) - \frac{\partial f}{\partial x}(\tilde{x}_{t-1}) \right\|_2 \leq L_{\nabla f} \|\mu_{t-1|t-1} - \tilde{x}_{t-1}\|_2 \qquad \forall\, 1 \leq t \leq \tau \tag{30}$$

Moreover, applying the triangle inequality along with Hypothesis 1 and Assumption 3 yields:

$$\begin{aligned} \|\mu_{t-1|t-1} - \tilde{x}_{t-1}\|_2 \leq\ & \underbrace{\|\mu_{t-1|t-1} - x_{t-1}\|_2}_{\text{Norm of Estimation Error, } \leq \epsilon_{\sup}} + \underbrace{\|x_{t-1} - \tilde{x}_{t-1}\|_2}_{\text{Norm of Trajectory Drift, } \leq \epsilon_v} \\ \iff \|\mu_{t-1|t-1} - \tilde{x}_{t-1}\|_2 \leq\ & \epsilon_{\sup} + \epsilon_v \qquad \forall\, 1 \leq t \leq \tau \end{aligned} \tag{31}$$

Applying this to the Eq. (30) confirms the bound for F_t.

Arguing similarly, using the containment of $\mu_{t|t-1} \in \Gamma(\mathcal{X}, f) \subseteq D(\mathcal{X}, f)$, we have:

$$\|H_t - \tilde{H}_t\|_2 \leq L_{\nabla h}\|\mu_{t|t-1} - \tilde{x}_t\|_2 \qquad \forall\, 1 \leq t \leq \tau \tag{32}$$

From Algorithm 2, we have $\mu_{t|t-1} = f(\mu_{t-1|t-1}) + \mu_{\text{BRC},\, t|t-1}$, with the neural residual $\|\mu_{\text{BRC},\, t|t-1}\|_2 \leq c_{\max}$ by construction (Eq. (1)). Applying Definition 1 to this, we have $\|\mu_{t|t-1} - \tilde{x}_t\|_2 = \|f(\mu_{t-1|t-1}) - f(\tilde{x}_{t-1}) + \mu_{\text{BRC},\, t|t-1}\|_2$. Applying this result and the triangle inequality to Eq. (32) yields:

$$\|H_t - \tilde{H}_t\|_2 \leq L_{\nabla h}(\|f(\mu_{t-1|t-1}) - f(\tilde{x}_{t-1})\|_2 + c_{\max}) \qquad \forall\, 1 \leq t \leq \tau \tag{33}$$

The Lipschitz continuity property of the twice continuously differentiable transition function over $D(\mathcal{X}, f)$ along with Lemma L.2 grants that $\|f(\mu_{t-1|t-1}) - f(\tilde{x}_{t-1})\|_2 \leq \bar{F}\|\mu_{t-1|t-1} - \tilde{x}_{t-1}\|_2$. Using this result and Eqs. (31) and (33) completes the proof. □

Proposition P.1. *Let $\{A_j\}$ and $\{B_j\}$ be two sequences of square matrices. Under the empty-product convention that $\prod_{j=t}^{t-1} \cdot \equiv I$, let Λ_ℓ and $\tilde{\Lambda}_\ell$ be defined such that for any window length $\ell \geq 0$:*

$$\Lambda_\ell := \prod_{m=t}^{t+\ell-1} A_m := A_{t+\ell-1}A_{t+\ell-2}\cdots A_t, \qquad \tilde{\Lambda}_\ell := \prod_{n=t}^{t+\ell-1} B_n := B_{t+\ell-1}B_{t+\ell-2}\cdots B_t \tag{34}$$

Under the empty-product convention above and the empty-sum convention, that $\sum_{j=t}^{t-1} \cdot \equiv 0$, the difference between the products Λ_ℓ and $\tilde{\Lambda}_\ell$ satisfies:

$$\Lambda_\ell - \tilde{\Lambda}_\ell = \sum_{j=t}^{t+\ell-1} \left(\prod_{m=j+1}^{t+\ell-1} A_m \right) (A_j - B_j)\tilde{\Lambda}_{j-t} \qquad \forall\, \ell \geq 0 \tag{35}$$

Proof. We proceed by induction on ℓ. In the *Base Case ($\ell = 0$)*, the left-hand side is $\Lambda_0 - \tilde{\Lambda}_0 = I - I = 0$. The right-hand side is an empty sum ($j = t$ to $t-1$), which is 0 by convention. Thus, the identity holds.

Inductive Step: Assume the identity holds for ℓ. For $\ell + 1$, we observe that $\Lambda_{\ell+1} = A_{t+\ell}\Lambda_\ell$ and $\tilde{\Lambda}_{\ell+1} = B_{t+\ell}\tilde{\Lambda}_\ell$. We expand the difference:

$$\Lambda_{\ell+1} - \tilde{\Lambda}_{\ell+1} = A_{t+\ell}\Lambda_\ell - B_{t+\ell}\tilde{\Lambda}_\ell = A_{t+\ell}(\Lambda_\ell - \tilde{\Lambda}_\ell) + (A_{t+\ell} - B_{t+\ell})\tilde{\Lambda}_\ell \tag{36}$$

Substituting the inductive hypothesis for $(\Lambda_\ell - \tilde{\Lambda}_\ell)$:

$$\begin{aligned}\Lambda_{\ell+1} - \tilde{\Lambda}_{\ell+1} &= A_{t+\ell}\left[\sum_{j=t}^{t+\ell-1} \left(\prod_{m=j+1}^{t+\ell-1} A_m \right) (A_j - B_j)\tilde{\Lambda}_{j-t} \right] + (A_{t+\ell} - B_{t+\ell})\tilde{\Lambda}_\ell \\ &= \sum_{j=t}^{t+\ell-1} \left(\prod_{m=j+1}^{t+\ell} A_m \right) (A_j - B_j)\tilde{\Lambda}_{j-t} + I(A_{t+\ell} - B_{t+\ell})\tilde{\Lambda}_\ell \end{aligned} \tag{37}$$

Representing the identity matrix, I, as the empty-product, $\prod_{m=t+\ell+1}^{t+\ell} A_m$, we can write:

$$\Lambda_{\ell+1} - \tilde{\Lambda}_{\ell+1} = \sum_{j=t}^{t+\ell-1} \left(\prod_{m=j+1}^{t+\ell} A_m \right) (A_j - B_j)\tilde{\Lambda}_{j-t} + \left(\prod_{m=t+\ell+1}^{t+\ell} A_m \right) (A_{t+\ell} - B_{t+\ell})\tilde{\Lambda}_\ell$$
$$= \sum_{j=t}^{t+\ell} \left(\prod_{m=j+1}^{t+\ell} A_m \right) (A_j - B_j)\tilde{\Lambda}_{j-t} \tag{38}$$

Thus, the identity holds for the case $\ell + 1$, completing the proof by induction. □

Definition 6 (ResEKF Trajectory Mappings)**.** *Let F_t and H_t be the Jacobian matrices evaluated by ResEKF in Algorithm 2. For any $i \geq t$, the ResEKF trajectory matrix $\Phi_{i,t}$ is defined by the product:*

$$\Phi_{i,t} := \begin{cases} \prod_{j=t}^{i-1} F_j := F_{i-1}F_{i-2}\cdots F_t, & \text{when } i > t \\ I, & \text{when } i = t \end{cases} \tag{39}$$

The corresponding ResEKF observability Gramian over a window k is defined as:

$$M_{t+k,t} := \sum_{i=t}^{t+k} (H_i\Phi_{i,t})^\top H_i\Phi_{i,t} \tag{40}$$

Lemma L.8 (Trajectory Matrix Perturbation)**.** *Let $\Phi_{t+k,t}$ and $\tilde{\Phi}_{t+k,t}$ be the ResEKF and deterministic trajectory matrices, respectively defined in Definitions 6 and 5. Under the conditions of Lemma L.4, these two matrices satisfy:*

$$\|\Phi_{t+k,t} - \tilde{\Phi}_{t+k,t}\|_2 \leq L_\Phi(k)(\epsilon_{\sup} + \epsilon_v) \qquad \forall\, 1 \leq t \leq \tau \text{ and } 1 \leq k \leq \tau + 1 - t \tag{41}$$

where $\epsilon_{\sup}$ is the estimation error bound from Hypothesis 1, ϵ_v is the trajectory deviation constant from Assumption 3 and $L_\Phi(k) := k\bar{F}^{k-1}L_{\nabla f}$ is a finite, strictly positive scalar, with $\bar{F}$ as the transition Jacobian matrix bound defined in Lemma L.2 and $L_{\nabla f}$ as the transition Jacobian matrix perturbation bound defined in Lemma L.7.

Proof. We start by using Proposition P.1, while setting the window length $\ell = k$ and identifying the general sequences $\{A_j\}$ and $\{B_j\}$ as the ResEKF and deterministic Jacobians $\{F_j\}$ and $\{\tilde{F}_j\}$, respectively. This identification yields the following mappings for the matrix products:

$$\Lambda_\ell \to \Phi_{t+k,t}, \qquad \tilde{\Lambda}_\ell \to \tilde{\Phi}_{t+k,t}, \qquad \text{and} \qquad \tilde{\Lambda}_{j-t} \to \tilde{\Phi}_{j,t} \tag{42}$$

Substituting these into the identity from Proposition P.1:

$$\Phi_{t+k,t} - \tilde{\Phi}_{t+k,t} = \sum_{j=t}^{t+k-1} \left(\prod_{m=j+1}^{t+k-1} F_m \right) (F_j - \tilde{F}_j)\tilde{\Phi}_{j,t} \tag{43}$$

Taking the spectral norm and applying the triangle inequality and sub-multiplicative property of the norm:

$$\|\Phi_{t+k,t} - \tilde{\Phi}_{t+k,t}\|_2 \leq \sum_{j=t}^{t+k-1} \left(\prod_{m=j+1}^{t+k-1} \|F_m\|_2 \right) \|F_j - \tilde{F}_j\|_2 \|\tilde{\Phi}_{j,t}\|_2 \tag{44}$$

From Lemma L.4, we have the uniform bounds $\|F_m\|_2 \leq \bar{F}$ for all $1 \leq m \leq \tau$. Furthermore, from Lemma L.5, we have $\|\tilde{F}_j\|_2 \leq \bar{F}$ for all $j \geq 0$, which can be combined with the sub-multiplicative

property of the norm to arrive at $\|\tilde{\Phi}_{j,t}\|_2 \leq \bar{F}^{j-t}$ for all $j \geq 0$. Moreover, from Lemma L.7, we have $\|F_j - \tilde{F}_j\|_2 \leq L_{\nabla f}(\epsilon_{\sup} + \epsilon_v)$ for all $1 \leq j \leq \tau$. Applying these results to Eq. (44), we obtain:

$$\begin{aligned}\|\Phi_{t+k,t} - \tilde{\Phi}_{t+k,t}\|_2 &\leq \sum_{j=t}^{t+k-1} \bar{F}^{t+k-1-j} \cdot L_{\nabla f}(\epsilon_{\sup} + \epsilon_v) \cdot \bar{F}^{j-t} \\ &= \sum_{n=0}^{k-1} \bar{F}^{k-1} \cdot L_{\nabla f}(\epsilon_{\sup} + \epsilon_v) \\ &\leq k\bar{F}^{k-1} \cdot L_{\nabla f}(\epsilon_{\sup} + \epsilon_v) \qquad \forall\, 1 \leq t \leq \tau \text{ and } 1 \leq k \leq \tau + 1 - t \end{aligned} \tag{45}$$

Defining $L_\Phi(k) := k\bar{F}^{k-1} L_{\nabla f}$ yields:

$$\|\Phi_{t+k,t} - \tilde{\Phi}_{t+k,t}\|_2 \leq L_\Phi(k)(\epsilon_{\sup} + \epsilon_v) \qquad \forall\, 1 \leq t \leq \tau \text{ and } 1 \leq k \leq \tau + 1 - t \tag{46}$$

We can deduce that $L_\Phi(k)$ is finite and strictly positive, thus concluding the proof. □

Lemma L.9 (Gramian Perturbation)**.** *Let $M_{t+k,t}$ and $\tilde{M}_{t+k,t}$ be the ResEKF and deterministic observability Gramians over a window k, defined in Definitions 6 and 5 respectively. Under the conditions of Lemma L.8, their difference satisfies:*

$$\|M_{t+k,t} - \tilde{M}_{t+k,t}\|_2 \leq \mathcal{L}_M(k)(\epsilon_{sup} + \epsilon_v) + \mathcal{C}_M(k) c_{\max} \qquad \forall\, 1 \leq t \leq \tau \text{ and } 1 \leq k \leq \tau - t \tag{47}$$

where ϵ_{sup} is the estimation error bound from Hypothesis 1, ϵ_v is the trajectory deviation constant from Assumption 3 and $\mathcal{L}_M(k)$ and $\mathcal{C}_M(k)$ are finite, strictly positive scalar over the window k, given by:

$$\mathcal{L}_M(k) := 2\sum_{i=0}^{k} (\bar{H}\bar{F}^{2i+1} L_{\nabla h} + \bar{H}^2 \bar{F}^i L_\Phi(i)) \quad \text{and} \quad \mathcal{C}_M(k) := 2\bar{H} L_{\nabla h} \sum_{i=0}^{k} \bar{F}^{2i} \tag{48}$$

with $\bar{F}$ and $\bar{H}$ as the transition and measurement Jacobian matrix bounds defined in Lemma L.2, $L_{\nabla h}$ as the Lipschitz constant of the measurement Jacobian matrix defined in Lemma L.7 and $L_\Phi(i)$ as defined in Lemma L.8.

Proof. By their definitions, we can write the difference between the ResEKF and deterministic observability Gramians as:

$$M_{t+k,t} - \tilde{M}_{t+k,t} = \sum_{i=0}^{k} \left(\underbrace{\Phi_{t+i,t}^\top H_{t+i}^\top H_{t+i} \Phi_{t+i,t}}_{:=G_i} - \underbrace{\tilde{\Phi}_{t+i,t}^\top \tilde{H}_{t+i}^\top \tilde{H}_{t+i} \tilde{\Phi}_{t+i,t}}_{:=\tilde{G}_i} \right) \tag{49}$$

Let G_i and $\tilde{G}_i$ denote the i-th terms of the ResEKF and deterministic sums, respectively. We decompose the difference $\|G_i - \tilde{G}_i\|_2$ by introducing the cross term $G_i' = \Phi_{t+i,t}^\top \tilde{H}_{t+i}^\top \tilde{H}_{t+i} \Phi_{t+i,t}$ and applying the triangle inequality:

$$\begin{aligned}\|G_i - \tilde{G}_i\|_2 \leq\ & \|\Phi_{t+i,t}^\top (H_{t+i}^\top H_{t+i} - \tilde{H}_{t+i}^\top \tilde{H}_{t+i}) \Phi_{t+i,t}\|_2 \\ &+ \|(\tilde{H}_{t+i}\Phi_{t+i,t})^\top \tilde{H}_{t+i}\Phi_{t+i,t} - (\tilde{H}_{t+i}\tilde{\Phi}_{t+i,t})^\top \tilde{H}_{t+i}\tilde{\Phi}_{t+i,t}\|_2 \end{aligned} \tag{50}$$

Using the identity $A^\top A - B^\top B = A^\top (A - B) + (A - B)^\top B$ for matrices A and B of similar dimensions, we have:

$$\|G_i - \tilde{G}_i\|_2 \leq \|\Phi_{t+i,t}^\top (H_{t+i}^\top (H_{t+i} - \tilde{H}_{t+i}) - (H_{t+i} - \tilde{H}_{t+i})^\top \tilde{H}_{t+i}) \Phi_{t+i,t}\|_2$$
$$+ \|(\tilde{H}_{t+i}\Phi_{t+i,t})^\top (\tilde{H}_{t+i}\Phi_{t+i,t} - \tilde{H}_{t+i}\tilde{\Phi}_{t+i,t}) - (\tilde{H}_{t+i}\Phi_{t+i,t} - \tilde{H}_{t+i}\tilde{\Phi}_{t+i,t})^\top \tilde{H}_{t+i}\tilde{\Phi}_{t+i,t}\|_2 \quad (51)$$

Simplifying, we have:

$$\|G_i - \tilde{G}_i\|_2 \leq \|\Phi_{t+i,t}^\top (H_{t+i}^\top (H_{t+i} - \tilde{H}_{t+i}) + (H_{t+i} - \tilde{H}_{t+i})^\top \tilde{H}_{t+i}) \Phi_{t+i,t}\|_2$$
$$+ \|(\tilde{H}_{t+i}\Phi_{t+i,t})^\top \tilde{H}_{t+i}(\Phi_{t+i,t} - \tilde{\Phi}_{t+i,t}) + (\Phi_{t+i,t} - \tilde{\Phi}_{t+i,t})^\top \tilde{H}_{t+i}^\top \tilde{H}_{t+i}\tilde{\Phi}_{t+i,t}\|_2 \quad (52)$$

Using the triangle inequality, sub-multiplicative property of the spectral norm and the identity that $\|A\|_2 = \|A^\top\|_2$ for any matrix A, we have:

$$\begin{aligned}\|G_i - \tilde{G}_i\|_2 \leq \|\Phi_{t+i,t}\|_2 (\|H_{t+i}\|_2 \|H_{t+i} - \tilde{H}_{t+i}\|_2 + \|H_{t+i} - \tilde{H}_{t+i}\|_2 \|\tilde{H}_{t+i}\|_2) \|\Phi_{t+i,t}\|_2 \\ + \|\tilde{H}_{t+i}\|_2 \|\Phi_{t+i,t}\|_2 \|\tilde{H}_{t+i}\|_2 \|\Phi_{t+i,t} - \tilde{\Phi}_{t+i,t}\|_2 \\ + \|\Phi_{t+i,t} - \tilde{\Phi}_{t+i,t}\|_2 \|\tilde{H}_{t+i}\|_2 \|\tilde{H}_{t+i}\|_2 \|\tilde{\Phi}_{t+i,t}\|_2 \quad (53)\end{aligned}$$

This simplifies to:

$$\begin{aligned}\|G_i - \tilde{G}_i\|_2 \leq \|\Phi_{t+i,t}\|_2^2 (\|H_{t+i}\|_2 + \|\tilde{H}_{t+i}\|_2) \|H_{t+i} - \tilde{H}_{t+i}\|_2 \\ + \|\tilde{H}_{t+i}\|_2^2 (\|\Phi_{t+i,t}\|_2 + \|\tilde{\Phi}_{t+i,t}\|_2) \|\Phi_{t+i,t} - \tilde{\Phi}_{t+i,t}\|_2 \quad (54)\end{aligned}$$

Now, consider the components of Eq. (54). Using the sub-multiplicative property of the spectral norm of matrices which are a product of other matrices, we have for all $1 \leq t \leq \tau$:

$$\|\Phi_{t+i,t}\|_2 \leq \bar{F}^i \qquad \forall\, 2 - t \leq i \leq \tau + 1 - t \text{ [Lemma L.4]} \quad (55a)$$
$$\|H_{t+i}\|_2 \leq \bar{H} \qquad \forall\, 1 - t \leq i \leq \tau - t \text{ [Lemma L.4]} \quad (55b)$$
$$\|\tilde{H}_{t+i}\|_2 \leq \bar{H} \qquad \forall\, i \geq 1 \text{ [Lemma L.5]} \quad (55c)$$
$$\|\tilde{\Phi}_{t+i,t}\|_2 \leq \bar{F}^i \qquad \forall\, i \geq 1 \text{ [Lemma L.5]} \quad (55d)$$
$$\|H_{t+i} - \tilde{H}_{t+i}\|_2 \leq L_{\nabla h}(\bar{F}(\epsilon_{\text{sup}} + \epsilon_v) + c_{\max}) \qquad \forall\, 1 - t \leq i \leq \tau - t \text{ [Lemma L.7]} \quad (55e)$$
$$\|\Phi_{t+i,t} - \tilde{\Phi}_{t+i,t}\|_2 \leq L_\Phi(i)(\epsilon_{\text{sup}} + \epsilon_v) \qquad \forall\ 1 \leq i \leq \tau + 1 - t \text{ [Lemma L.8]} \quad (55f)$$

Substituting these in Eq. (54), we have for all $1 \leq t \leq \tau$ and $1 \leq i \leq \tau - t$:

$$\|G_i - \tilde{G}_i\|_2 \leq \bar{F}^{2i}(\bar{H} + \bar{H}) L_{\nabla h}(\bar{F}(\epsilon_{\text{sup}} + \epsilon_v) + c_{\max}) + \bar{H}^2(\bar{F}^i + \bar{F}^i) L_\Phi(i)(\epsilon_{\text{sup}} + \epsilon_v)$$
$$\iff \|G_i - \tilde{G}_i\|_2 \leq 2\bar{H}\bar{F}^i \left[\left(\bar{F}^{i+1} L_{\nabla h} + \bar{H} L_\Phi(i)\right)(\epsilon_{\text{sup}} + \epsilon_v) + \bar{F}^i L_{\nabla h} c_{\max} \right] \quad (56)$$

Defining $\mathcal{L}_M(k) := 2\bar{H} \sum_{i=0}^{k} (\bar{F}^{2i+1} L_{\nabla h} + \bar{H}\bar{F}^i L_\Phi(i))$ and $\mathcal{C}_M(k) := 2\bar{H} L_{\nabla h} (\sum_{i=0}^{k} \bar{F}^{2i})$ summing the bounds in Eq. (56) over $i \in [0, k]$ yields the bound below via the triangle inequality:

$$\|M_{t+k,t} - \tilde{M}_{t+k,t}\|_2 \leq \mathcal{L}_M(k)(\epsilon_{\text{sup}} + \epsilon_v) + \mathcal{C}_M(k) c_{\max} \qquad \forall\, 1 \leq t \leq \tau \text{ and } 1 \leq k \leq \tau - t \quad (57)$$

Finally, since their components are strictly positive and finite, and the summations are over a finite window k, $\mathcal{L}_M(k)$ and $\mathcal{C}_M(k)$ are both finite and strictly positive, thus concluding the proof. □

Lemma L.10 (ResEKF Discrete-Time Riccati Equation)**.** *ResEKF* a priori *state covariance estimates* $\{P_{t|t-1}\}_{t \geq 1}$ *satisfy a discrete-time Riccati equation operating with a modified effective process noise matrix.*

Proof. From Algorithm 2, the *a priori* covariance at time step t is computed as $P_{t|t-1} = F_t P_{t-1|t-1} F_t^\top + Q_t + P_{\text{BRC},\, t|t-1}$ and the *a posteriori* covariance at time step $t-1$ as $P_{t-1|t-1} = (I - K_{t-1}H_{t-1})P_{t-1|t-2} + P_{\text{BRC},\, t-1|t-1}$. Substituting the latter in the former yields:

$$\begin{aligned} P_{t|t-1} &= F_t(I - K_{t-1}H_{t-1})P_{t-1|t-2}F_t^\top + \bar{Q}_{t-1}, \\ \text{with} \quad \bar{Q}_{t-1} &:= F_t P_{\text{BRC},\, t-1|t-1} F_t^\top + Q_t + P_{\text{BRC},\, t|t-1} \end{aligned} \tag{58}$$

Since $P_{\text{BRC},\, t|\cdot}$ are parameterised to be positive semidefinite and Q_t is positive definite by Assumption 4, $\bar{Q}_{t-1}$ is positive definite. Now, from Algorithm 4, it can be worked out that:

$$K_{t-1} = P_{t-1|t-2}H_{t-1}^\top (H_{t-1}P_{t-1|t-2}H_{t-1}^\top + \Sigma_{t-1})^{-1} \tag{59}$$

Substituting Eq. (59) into Eq. (58) yields:

$$\begin{aligned} P_{t|t-1} = F_t \Big[& P_{t-1|t-2} \\ & - P_{t-1|t-2}H_{t-1}^\top (H_{t-1}P_{t-1|t-2}H_{t-1}^\top + \Sigma_{t-1})^{-1} H_{t-1}P_{t-1|t-2} \Big] F_t^\top + \bar{Q}_{t-1} \end{aligned} \tag{60}$$

Equation (60) is algebraically identical to the canonical discrete-time Riccati equation (Anderson & Moore, 1979, p. 39), with the *effective* process noise matrix $\bar{Q}_{t-1}$. □

Lemma L.11 (Bounded *A Priori* Covariance)**.** *Let Assumptions 4 and 5 and the conditions of Lemma L.9 hold. Further, consider the following additional condition: the trajectory deviation constant from Assumption 3, ϵ_v, and the estimation error bound from Hypothesis 1, ϵ_{sup}, satisfy:*

$$\textit{(Covariance Boundedness Condition)} \qquad \mathcal{L}(s)(\epsilon_{sup} + \epsilon_v) + \mathcal{C}(s)c_{\max} < \mathcal{M}(s) \tag{61}$$

where s is the nominal state dimension and the terms $\mathcal{L}(s)$, $\mathcal{C}(s)$ and $\mathcal{M}(s)$ are strictly positive scalars defined as:

$$\mathcal{L}(s) = \begin{cases} \mathcal{L}_M(s-1) & s > 1 \\ \bar{F}L_{\nabla h} & s = 1 \end{cases}, \mathcal{C}(s) = \begin{cases} \mathcal{C}_M(s-1) & s > 1 \\ L_{\nabla h} & s = 1 \end{cases}, \textit{ and } \mathcal{M}(s) = \begin{cases} m_{det}(s-1) & s > 1 \\ \sqrt{m_{det}(0)} & s = 1 \end{cases} \tag{62}$$

where $m_{\text{det}}(\cdot)$ is the deterministic observability Gramian bound defined in Lemma L.6, specifically Eq. (28), $\mathcal{L}_M(\cdot)$ and $\mathcal{C}_M(\cdot)$ are Gramian perturbation coefficients defined in Lemma L.9, and $L_{\nabla h}$ is the measurement Jacobian Lipschitz constant defined in Lemma L.5.

Provided all the above conditions hold, there exist strictly positive, finite scalars $\underline{p}$ and $\bar{p}$ such that ResEKF's a priori *state covariance estimate $P_{t|t-1}$, the solution of its Riccati equation (60), is bounded according to:*

$$\underline{p}I \preceq P_{t|t-1} \preceq \bar{p}I, \quad \forall\, 1 \le t \le \tau \tag{63}$$

Proof. We prove the lemma by establishing the bounds across two distinct cases for the nominal state dimension: $s = 1$ and $s > 1$.

Scalar State Case ($s = 1$): In this case, we first establish the upper bound, $\bar{p}$. We start by applying the reverse triangle inequality to measurement function Jacobian matrix. Specifically, we note that $|\tilde{H}_{t-1} - H_{t-1}| \ge |\tilde{H}_{t-1}| - |H_{t-1}|$. However, Lemma L.6 grants that $|\tilde{H}_{t-1}| \ge \sqrt{m_{\text{det}}(0)} > 0$ and Lemma L.7, that $\|H_t - \tilde{H}_t\|_2 \le L_{\nabla h}(\bar{F}(\epsilon_{\text{sup}} + \epsilon_v) + c_{\max})$ for all $1 \le t \le \tau$. Thus, we have:

$$|H_{t-1}| \ge |\tilde{H}_{t-1}| - |H_{t-1} - \tilde{H}_{t-1}| \ge \sqrt{m_{\text{det}}(0)} - L_{\nabla h}(\bar{F}(\epsilon_{\text{sup}} + \epsilon_v) + c_{\max}) := \underline{h} \quad \forall\, 1 \le t \le \tau \tag{64}$$

The lemma's explicit condition that $\bar{F}(\epsilon_{\text{sup}} + \epsilon_v) + c_{\max} < \sqrt{m_{\text{det}}(0)}/L_{\nabla h}$ strictly guarantees that $\underline{h} > 0$. Expressing the scalar *a posteriori* model-based variance in information form, and recognizing that the strictly positive *a priori* term $P_{t-1|t-2}^{-1} > 0$ only decreases the total variance, we obtain a strict upper bound:

$$P_{\text{KF},\, t-1|t-1} = \left(P_{t-1|t-2}^{-1} + H_{t-1}^2 \Sigma_{t-1}^{-1}\right)^{-1} < \Sigma_{t-1} H_{t-1}^{-2} \leq \frac{\Sigma_{t-1}}{\underline{h}^2} \leq \frac{\bar{s}}{\underline{h}^2} \quad \forall\, 1 \leq t \leq \tau \tag{65}$$

By Eq. (1), the ResEKF *a posteriori* variance adds a bounded neural residual to its model-based counterpart, with the residual magnitude bounded by $c_{\max}$. Thus:

$$P_{t-1|t-1} = P_{\text{KF},\, t-1|t-1} + P_{\text{BRC},\, t-1|t-1} < \frac{\bar{s}}{\underline{h}^2} + c_{\max} \quad \forall\, 1 \leq t \leq \tau \tag{66}$$

Substituting this into the scalar time update and applying the transition Jacobian bound $F_t^2 \leq \bar{F}^2$ (Lemma L.4), we have:

$$P_{\text{KF},\, t|t-1} = F_t^2 P_{t-1|t-1} + Q_t < \bar{F}^2 \left(\frac{\bar{s}}{\underline{h}^2} + c_{\max}\right) + \bar{q} \quad \forall\, 1 \leq t \leq \tau \tag{67}$$

Again, adding the bounded neural residual to the model-based *a priori* variance via Eq. (1), we have:

$$P_{t|t-1} = P_{\text{KF},\, t|t-1} + P_{\text{BRC},\, t|t-1} < \underbrace{\bar{F}^2 \left(\frac{\bar{s}}{\underline{h}^2} + c_{\max}\right) + \bar{q} + c_{\max}}_{:=\bar{p}} \quad \forall\, 1 \leq t \leq \tau \tag{68}$$

Next, we establish the lower bound, $\underline{p}$. The term $F_t^2 P_{t-1|t-1}$ in the scalar time update is a product of squared values and a previous variance estimate (which is non-negative by induction), it is strictly non-negative. Consequently, the model-based *a priori* variance is bounded from below by the process noise covariance Q_t, which is positive by Assumption 4, and the added neural residual, $P_{\text{BRC},\, t|t-1}$ is also non-negative by Eq. (1). Thus we have:

$$P_{t|t-1} = P_{\text{KF},\, t|t-1} + P_{\text{BRC},\, t|t-1} = F_t^2 P_{t-1|t-1} + Q_t + P_{\text{BRC},\, t|t-1} \geq Q_t \geq \underline{q} := \underline{p} \tag{69}$$

Since $\underline{q}$ is strictly positive by Assumption 4, and the previously defined $\bar{p}$ is a finite sum of strictly positive, finite parameters, we have successfully established that $0 < \underline{p} \leq P_{t|t-1} \leq \bar{p} < \infty$ for all $1 \leq t \leq \tau$.

Multivariate State Case ($s > 1$): In this case, the proof proceeds by partitioning the timeline into two phases: a finite transient phase ($1 \leq t < s$) and a steady-state phase ($s \leq t \leq \tau$).[5], subsequently unifying their respective bounds. We first establish the boundedness and positive definiteness of the initial covariance (the 'initial condition') and effective noise covariances. These properties guarantee that the covariance remains well-behaved during the transient phase, while simultaneously satisfying Conditions 1 and 3 of Lemma 4.1 in Reif et al. (1999), which we require for the steady-state phase.

Positive Definite Initial Condition:
The initial condition of the ResEKF Riccati equation (60) is $P_{1|0}$. From Algorithm 2, $P_{1|0} = F_1 P_{0|0} F_1^\top + Q_1 + P_{\text{BRC},\, 1|0}$. By Assumption 5, $P_{0|0}$ is strictly positive definite. Since $Q_1 \succ 0$ and $P_{\text{BRC},\, 1|0} \succeq 0$, it follows that $P_{1|0} \succ 0$.

Bounded Noise Matrices:
By definition, the neural covariance residuals $P_{\text{BRC},\, t|\cdot}$ are positive semi-definite and bounded so that $0 \preceq P_{\text{BRC},\, t|\cdot} \preceq c_{\max} I$ for all t. Moreover, by Lemma L.4, the spectral norm of the transition

[5]Note that the steady-state phase exists only in the case when $\tau \geq s$. We proceed with the assumption that $\tau \geq s$ without loss of generality.

Jacobian is bounded by $\|F_t\|_2 \leq \bar{F}$ for all $1 \leq t \leq \tau$. Since the nominal process noise Q_t satisfies $\underline{q}I \preceq Q_t \preceq \bar{q}I$ for all $t \geq 0$ (Assumption 4) and spectral norm is sub-multiplicative, the *effective* process noise $\bar{Q}_{t-1} := F_t P_{\text{BRC},\, t-1|t-1} F_t^\top + Q_t + P_{\text{BRC},\, t|t-1}$ (Eq. (58)) is strictly positive definite and uniformly bounded. We therefore have:

$$\underline{q}I \preceq \bar{Q}_{t-1} \preceq (\bar{F}^2 c_{\max} + \bar{q} + c_{\max})I \qquad \forall 1 \leq t \leq \tau \tag{70}$$

The measurement noise covariance Σ_t is also bounded by Assumption 4, $\underline{s}I \preceq \Sigma_t \preceq \bar{s}I$. Thus, both noise covariances $\bar{Q}_{t-1}$ and Σ_t are bounded for all $1 \leq t \leq \tau$.

Transient Phase:
During the transient phase ($1 \leq t < s$), the covariance $P_{t|t-1}$ is generated by a strictly finite number of Riccati updates. Because the initial condition $P_{1|0}$ is strictly positive definite, the effective noise covariances $\bar{Q}_{t-1}$ and Σ_t are strictly positive definite and uniformly bounded, and the process and measurement Jacobian matrices are bounded (Lemma L.4), $P_{t|t-1}$ can neither diverge to infinity nor become singular in a finite number of steps. Consequently, there exist strictly positive, finite scalars $\underline{p}_{\text{tr}}$ and $\bar{p}_{\text{tr}}$ such that:

$$\underline{p}_{\text{tr}} I \preceq P_{t|t-1} \preceq \bar{p}_{\text{tr}} I \qquad \forall\, 1 \leq t < s \tag{71}$$

Steady-State Phase:
To establish the steady-state bounds for $t \geq s$, we apply the Riccati stability results of Lemma 4.1 in Reif et al. (1999) to ResEKF's Riccati equation (60). We now verify that the remaining condition of Uniform Observability required by this lemma is met, under Assumptions 4 and 5 and the conditions of Lemma L.9.

Uniform Observability (in the Steady-State Phase):
Uniform Observability requires that there exist an (integer) window $k > 0$ and strictly positive, finite scalars $\underline{m}(k)$ and $\bar{m}(k)$ such that ResEKF observability Gramian $M_{t+k,t}$ defined by Definition 6 satisfies:

$$\underline{m}(k)I \preceq M_{t+k,t} \preceq \bar{m}(k)I \qquad \forall\, s \leq t \leq \tau \tag{72}$$

with:

$$M_{t+k,t} := \sum_{i=t}^{t+k} (H_i \Phi_{i,t})^\top H_i \Phi_{i,t}, \qquad \Phi_{i,t} := \begin{cases} \prod_{j=t}^{i-1} F_j := F_{i-1} \cdots F_t, & \text{when } i > t \\ I, & \text{when } i = t \end{cases} \tag{73}$$

To establish the upper bound coefficient $\bar{m}(k)$, we make use of Eq. (55), which, in turn, uses Lemma L.4 and the sub-multiplicative property of the spectral norm of matrices, to get $\|\Phi_{t+i,t}\|_2 \leq \bar{F}^i$ and $\|H_{t+i}\|_2 \leq \bar{H}$ for all $1 \leq t \leq \tau$ and $1 \leq i \leq \tau + 1 - t$. Here, $\bar{F}$ and $\bar{H}$ are upper bounds on the spectral norms of the process and measurement Jacobian matrices under Hypothesis 1. Thus, we have:

$$\begin{aligned} \|M_{t+k,t}\|_2 \leq \sum_{i=t}^{t+k} (\bar{H}\bar{F}^{i-t})^2 &= \underbrace{\bar{H}^2 \sum_{j=0}^{k} \bar{F}^{2j}}_{:=\bar{m}(k)} \\ \implies M_{t+k,t} &\preceq \bar{m}(k) I \qquad \forall\, 1 \leq k \leq \tau - t \text{ and } s \leq t \leq \tau \end{aligned} \tag{74}$$

Since $\bar{H}$ and $\bar{F}$ are both finite and positive and k is also finite, we have $\infty > \bar{m}(k) > 0$. This proves that there exist k and a finite, strictly positive $\bar{m}(k)$ satisfying Eq. (72).

To establish the lower bound coefficient $\underline{m}(k)$, we proceed as follows. By definition, both matrices $M_{t+k,t}$ and $\tilde{M}_{t+k,t}$ are symmetric. Now, by Weyl's Perturbation Theorem for symmetric matrices,

the absolute difference between the corresponding eigenvalues of two matrices is bounded by the spectral norm of their difference. Applying this specifically to the minimum eigenvalues of ResEKF and deterministic observability Gramians yields:

$$|\lambda_{\min}(M_{t+k,t}) - \lambda_{\min}(\tilde{M}_{t+k,t})| \leq \|M_{t+k,t} - \tilde{M}_{t+k,t}\|_2 \tag{75}$$

From Lemma L.9, we have the following Gramian perturbation bound: $\|M_{t+k,t} - \tilde{M}_{t+k,t}\|_2 \leq \mathcal{L}_M(k)(\epsilon_{\text{sup}} + \epsilon_v) + \mathcal{C}_M(k)c_{\max}$ for all $1 \leq t \leq \tau$ and $1 \leq k \leq \tau - t$, with $\mathcal{L}_M(k), \mathcal{C}_M(k) > 0$, $\epsilon_{\text{sup}} > 0$ is the estimation error bound from Hypothesis 1, $\epsilon_v > 0$ is the trajectory deviation constant from Assumption 3. Substituting this bound above yields:

$$\lambda_{\min}(M_{t+k,t}) \geq \lambda_{\min}(\tilde{M}_{t+k,t}) - \mathcal{L}_M(k)(\epsilon_{\text{sup}} + \epsilon_v) - \mathcal{C}_M(k)c_{\max} \quad \forall\, s \leq t \leq \tau \text{ and } 1 \leq k \leq \tau - t \tag{76}$$

Moreover, from Lemma L.6 concerning the deterministic observability Gramian $\tilde{M}_{t+k,t}$ defined in Definition 5, we have $\tilde{M}_{t+k,t} \succeq m_{\text{det}}(k)I$, implying that the lowest eigenvalue of this Gramian satisfies $\lambda_{\min}(\tilde{M}_{t+k,t}) \geq m_{\text{det}}(k)$ for all $t \geq 0$ and $k \geq s - 1$. Substituting this bound in Eq. (76) yields:

$$\begin{aligned}\lambda_{\min}(M_{t+k,t}) &\geq \underbrace{m_{\text{det}}(k) - \mathcal{L}_M(k)(\epsilon_{\text{sup}} + \epsilon_v) - \mathcal{C}_M(k)c_{\max}}_{:=\underline{m}(k)} \\ &\implies M_{t+k,t} \succeq \underline{m}(k)I \qquad \forall\, s-1 \leq k \leq \tau - t \text{ and } s \leq t \leq \tau\end{aligned} \tag{77}$$

Now, comparing Eqs. (74) and (77), we see that both bounds $\bar{m}(k)$ and $\underline{m}(k)$ exist for $s - 1 \leq k \leq \tau - t$. Setting the window k to the minimum, $s - 1$, eliminates dependence on k. For $k = s - 1$, we established finiteness and strict positivity of $\bar{m}(s-1)$ in the discussion following Eq. (74).

Now, by definition, $m_{\text{det}}(s-1)$, $\mathcal{L}_M(s-1)$, $\mathcal{C}_M(s-1)$, ϵ_{sup} and ϵ_v are all finite and strictly positive. This, combined with the fact that the lemma's explicit condition grants that $\mathcal{L}_M(s-1)(\epsilon_{\text{sup}} + \epsilon_v) + \mathcal{C}_M(s-1)c_{\max} < m_{\text{det}}(s-1)$ ensures that $\underline{m}(s-1)$ is finite and strictly positive.

Thus, we have strictly positive, finite scalars $\underline{m}(s-1)$ and $\bar{m}(s-1)$ for the window $k = s - 1$, satisfying the Uniform Observability condition, Eq. (72).

Stable Riccati Equation in the Steady-State Phase:
Having satisfied all three conditions of Lemma 4.1 of Reif et al. (1999) for the Steady State ($s \leq t \leq \tau$), we prove that ResEKF's Riccati equation (60) is stable and obtain strictly positive, finite scalars:

$$\underline{p}_{\text{ss}} I \preceq P_{t|t-1} \preceq \bar{p}_{\text{ss}} I \qquad \forall\, s \leq t \leq \tau \tag{78}$$

Combining Results over the Transient and Steady-State Phases:
From Eqs. (71) and (78), we obtain strictly positive and finite scalars $\underline{p} := \min(\underline{p}_{\text{tr}},\ \underline{p}_{\text{ss}})$ and $\bar{p} := \max(\bar{p}_{\text{tr}},\ \bar{p}_{\text{ss}})$, satisfying Eq. (63), thus concluding the proof.

□

Lemma L.12 (Bounded Inverse *A Posteriori* Covariance)**.** *Under the conditions of Lemma L.11, there exist strictly positive, finite scalars $\underline{\pi}$ and $\bar{\pi}$ such that the inverse of the ResEKF* a posteriori *state covariance estimate produced by Algorithm 2, $P_{t|t}^{-1}$, is bounded according to:*

$$\underline{\pi} I \preceq P_{t|t}^{-1} \preceq \bar{\pi} I, \quad \forall\, 1 \leq t \leq \tau \tag{79}$$

Proof. Expressed in information form (see e.g., (Anderson & Moore, 1979, p. 139)), the inverse of this model-based *a posteriori* covariance produced by Algorithm 2 is given by:

$$P_{\text{KF},\,t|t}^{-1} = P_{t|t-1}^{-1} + H_t^\top \Sigma_t^{-1} H_t \tag{80}$$

To establish the lower bound coefficient $\underline{\pi}$, we bound $P_{t|t}$ from above. Since the measurement noise covariance Σ_t is positive definite (Assumption 4), its inverse is positive definite, making $H_t^\top \Sigma_t^{-1} H_t$ strictly positive semi-definite ($H_t^\top \Sigma_t^{-1} H_t \succeq 0$). Consequently, from Eq. (80), we have $P_{\text{KF},\,t|t}^{-1} \succeq P_{t|t-1}^{-1}$. Inverting this relationship implies $P_{\text{KF},\,t|t} \preceq P_{t|t-1}$.

From Lemma L.11, $P_{t|t-1} \preceq \bar{p} I$, which gives $P_{\text{KF},\,t|t} \preceq \bar{p} I$ for all $1 \le t \le \tau$. Furthermore, by Eq. (1), the neural residual is bounded such that $P_{\text{BRC},\,t|t} \preceq c_{\max} I$. Substituting these upper bounds into Eq. (1) yields:

$$P_{t|t} \preceq (\bar{p} + c_{\max}) I \implies P_{t|t}^{-1} \succeq \underbrace{\frac{1}{\bar{p} + c_{\max}}}_{:=\underline{\pi}} I \qquad \forall\, 1 \le t \le \tau \tag{81}$$

Defining $\underline{\pi} := \frac{1}{\bar{p}+c_{\max}}$, we obtain a strictly positive, finite scalar lower bound, satisfying the left-hand side of Eq. (79), since $\bar{p}$ and $c_{\max}$ are both strictly positive, finite scalars.

To establish the upper bound coefficient $\bar{\pi}$, we bound $P_{t|t}^{-1}$ from above. Because the neural residual $P_{\text{BRC},\,t|t}$ is positive semi-definite, we have $P_{t|t} \succeq P_{\text{KF},\,t|t}$. Inverting this relationship yields:

$$P_{t|t}^{-1} \preceq P_{\text{KF},\,t|t}^{-1} = P_{t|t-1}^{-1} + H_t^\top \Sigma_t^{-1} H_t \tag{82}$$

From Lemma L.11, $P_{t|t-1} \succeq \underline{p} I$ implies $P_{t|t-1}^{-1} \preceq \frac{1}{\underline{p}} I$ for all $1 \le t \le \tau$. For the measurement update term, $H_t^\top \Sigma_t^{-1} H_t$, Lemma L.4 yields $\|H_t\|_2 \le \bar{H}$, implying $H_t^\top H_t \preceq \bar{H}^2 I$ for all $t \le \tau$. By Assumption 4, $\Sigma_t \succeq \underline{s} I$, which implies $\Sigma_t^{-1} \preceq \frac{1}{\underline{s}} I$. Using the sub-multiplicative property of spectral norms, we have $\|H_t^\top \Sigma_t^{-1} H_t\|_2 \le \|H_t\|_2^2 \|\Sigma_t^{-1}\|_2 \le \frac{\bar{H}^2}{\underline{s}}$. Because $H_t^\top \Sigma_t^{-1} H_t$ is symmetric and positive semi-definite, its maximum eigenvalue is bounded by its spectral norm, yielding $H_t^\top \Sigma_t^{-1} H_t \preceq \frac{\bar{H}^2}{\underline{s}} I$. Substituting these bounds in the above inequality, we have:

$$P_{t|t}^{-1} \preceq \underbrace{\left(\frac{1}{\underline{p}} + \frac{\bar{H}^2}{\underline{s}}\right)}_{:=\bar{\pi}} I \tag{83}$$

Defining $\bar{\pi} := \frac{1}{\underline{p}} + \frac{\bar{H}^2}{\underline{s}}$, we obtain a strictly positive and finite scalar, satisfying the right-hand side of Eq. (79), since $\underline{p}$, $\bar{H}$ and $\underline{s}$ are all finite, strictly positive scalars. This concludes the proof. □

Definition 7 (Lyapunov Function Candidate)**.** *For our analysis, we adopt the standard Lyapunov function candidate for the estimation error e_t:*

$$V_t(e_t) = e_t^\top P_{t|t}^{-1} e_t \tag{84}$$

where $P_{t|t}$ is the a posteriori state estimate covariance matrix computed by Algorithm 2 at time t.

Note that we sometimes denote $V_t(e_t)$ by the shorthand V_t.

Lemma L.13 (Rayleigh Quotient Bounds)**.** *Under the conditions of Lemma L.12, the Lyapunov function candidate $V_t(e_t)$ defined in Eq. (84) is bounded by the squared Euclidean norm of the estimation error according to:*

$$\underline{\pi} \|e_t\|_2^2 \le V_t(e_t) \le \bar{\pi} \|e_t\|_2^2 \quad \forall\, 1 \le t \le \tau \tag{85}$$

where $\underline{\pi}$ and $\bar{\pi}$ are the strictly positive, finite scalars established in Lemma L.12.

Proof. By Definition 7, the Lyapunov function is the quadratic form $V_t(e_t) = e_t^\top P_{t|t}^{-1} e_t$. From standard properties of the Rayleigh quotient for symmetric matrices, any quadratic form $x^\top Ax$ is bounded by the minimum and maximum eigenvalues of A as $\lambda_{\min}(A)\|x\|_2^2 \le x^\top Ax \le \lambda_{\max}(A)\|x\|_2^2$.

From Lemma L.12, we have the inequality $\underline{\pi} I \preceq P_{t|t}^{-1} \preceq \bar{\pi} I$ for all $1 \le t \le \tau$. This inequality implies that $\lambda_{\min}(P_{t|t}^{-1}) \ge \underline{\pi}$ and $\lambda_{\max}(P_{t|t}^{-1}) \le \bar{\pi}$. Pre-multiplying this inequality by $e_t^\top$ and post-multiplying by e_t yields:

$$e_t^\top(\underline{\pi} I)e_t \le e_t^\top P_{t|t}^{-1} e_t \le e_t^\top(\bar{\pi} I)e_t \quad \forall\, 1 \le t \le \tau \tag{86}$$

Since $e_t^\top I e_t = e_t^\top e_t = \|e_t\|_2^2$, we obtain:

$$\underline{\pi}\|e_t\|_2^2 \le V_t(e_t) \le \bar{\pi}\|e_t\|_2^2 \quad \forall\, 1 \le t \le \tau \tag{87}$$

□

Proposition P.2. *If X and Y are matrices such that the inverse of their sum $(X+Y)$ exists, they satisfy:*

$$X - X(X+Y)^{-1}X = Y - Y(X+Y)^{-1}Y \tag{88}$$

Proof. Using the trivial identity $X = (X+Y) - Y$, we have:

$$\begin{aligned}
X - X(X+Y)^{-1}X &= X - \Big[(X+Y) - Y\Big](X+Y)^{-1}X \\
&= X - \Big[(X+Y)(X+Y)^{-1}X - Y(X+Y)^{-1}X\Big] \\
&= X - \Big[X - Y(X+Y)^{-1}X\Big] \\
&= Y(X+Y)^{-1}X \\
&= Y(X+Y)^{-1}\Big[(X+Y) - Y\Big] \\
&= Y(X+Y)^{-1}(X+Y) - Y(X+Y)^{-1}Y \\
&= Y - Y(X+Y)^{-1}Y
\end{aligned} \tag{89}$$

□

Lemma L.14 (Unforced Error Dissipation)**.** *Let $E_t := (I - K_t H_t)F_t$ be the error dynamics matrix of the ResEKF as defined in Eq. (19a). Under the conditions of Lemma L.12, there exists a scalar $\alpha \in (0,1)$ such that:*

$$E_t^\top P_{t|t}^{-1} E_t \preceq (1-\alpha) P_{t-1|t-1}^{-1} \qquad \forall\, 1 \le t \le \tau \tag{90}$$

where $P_{\cdot|\cdot}$ are ResEKF a posteriori state covariances evaluated by Algorithm 2.

Proof. We start with Eq. (82), from which we have $P_{t|t}^{-1} \preceq P_{\text{KF},\, t|t}^{-1}$, which yields:

$$(I - K_t H_t)^\top P_{t|t}^{-1}(I - K_t H_t) \preceq (I - K_t H_t)^\top P_{\text{KF},\, t|t}^{-1}(I - K_t H_t) \tag{91}$$

We then prove the identity that $P_{t|t-1}^{-1} = P_{\text{KF},\, t|t}^{-1}(I - K_t H_t)$ as shown below:

$$\begin{aligned}
& P_{\text{KF},\, t|t} = (I - K_t H_t) P_{t|t-1} && \text{[from Algorithm 2]} \\
\implies & P_{\text{KF},\, t|t} P_{t|t-1}^{-1} = I - K_t H_t && \text{[post-multiplying by } P_{t|t-1}^{-1}\text{]} \\
\implies & P_{t|t-1}^{-1} = P_{\text{KF},\, t|t}^{-1}(I - K_t H_t) && \text{[pre-multiplying by } P_{\text{KF},\, t|t}^{-1}\text{]}
\end{aligned} \tag{92}$$

Substituting this in Eq. (91), we have:

$$(I - K_t H_t)^\top P_{t|t}^{-1}(I - K_t H_t) \preceq (I - K_t H_t)^\top P_{t|t-1}^{-1} \tag{93}$$

Distributing terms in the right-hand side and using the definition $K_t := P_{t|t-1}H_t^\top(H_t P_{t|t-1}H_t^\top + \Sigma_t)^{-1}$ from Algorithm 2 yields:

$$\begin{aligned}(I - K_t H_t)^\top P_{t|t}^{-1}(I - K_t H_t) &\preceq P_{t|t-1}^{-1} - H_t^\top(H_t P_{t|t-1}H_t^\top + \Sigma_t)^{-1}H_t \\ &\preceq P_{t|t-1}^{-1} \qquad \forall\, 1 \le t \le \tau \end{aligned} \tag{94}$$

where the final inequality holds because we are subtracting a positive semi-definite matrix from a positive definite matrix. Specifically, Lemma L.11 establishes $P_{t|t-1} \succ 0$, which guarantees that the quadratic form $H_t P_{t|t-1}H_t^\top \succeq 0$ for any measurement Jacobian H_t. Because the measurement noise covariance is strictly positive definite, $\Sigma_t \succ 0$ (Assumption 4), the combined innovation covariance $H_t P_{t|t-1}H_t^\top + \Sigma_t$ is strictly positive definite, ensuring its inverse is also strictly positive definite. Consequently, the outer quadratic form $H_t^\top(H_t P_{t|t-1}H_t^\top + \Sigma)^{-1}H_t \succeq 0$. Subtracting this positive semi-definite matrix from $P_{t|t-1}^{-1}$ guarantees a resulting matrix $\preceq P_{t|t-1}^{-1}$.

We pre-multiply by $F_t^\top$ and post-multiply by F_t the above inequality and use the definition $E_t := (I - K_t H_t)F_t$ to obtain:

$$E_t^\top P_{t|t}^{-1} E_t \preceq F_t^\top P_{t|t-1}^{-1} F_t \qquad \forall\, 1 \le t \le \tau \tag{95}$$

Now, from Algorithm 2, we have $P_{t|t-1} = F_t P_{t-1|t-1}F_t^\top + \tilde{Q}_t$, where $\tilde{Q}_t := Q_t + P_{\text{BRC},\, t|t-1}$. Furthermore, since $Q_t \succ 0$ (Assumption 4) and $P_{\text{BRC},\, t|t-1} \succeq 0$ by construction (Eq. (1)), $\tilde{Q}_t$ is invertible. Moreover, from Lemma L.12, we have $P_{t-1|t-1} \succ 0$ as well, for all $1 \le t \le \tau$. Thus, we can use the Matrix Inversion Lemma as formulated by Anderson & Moore (1979, Eq. 3.1, p. 139), i.e.,:

$$(A^{-1} + BC^{-1}B^\top)^{-1} = A - AB(B^\top AB + C)^{-1}B^\top A \tag{96}$$

setting $A^{-1} = \tilde{Q}_t$, $B = F_t$ and $C^{-1} = P_{t-1|t-1}$, to obtain:

$$\begin{aligned}P_{t|t-1}^{-1} &= (F_t P_{t-1|t-1}F_t^\top + \tilde{Q}_t)^{-1} \\ &= \tilde{Q}_t^{-1} - \tilde{Q}_t^{-1}F_t(F_t^\top \tilde{Q}_t^{-1}F_t + P_{t-1|t-1}^{-1})^{-1}F_t^\top \tilde{Q}_t^{-1} \qquad \forall\, 1 \le t \le \tau \end{aligned} \tag{97}$$

Pre-multiplying by $F_t^\top$ and post-multiplying by F_t, we have for all $1 \le t \le \tau$:

$$F_t^\top P_{t|t-1}^{-1} F_t = (F_t^\top \tilde{Q}_t^{-1} F_t) - (F_t^\top \tilde{Q}_t^{-1} F_t)((F_t^\top \tilde{Q}_t^{-1} F_t) + P_{t-1|t-1}^{-1})^{-1}(F_t^\top \tilde{Q}_t^{-1} F_t) \tag{98}$$

Applying Proposition P.2, setting $X = F_t^\top \tilde{Q}_t^{-1} F_t$ and $Y = P_{t-1|t-1}^{-1}$, we have:

$$F_t^\top P_{t|t-1}^{-1} F_t = P_{t-1|t-1}^{-1} - P_{t-1|t-1}^{-1}(P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t)^{-1} P_{t-1|t-1}^{-1} \qquad \forall\, 1 \le t \le \tau \tag{99}$$

Substituting this back in Eq. (95), we finally obtain a relationship between $E_t^\top P_{t|t}^{-1} E_t$ and $P_{t-1|t-1}^{-1}$, i.e.:

$$E_t^\top P_{t|t}^{-1} E_t \preceq P_{t-1|t-1}^{-1} - P_{t-1|t-1}^{-1}(P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t)^{-1} P_{t-1|t-1}^{-1} \qquad \forall\, 1 \le t \le \tau \tag{100}$$

Now, to obtain the scalar α, we first upper-bound the inner matrix sum $P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t$. Lemma L.12 establishes that $\underline{\pi} I \preceq P_{t-1|t-1}^{-1} \preceq \bar{\pi} I$ for all $1 \leq t \leq \tau$. Next, we lower-bound $\tilde{Q}_t$ and then invert it, as follows. By construction, $P_{\text{BRC},\, t|t-1} \succeq 0$ (Eq. (1)). Adding the nominal process noise Q_t to both sides of this inequality yields $\tilde{Q}_t \succeq Q_t$. By Assumption 4, we have $\tilde{Q}_t \succeq Q_t \succeq \underline{q} I$. Inverting both sides of this inequality reverses the spectral bound, yielding $\tilde{Q}_t^{-1} \preceq \underline{q}^{-1} I$.

Lemma L.4 guarantees that the transition Jacobian matrix, F_t, is bounded such that $\|F_t\|_2 \leq \bar{F}$ for all $1 \leq t \leq \tau$. Using the sub-multiplicative property of the spectral norm, we have $\|F_t^\top \tilde{Q}_t^{-1} F_t\|_2 \leq \|F_t^\top\|_2 \|\tilde{Q}_t^{-1}\|_2 \|F_t\|_2 \leq \frac{\bar{F}^2}{\underline{q}}$. By applying the triangle inequality for spectral norms, we bound the sum of the inner matrices: $\|P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t\|_2 \leq \bar{\pi} + \frac{\bar{F}^2}{\underline{q}}$. Because both $P_{t-1|t-1}^{-1}$ and $F_t^\top \tilde{Q}_t^{-1} F_t$ are symmetric matrices, their sum is also symmetric. Thus, the spectral norm bound on $P_{t-1|t-1}^{-1} + F_t^\top (Q_t')^{-1} F_t$ yields:

$$
\begin{aligned}
& P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t \preceq \left(\bar{\pi} + \frac{\bar{F}^2}{\underline{q}} \right) I \\
\implies & (P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t)^{-1} \succeq \left(\bar{\pi} + \frac{\bar{F}^2}{\underline{q}} \right)^{-1} I \qquad \forall\, 1 \leq t \leq \tau
\end{aligned} \tag{101}
$$

We then pre- and post-multiply this inequality by $P_{t-1|t-1}^{-1}$; because $P_{t-1|t-1}^{-1}$ is symmetric, this operation preserves the positive semi-definite ordering. Thus, for all $1 \leq t \leq \tau$, we obtain:

$$
P_{t-1|t-1}^{-1} (P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t)^{-1} P_{t-1|t-1}^{-1} \succeq \left(\bar{\pi} + \frac{\bar{F}^2}{\underline{q}} \right)^{-1} P_{t-1|t-1}^{-1} P_{t-1|t-1}^{-1} \tag{102}
$$

From Lemma L.12, we have the strict lower bound $P_{t-1|t-1}^{-1} \succeq \underline{\pi} I$. Because $P_{t-1|t-1}^{-1}$ is positive definite and commutes with itself, multiplying both sides of this lower bound by $P_{t-1|t-1}^{-1}$ yields $P_{t-1|t-1}^{-1} P_{t-1|t-1}^{-1} \succeq \underline{\pi} P_{t-1|t-1}^{-1}$. Substituting this into the right-hand side of our inequality, we obtain:

$$
P_{t-1|t-1}^{-1} (P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t)^{-1} P_{t-1|t-1}^{-1} \succeq \left(\frac{\underline{\pi}}{\bar{\pi} + \frac{\bar{F}^2}{\underline{q}}} \right) P_{t-1|t-1}^{-1} \qquad \forall\, 1 \leq t \leq \tau \tag{103}
$$

At this stage, we define α as:

$$
\alpha := \frac{\underline{\pi}}{\bar{\pi} + \frac{\bar{F}^2}{\underline{q}}} \tag{104}
$$

Multiplying Eq. (103) by -1 reverses the bound and adding $P_{t-1|t-1}^{-1}$ to both sides yields:

$$
P_{t-1|t-1}^{-1} - P_{t-1|t-1}^{-1} (P_{t-1|t-1}^{-1} + F_t^\top \tilde{Q}_t^{-1} F_t)^{-1} P_{t-1|t-1}^{-1} \preceq (1-\alpha) P_{t-1|t-1}^{-1} \; \forall\, 1 \leq t \leq \tau \tag{105}
$$

Substituting this in Eq. (100) yields:

$$
E_t^\top P_{t|t}^{-1} E_t \preceq (1-\alpha) P_{t-1|t-1}^{-1} \qquad \forall\, 1 \leq t \leq \tau \tag{106}
$$

What remains to be proved is that $0 < \alpha < 1$. By definition, the scalar components $\underline{\pi}$, $\bar{\pi}$, $\bar{F}$, and $\underline{q}$ are all strictly positive and finite, which immediately guarantees that $\alpha > 0$. Furthermore, Lemma

L.12 establishes the bounds $\underline{\pi} I \preceq P_{t|t}^{-1} \preceq \bar{\pi} I$, which necessitates that $\underline{\pi} \leq \bar{\pi}$. Because the additive term $\frac{\bar{F}^2}{\underline{q}}$ is strictly greater than zero, the denominator of α must be strictly greater than its numerator ($\bar{\pi} + \frac{\bar{F}^2}{\underline{q}} > \underline{\pi}$). Therefore, α is strictly less than 1, satisfying $0 < \alpha < 1$ and concluding the proof. □

Proposition P.3 (Matrix Young's Inequality). *Let $a, b \in \mathbb{R}^n$ be arbitrary vectors, $M \in \mathbb{R}^{n \times n}$ be a positive-definite matrix. Then the following inequality holds for any $\eta > 0$:*

$$2a^\top M b \leq \eta a^\top M a + \frac{1}{\eta} b^\top M b \tag{107}$$

Proof. Let $c := \sqrt{\eta} a - \frac{1}{\sqrt{\eta}} b$. Since $M \succ 0$, we have:

$$\begin{aligned} & 0 \leq c^\top M c \\ \implies & 0 \leq (\sqrt{\eta} a - \frac{1}{\sqrt{\eta}} b)^\top M (\sqrt{\eta} a - \frac{1}{\sqrt{\eta}} b) \\ \implies & 0 \leq \eta a^\top M a - a^\top M b - b^\top M a + \frac{1}{\eta} b^\top M b \end{aligned} \tag{108}$$

Because M is positive-definite, it is also symmetric and thus, the inner products $a^\top M b$ and $b^\top M a$ are equal. Substituting this and rearranging the terms yields:

$$2a^\top M b \leq \eta a^\top M a + \frac{1}{\eta} b^\top M b \tag{109}$$

which completes the proof. □

Lemma L.15 (Implicit Lyapunov Dissipation). *Let $V_t(e_t) := e_t^\top P_{t|t}^{-1} e_t$ be the Lyapunov candidate for the ResEKF, as defined in Definition 7. Under the conditions of Lemma L.14, there exists a scalar $0 < \underline{\alpha} < \alpha < 1$ satisfying:*

$$V_t \leq (1 - \underline{\alpha}) V_{t-1} + \left(\frac{1}{\alpha - \underline{\alpha}} \right) \bar{\pi} \|q_t\|_2^2 \qquad \forall\, 1 \leq t \leq \tau \tag{110}$$

where $\bar{\pi}$ and α are scalars defined by Lemmas L.12 and L.14 respectively and $q_t := s_t + r_t + u_t$ is the total disturbance term of ResEKF's exact error dynamics (Eq. (18)), defined by Eqs. (19a)-(19d).

Proof. We begin by substituting ResEKF's exact error dynamics $e_t = E_t e_{t-1} + q_t$ into the Lyapunov function:

$$\begin{aligned} V_t &= (E_t e_{t-1} + q_t)^\top P_{t|t}^{-1} (E_t e_{t-1} + q_t) \\ &= e_{t-1}^\top E_t^\top P_{t|t}^{-1} E_t e_{t-1} + 2 e_{t-1}^\top E_t^\top P_{t|t}^{-1} q_t + q_t^\top P_{t|t}^{-1} q_t \qquad \forall\, 1 \leq t \leq \tau \end{aligned} \tag{111}$$

To decouple the cross-term on the right-hand side of Eq. (111), we use a Matrix version of Young's Inequality (Proposition P.3) with $a = E_t e_{t-1}$, $b = q_t$, and $M = P_{t|t}^{-1}$ to obtain the below for any scalar $\eta > 0$ and time $1 \leq t \leq \tau$:

$$2 e_{t-1}^\top E_t^\top P_{t|t}^{-1} q_t \leq \eta (e_{t-1}^\top E_t^\top P_{t|t}^{-1} E_t e_{t-1}) + \eta^{-1} (q_t^\top P_{t|t}^{-1} q_t) \tag{112}$$

Substituting this in Eq. (111), we have for any scalar $\eta > 0$:

$$V_t \leq (1+\eta) e_{t-1}^\top \underbrace{E_t^\top P_{t|t}^{-1} E_t}_{:=T_1} e_{t-1} + (1+\eta^{-1}) q_t^\top \underbrace{P_{t|t}^{-1}}_{:=T_2} q_t \qquad \forall\, 1 \leq t \leq \tau \tag{113}$$

Applying Lemmas L.14 and L.12 respectively to terms T_1 and T_2 above, we have for any scalar $\eta > 0$:

$$V_t \leq (1+\eta)(1-\alpha) V_{t-1} + (1+\eta^{-1}) \bar{\pi} \|q_t\|_2^2 \qquad \forall\, 1 \leq t \leq \tau \tag{114}$$

We proceed by restricting η to the interval $(0, \frac{\alpha}{1-\alpha})$, with $\alpha \in (0,1)$ (Lemma L.14) and by defining $\underline{\alpha}$ as follows:

$$\underline{\alpha} := \alpha - \eta(1-\alpha) \text{ with } 0 < \eta < \frac{\alpha}{1-\alpha} \tag{115}$$

It follows that the following three results hold:

$$\frac{1}{\eta(1-\alpha)} = \frac{1}{\alpha - \underline{\alpha}} \tag{116}$$

$$(1+\eta)(1-\alpha) = 1 - \underline{\alpha} \tag{117}$$

$$0 < \underline{\alpha} < \alpha < 1 \tag{118}$$

Equations (116) and (117) follow from rearranging Eq. (115), which gives $\eta(1-\alpha) = \alpha - \underline{\alpha} \implies \eta = \frac{\alpha - \underline{\alpha}}{1-\alpha}$. Equation (118) follows because the restriction $0 < \eta < \frac{\alpha}{1-\alpha}$ ensures $0 < \eta(1-\alpha) < \alpha$, which implies $\alpha > \alpha - \eta(1-\alpha) > 0$.

Now, since $\underline{\alpha} > 0$, it follows that $1 - \underline{\alpha} < 1$. Substituting Eq. (117), this yields $1 + \eta < \frac{1}{1-\alpha}$. Dividing both sides by the positive scalar η and using Eq. (116), we have:

$$\begin{aligned} 1 + \eta^{-1} &< \frac{1}{\eta(1-\alpha)} = \frac{1}{\alpha - \underline{\alpha}} \\ \implies (1+\eta^{-1}) \bar{\pi} \|q_t\|_2^2 &\leq \left(\frac{1}{\alpha - \underline{\alpha}} \right) \bar{\pi} \|q_t\|_2^2 \qquad \forall\, 1 \leq t \leq \tau \end{aligned} \tag{119}$$

where equality only occurs if $\|q_t\|_2^2 = 0$.

Substituting Eqs. (117) and (119) back in Eq. (114), we have:

$$V_t \leq (1 - \underline{\alpha}) V_{t-1} + \left(\frac{1}{\alpha - \underline{\alpha}} \right) \bar{\pi} \|q_t\|_2^2 \qquad \forall\, 1 \leq t \leq \tau \tag{120}$$

which, along with Eq. (118) completes the proof. □

Lemma L.16 (ResEKF Kalman Gain Bound)**.** *Under the conditions of Lemma L.11, there exists a finite, strictly positive scalar $\bar{k}$ such that the spectral norm of the Kalman gain K_t evaluated by Algorithm 4 (hereafter termed the Res-Kalman gain to distinguish it from the standard Kalman gain) satisfies:*

$$\|K_t\|_2 \leq \bar{k} := \frac{\bar{p} \bar{H}}{\underline{s}} \qquad \forall\, 1 \leq t \leq \tau \tag{121}$$

where $\bar{H}$ is the measurement Jacobian bound defined in Lemma L.4, $\bar{p}$ is the a priori *covariance upper bound defined in Lemma L.11, and $\underline{s}$ is the minimum eigenvalue of the measurement noise covariance matrix.*

Proof. From Algorithm 4, the Res-Kalman gain at time $t \geq 0$ is defined as:

$$K_t := P_{t|t-1} H_t^\top (H_t P_{t|t-1} H_t^\top + \Sigma_t)^{-1} \tag{122}$$

Applying the sub-multiplicative property of the spectral norm to Eq. (122) yields:

$$\|K_t\|_2 \leq \|P_{t|t-1}\|_2 \|H_t^\top\|_2 \|(H_t P_{t|t-1} H_t^\top + \Sigma_t)^{-1}\|_2 \tag{123}$$

We bound each of the three terms on the right-hand side of Eq. (123). The bounds $\|P_{t|t-1}\|_2 \leq \bar{p}$ and $\|H_t\|_2 \leq \bar{H}$ are made available by Lemmas L.11 and L.4 respectively.

As for the remaining term, we note that since $P_{t|t-1}$ is positive definite, the quadratic form $H_t P_{t|t-1} H_t^\top$ is positive semi-definite ($H_t P_{t|t-1} H_t^\top \succeq 0$). Adding the strictly positive definite measurement noise covariance Σ_t (Assumption 4) to both sides preserves the inequality:

$$H_t P_{t|t-1} H_t^\top + \Sigma_t \succeq \Sigma_t \succeq \underline{s} I \tag{124}$$

Inverting this reverses the spectral bound, providing an upper bound on the inverse:

$$(H_t P_{t|t-1} H_t^\top + \Sigma_t)^{-1} \preceq \frac{1}{\underline{s}} I \implies \|(H_t P_{t|t-1} H_t^\top + \Sigma_t)^{-1}\|_2 \leq \frac{1}{\underline{s}} \tag{125}$$

Substituting these three individual bounds back into Eq. (123) yields:

$$\|K_t\|_2 \leq \bar{p} \cdot \bar{H} \cdot \frac{1}{\underline{s}} = \frac{\bar{p}\bar{H}}{\underline{s}} \tag{126}$$

Defining $\bar{k} := \frac{\bar{p}\bar{H}}{\underline{s}}$ provides the required constructive bound. Because $\bar{p}$, $\bar{H}$, and $\underline{s}$ are all strictly positive and finite scalars, $\bar{k}$ is also a finite, strictly positive scalar, which completes the proof. □

Lemma L.17 (Expected Squared Noise Bound)**.** *Let r_t be the noise term of the exact ResEKF error dynamics, defined in Eq. (19c). Under the conditions of Lemma L.16, there exists a strictly positive, finite scalar ρ satisfying:*

$$\mathbb{E}[\|r_t\|_2^2] \leq \rho := (1 + \bar{k}\bar{H})^2 \epsilon_v^2 (1 + \bar{F})^2 + \bar{k}^2 d\bar{s} \qquad \forall\, 1 \leq t \leq \tau \tag{127}$$

where ϵ_v is the maximum deviation from deterministic trajectory (Assumption 3), $\bar{F}$ and $\bar{H}$ are the universal Jacobian bounds (Lemma L.2), $\bar{k}$ is the Res-Kalman gain bound (Lemma L.16), and $\bar{s}$ is the uniform upper-bound on Σ_t's largest eigenvalue (Assumption 4).

Proof. We first construct a deterministic upper bound for the process noise v_t. From the nominal model-based model (Eq. (2)) and the deterministic process model (Definition 1), we have:

$$\begin{aligned} & v_t = x_t - f(x_{t-1}) \quad \text{and} \quad \tilde{x}_t = f(\tilde{x}_{t-1}) \\ \implies\ & v_t = x_t - \tilde{x}_t + \tilde{x}_t - f(x_{t-1}) \\ & \quad = (x_t - \tilde{x}_t) + (f(\tilde{x}_{t-1}) - f(x_{t-1})) \end{aligned} \tag{128}$$

Taking the Euclidean norm and applying the triangle inequality yields:

$$\|v_t\|_2 \leq \|x_t - \tilde{x}_t\|_2 + \|f(\tilde{x}_{t-1}) - f(x_{t-1})\|_2 \tag{129}$$

By the Bounded State Assumption (Assumption 3), the nominal state x_t and deterministic solution $\tilde{x}_t$ both live in the compact domain $\mathcal{X}$ such that $\|x_t - \tilde{x}_t\|_2 \leq \epsilon_v$ for all $t \geq 0$. Furthermore,

by Assumption 1, f is continuously differentiable. Applying the Mean Value Theorem over the compact domain $\mathcal{X}$ and leveraging the universal transition Jacobian bound $\bar{F}$ from Lemma L.2:

$$\|f(\tilde{x}_{t-1}) - f(x_{t-1})\|_2 \leq \bar{F}\|\tilde{x}_{t-1} - x_{t-1}\|_2 \leq \bar{F}\epsilon_v \tag{130}$$

Substituting these back into Eq. (129), we explicitly and deterministically bound the process noise:

$$\|v_t\|_2 \leq \epsilon_v + \bar{F}\epsilon_v = \epsilon_v(1+\bar{F}) \qquad \forall\, t \geq 1 \tag{131}$$

Next, we bound the total noise term r_t. Taking the expected squared Euclidean norm of r_t introduces a cross-term:

$$\begin{aligned}
\mathbb{E}[\|r_t\|_2^2] &= \mathbb{E}[\|(I - K_tH_t)v_t - K_tw_t\|_2^2] \\
&= \mathbb{E}[((I - K_tH_t)v_t - K_tw_t)^\top((I - K_tH_t)v_t - K_tw_t)] \\
&= \mathbb{E}[\|(I - K_tH_t)v_t\|_2^2 - 2v_t^\top(I - K_tH_t)^\top K_tw_t + \|K_tw_t\|_2^2] \\
&= \mathbb{E}[\|(I - K_tH_t)v_t\|_2^2] - 2\mathbb{E}[v_t^\top(I - K_tH_t)^\top K_tw_t] + \mathbb{E}[\|K_tw_t\|_2^2]
\end{aligned} \tag{132}$$

By Assumption 4, the measurement noise w_t is zero-mean and strictly independent of the process noise v_t and the past measurement history (upon which K_t and H_t depend). Consequently, the expectation of the cross-term vanishes identically, reducing the expression to:

$$\mathbb{E}[\|r_t\|_2^2] = \mathbb{E}[\|(I - K_tH_t)v_t\|_2^2] + \mathbb{E}[\|K_tw_t\|_2^2] \tag{133}$$

For the first term, we apply the sub-multiplicative property, the spectral norm bounds $\|H_t\|_2 \leq \bar{H}$ (Lemma L.4) and $\|K_t\|_2 \leq \bar{k}$ (Lemma L.16), and the deterministic bound for v_t (Eq. (131)):

$$\begin{aligned}
\mathbb{E}[\|(I - K_tH_t)v_t\|_2^2] &\leq \mathbb{E}[(\|I - K_tH_t\|_2\|v_t\|_2)^2] \\
&\leq (1 + \bar{k}\bar{H})^2\epsilon_v^2(1+\bar{F})^2
\end{aligned} \tag{134}$$

For the second term, we rewrite the squared Euclidean norm using the matrix trace operator and the cyclic property of the trace:

$$\mathbb{E}[\|K_tw_t\|_2^2] = \mathbb{E}[\mathrm{Tr}(w_t^\top K_t^\top K_tw_t)] = \mathbb{E}[\mathrm{Tr}(K_t^\top K_tw_tw_t^\top)] \tag{135}$$

Because K_t is independent of w_t, we can pass the expectation to the noise outer product. By Assumption 4, $\mathbb{E}[w_tw_t^\top] = \Sigma_t$. Using the deterministic upper bound $\|K_t\|_2 \leq \bar{k}$, we have $K_t^\top K_t \preceq \bar{k}^2 I$. Applying this to the trace along with Eq. (13) yields:

$$\mathbb{E}[\mathrm{Tr}(K_t^\top K_t\Sigma_t)] \leq \bar{k}^2 d\bar{s} \tag{136}$$

Substituting Eq. (134) and Eq. (136) into Eq. (133), we obtain the final bound:

$$\mathbb{E}[\|r_t\|_2^2] \leq (1 + \bar{k}\bar{H})^2\epsilon_v^2(1+\bar{F})^2 + \bar{k}^2 d\bar{s} := \rho \qquad \forall\, 1 \leq t \leq \tau \tag{137}$$

Because $\bar{k}$, $\bar{H}$, ϵ_v, $\bar{F}$, and $\bar{s}$ are all finite, positive scalars, ρ is also a finite, strictly positive scalar, which concludes the proof. □

Lemma L.18 (Transition Linearisation Residue Bound)**.** *Let φ be the Taylor remainder of the transition function from Eq. (14). Further, let the conditions of Lemma L.16 hold. Provided the error bound $\epsilon_{\sup}$ in Hypothesis 1 satisfies the Transition Linearisation Boundedness (TLB) Condition:*

$$\textit{(TLB Condition)} \qquad \epsilon_{sup} \leq \epsilon_{\mathcal{X}} \tag{138}$$

there exists a finite, strictly positive scalar c_φ such that:

$$\|\varphi(x_{t-1}, \mu_{t-1|t-1})\|_2 \leq c_\varphi \epsilon_{sup}^2 \qquad \forall\, 1 \leq t \leq \tau \tag{139}$$

Proof. By Assumption 1, the transition function f is differentiable in $\mathbb{R}^s$, satisfying the conditions required for the Euclidean norm of residual of first-order Taylor series expansion to be bounded using (Dieudonné, 1969, Theorem 8.6.2). Setting the expansion point to $\mu_{t-1|t-1}$ and evaluating at $x_{t-1} = \mu_{t-1|t-1} + e_{t-1}$, the Euclidean norm of the residual $\varphi(x_{t-1}, \mu_{t-1|t-1})$ satisfies:

$$\|\varphi(x_{t-1}, \mu_{t-1|t-1})\|_2 \leq \|e_{t-1}\|_2 \sup_{\varrho \in [0,1]} \left\| \frac{\partial f}{\partial x}(\mu_{t-1|t-1} + \varrho e_{t-1}) - \frac{\partial f}{\partial x}(\mu_{t-1|t-1}) \right\|_2 \tag{140}$$

Now, by Assumption 3, $\mathcal{X}$ contains all points within distance $\epsilon_{\mathcal{X}}$ of the nominal trajectory. Further, by Hypothesis 1, the estimation error satisfies $\|e_{t-1}\|_2 \leq \epsilon_{\text{sup}}$, and by the lemma's hypothesis (Eq. (138)), $\epsilon_{\text{sup}} \leq \epsilon_{\mathcal{X}}$. Consequently, every point on the line segment $\{\mu_{t-1|t-1} + \varrho e_{t-1} : \varrho \in [0,1]\}$ lies within distance $\epsilon_{\mathcal{X}}$ of $x_{t-1} \in \mathcal{X}$, and is therefore contained within $\mathcal{X} \subseteq D(\mathcal{X}, f)$ for all $1 \leq t \leq \tau$.

Further, by Assumption 1, the Jacobian mapping $\frac{\partial f}{\partial x}$ is continuously differentiable in $\mathbb{R}^s$. We apply the Mean Value Inequality to this mapping over the line segment $\mu_{t-1|t-1} + \varrho e_{t-1}$ with parameter $\varrho \in [0,1]$ connecting the nominal point $\mu_{t-1|t-1}$ to the evaluation point $x_{t-1} = \mu_{t-1|t-1} + e_{t-1}$ to obtain:

$$\left\| \frac{\partial f}{\partial x}(\mu_{t-1|t-1} + \varrho e_{t-1}) - \frac{\partial f}{\partial x}(\mu_{t-1|t-1}) \right\|_2 \leq \left(\sup_{\varrho \in [0,1]} \left\| \frac{\partial^2 f}{\partial x^2}(\mu_{t-1|t-1} + \varrho e_{t-1}) \right\|_2 \right) \|e_{t-1}\|_2 \tag{141}$$

Substituting this in Eq. (140) yields:

$$\|\varphi(x_{t-1}, \mu_{t-1|t-1})\|_2 \leq \|e_{t-1}\|_2^2 \sup_{\varrho \in [0,1]} \left\| \frac{\partial^2 f}{\partial x^2}(\mu_{t-1|t-1} + \varrho e_{t-1}) \right\|_2 \tag{142}$$

Since, by Assumption 1, the Hessian $\frac{\partial^2 f}{\partial x^2}$ is also continuous in $\mathbb{R}^s$, it also attains a finite maximum in the compact operating domain $D(\mathcal{X}, f)$ (Extreme Value Theorem). Thus, there exists a strictly positive, finite scalar c_φ satisfying:

$$\sup_{\varrho \in [0,1]} \left\| \frac{\partial^2 f}{\partial x^2}(\mu_{t-1|t-1} + \varrho e_{t-1}) \right\|_2 \leq c_\varphi \qquad \forall\, 1 \leq t \leq \tau \tag{143}$$

Substituting this in Eq. (142), we have:

$$\|\varphi(x_{t-1}, \mu_{t-1|t-1})\|_2 \leq c_\varphi \|e_{t-1}\|_2^2 \qquad \forall\, 1 \leq t \leq \tau \tag{144}$$

By Hypothesis 1, the estimation error satisfies $\|e_{t-1}\|_2 \leq \epsilon_{\text{sup}}$. Substituting this into Eq. (144) completes the proof. □

Lemma L.19 (Measurement Linearisation Residue Bound)**.** *Let χ be the Taylor remainder of the measurement function from Eq. (14). Further, let the conditions of Lemma L.18 hold. Provided the error bound ϵ_{sup} in Hypothesis 1 and the constant c_φ from Lemma L.18 satisfy the Measurement Linearisation Bound (MLB) Condition below:*

$$\textit{(MLB Condition)} \qquad \bar{F}\epsilon_{sup} + c_\varphi \epsilon_{sup}^2 + \epsilon_v(1 + \bar{F}) + c_{\max} \leq \epsilon_{\mathcal{X}} \tag{145}$$

where $\epsilon_{\mathcal{X}}$ is the trajectory neighbourhood radius from Assumption 3, $c_{\max}$ is the filter's norm-bound, and $\bar{F}$ is the universal transition Jacobian bound from Lemma L.4, there exists a finite, strictly positive scalar c_χ such that:

$$\|\chi(x_t, \mu_{t|t-1})\|_2 \leq c_\chi(\bar{F}\epsilon_{sup} + c_\varphi\epsilon_{sup}^2 + \epsilon_v(1+\bar{F}) + c_{\max})^2 \qquad \forall\, 1 \leq t \leq \tau \tag{146}$$

Proof. The proof proceeds similarly to that of Lemma L.18. By Assumption 1, the measurement function h is twice continuously differentiable in $\mathbb{R}^s$, satisfying the conditions required for the Euclidean norm of the residual of its first-order Taylor series expansion to be bounded using (Dieudonné, 1969, Theorem 8.6.2). Setting the expansion point to $\mu_{t|t-1}$ and evaluating at $x_t = \mu_{t|t-1} + (x_t - \mu_{t|t-1})$, the trivial expansion of x_t involving $\mu_{t|t-1}$, the Euclidean norm of the residual $\chi(x_t, \mu_{t|t-1})$ satisfies:

$$\|\chi(x_t, \mu_{t|t-1})\|_2 \leq \|x_t - \mu_{t|t-1}\|_2 \sup_{\varrho\in[0,1]} \left\| \frac{\partial h}{\partial x}(\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1})) - \frac{\partial h}{\partial x}(\mu_{t|t-1}) \right\|_2 \tag{147}$$

We now establish that the line segment $\{\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1}) : \varrho \in [0,1]\}$ is contained within $\mathcal{X}$ for all $1 \leq t \leq \tau$. From the exact ResEKF a priori error dynamics (Eq. (15)), we have:

$$x_t - \mu_{t|t-1} = F_t e_{t-1} + \varphi(x_{t-1}, \mu_{t-1|t-1}) + v_t - \mu_{\text{BRC},\, t|t-1} \tag{148}$$

Applying the triangle inequality, the sub-multiplicative property of the spectral norm, Lemma L.4 ($\|F_t\|_2 \leq \bar{F}$), Lemma L.18 ($\|\varphi\|_2 \leq c_\varphi\epsilon_{\text{sup}}^2$), Eq. (131) ($\|v_t\|_2 \leq \epsilon_v(1+\bar{F})$), Eq. (1) ($\|\mu_{\text{BRC},\, t|t-1}\|_2 \leq c_{\max}$), and Hypothesis 1 ($\|e_{t-1}\|_2 \leq \epsilon_{\text{sup}}$), we obtain:

$$\|x_t - \mu_{t|t-1}\|_2 \leq \bar{F}\epsilon_{\text{sup}} + c_\varphi\epsilon_{\text{sup}}^2 + \epsilon_v(1+\bar{F}) + c_{\max} \qquad \forall\, 1 \leq t \leq \tau \tag{149}$$

By the lemma's hypothesis (Eq. (145)), the right-hand side is at most $\epsilon_{\mathcal{X}}$. Consequently, every point on the line segment $\{\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1}) : \varrho \in [0,1]\}$ lies within distance $\epsilon_{\mathcal{X}}$ of $x_t \in \mathcal{X}$. By Assumption 3, $\mathcal{X}$ contains all points within distance $\epsilon_{\mathcal{X}}$ of the nominal trajectory; therefore, the entire segment is contained within $\mathcal{X}$ for all $1 \leq t \leq \tau$.

Now, by Assumption 1, the Jacobian mapping $\frac{\partial h}{\partial x}$ is continuously differentiable in $\mathbb{R}^s$. Applying the Mean Value Inequality to this mapping over the line segment $\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1})$ with parameter $\varrho \in [0,1]$ connecting the nominal point $\mu_{t|t-1}$ to the evaluation point x_t yields:

$$\begin{aligned} &\left\| \frac{\partial h}{\partial x}(\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1})) - \frac{\partial h}{\partial x}(\mu_{t|t-1}) \right\|_2 \\ &\qquad \leq \left(\sup_{\varrho\in[0,1]} \left\| \frac{\partial^2 h}{\partial x^2}(\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1})) \right\|_2 \right) \|x_t - \mu_{t|t-1}\|_2 \end{aligned} \tag{150}$$

Substituting this into Eq. (147) yields:

$$\|\chi(x_t, \mu_{t|t-1})\|_2 \leq \|x_t - \mu_{t|t-1}\|_2^2 \sup_{\varrho\in[0,1]} \left\| \frac{\partial^2 h}{\partial x^2}(\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1})) \right\|_2 \tag{151}$$

Since, by Assumption 1 the Hessian $\frac{\partial^2 h}{\partial x^2}$ is continuous on $\mathbb{R}^s$, it attains a finite maximum on the compact set $\mathcal{X}$ by the Extreme Value Theorem. Thus, there exists a strictly positive, finite scalar c_χ satisfying:

$$\sup_{\varrho\in[0,1]} \left\| \frac{\partial^2 h}{\partial x^2}(\mu_{t|t-1} + \varrho(x_t - \mu_{t|t-1})) \right\|_2 \leq c_\chi \qquad \forall\, 1 \leq t \leq \tau \tag{152}$$

Substituting this into Eq. (151) and using the bound from Eq. (149) yields:

$$\|\chi(x_t, \mu_{t|t-1})\|_2 \leq c_\chi(\bar{F}\epsilon_{\text{sup}} + c_\varphi\epsilon_{\text{sup}}^2 + \epsilon_v(1+\bar{F}) + c_{\max})^2 \qquad \forall\, 1 \leq t \leq \tau \tag{153}$$

which concludes the proof. □

Lemma L.20 (Linearisation Residue Bound)**.** *Let s_t be the linearisation residue of the exact ResEKF error dynamics, as defined in Eq. (19b). Further, let the conditions of Lemma L.19 hold. Then there exists a finite, strictly positive scalar ζ satisfying:*

$$\|s_t\|_2 \leq \zeta \qquad \forall\, 1 \leq t \leq \tau \tag{154}$$

Proof. Taking the Euclidean norm of s_t and applying the triangle and the sub-multiplicative property of the induced spectral norm yields:

$$\|s_t\|_2 \leq (1 + \|K_t\|_2\|H_t\|_2)\|\varphi(x_{t-1}, \mu_{t-1|t-1})\|_2 + \|K_t\|_2\|\chi(x_t, \mu_{t|t-1})\|_2 \tag{155}$$

The bounds $\|H_t\|_2 \leq \bar{H}$ and $\|K_t\|_2 \leq \bar{k}$ are made available by Lemmas L.4 and L.16 respectively for all $1 \leq t \leq \tau$. Substituting these results along with the bounds on the Taylor remainders φ and χ from Lemmas L.18 and L.19 yields:

$$\|s_t\|_2 \leq (1 + \bar{k}\bar{H})c_\varphi\epsilon_{\text{sup}}^2 + \bar{k}c_\chi(\bar{F}\epsilon_{\text{sup}} + c_\varphi\epsilon_{\text{sup}}^2 + \epsilon_v(1 + \bar{F}) + c_{\max})^2 \qquad \forall\, 1 \leq t \leq \tau \tag{156}$$

Defining ζ as:

$$\zeta := (1 + \bar{k}\bar{H})c_\varphi\epsilon_{\text{sup}}^2 + \bar{k}c_\chi(\bar{F}\epsilon_{\text{sup}} + c_\varphi\epsilon_{\text{sup}}^2 + \epsilon_v(1 + \bar{F}) + c_{\max})^2 \tag{157}$$

and noticing that ζ is strictly positive and finite, since its components are all strictly positive and finite, concludes the proof. □

Lemma L.21 (Neural Feedback Bound)**.** *Let u_t be the neural feedback term of the exact ResEKF error dynamics, as defined in Eq. (19d). Under the conditions of Lemma L.16, there exists a strictly positive, finite scalar υ such that:*

$$\|u_t\|_2 \leq \upsilon := c_{\max}(2 + \bar{k}\bar{H}) \qquad \forall\, 1 \leq t \leq \tau \tag{158}$$

where $c_{\max}$ is the ResEKF's norm-bound, $\bar{H}$ is the measurement Jacobian bound from Lemma L.4, and $\bar{k}$ is the Res-Kalman gain bound from Lemma L.16.

Proof. Taking the Euclidean norm of u_t and applying the triangle and the sub-multiplicative property of the induced spectral norm yields:

$$\|u_t\|_2 \leq (1 + \|K_t\|_2\|H_t\|_2)\|\mu_{\text{BRC},\, t|t-1}\|_2 + \|\mu_{\text{BRC},\, t|t-1}\|_2 \tag{159}$$

By construction (Eq. (1)), we have $\|\mu_{\text{BRC},\, t|\cdot}\|_2 \leq c_{\max}$ and the bounds $\|H_t\|_2 \leq \bar{H}$ and $\|K_t\|_2 \leq \bar{k}$ are made available by Lemmas L.4 and L.16 respectively for all $1 \leq t \leq \tau$. Substituting these results in the above inequality yields:

$$\|u_t\|_2 \leq c_{\max}(1 + \bar{k}\bar{H}) + c_{\max} = c_{\max}(2 + \bar{k}\bar{H}) \qquad \forall\, 1 \leq t \leq \tau \tag{160}$$

Defining $\upsilon := c_{\max}(2 + \bar{k}\bar{H})$ provides the required bound. Because $c_{\max}$, $\bar{k}$, and $\bar{H}$ are all strictly positive, finite scalars, υ is also a finite, strictly positive scalar, which concludes the proof. □

Lemma L.22 (Expected Lyapunov Dissipation)**.** *Let $V_t(e_t)$ be the ResEKF Lyapunov candidate, as defined in Definition 7. Under the conditions of Lemma L.17 its expected value satisfies:*

$$\mathbb{E}[V_t(e_t)] \leq (1 - \underline{\alpha})\mathbb{E}[V_{t-1}(e_{t-1})] + \frac{3\bar{\pi}}{\alpha - \underline{\alpha}}(\zeta^2 + \rho + \upsilon^2) \qquad \forall\, 1 \leq t \leq \tau \tag{161}$$

where $\underline{\alpha} \in (0, 1)$ is as defined in Lemma L.15, $\bar{\pi}$ and α are as defined in Lemmas L.12 and L.14 respectively, and ζ, ρ, υ are the respective bounds for the linearisation residue, expected squared noise, and neural feedback (Lemmas L.17–L.21).

Proof. Taking expectation of the implicit Lyapunov dissipation inequality from Lemma L.15 yields:

$$\mathbb{E}[V_t(e_t)] \leq (1-\underline{\alpha})\mathbb{E}[V_{t-1}(e_{t-1})] + \left(\frac{\bar{\pi}}{\alpha-\underline{\alpha}}\right)\mathbb{E}[\|q_t\|_2^2] \qquad \forall\, 1 \leq t \leq \tau \tag{162}$$

where the total disturbance is $q_t := s_t + r_t + u_t$.

Applying Jensen's inequality $\left(\frac{\|a+b+c\|_2}{3}\right)^2 \leq \frac{\|a\|_2^2+\|b\|_2^2+\|c\|_2^2}{3}$ to q_t and using Lemmas L.17–L.21, we obtain:

$$\begin{aligned} \|q_t\|_2^2 &\leq 3(\|s_t\|_2^2 + \|r_t\|_2^2 + \|u_t\|_2^2) \\ \mathbb{E}[\|q_t\|_2^2] &\leq 3(\mathbb{E}[\|s_t\|_2^2] + \mathbb{E}[\|r_t\|_2^2] + \mathbb{E}[\|u_t\|_2^2]) = 3(\zeta^2 + \rho + \upsilon^2) \qquad \forall\, 1 \leq t \leq \tau \end{aligned} \tag{163}$$

Substituting this in Eq. (162) completes the proof, as we obtain:

$$\mathbb{E}[V_t(e_t)] \leq (1-\underline{\alpha})\mathbb{E}[V_{t-1}(e_{t-1})] + \frac{3\bar{\pi}}{\alpha-\underline{\alpha}}(\zeta^2 + \rho + \upsilon^2) \qquad \forall\, 1 \leq t \leq \tau \tag{164}$$

□

Lemma L.23 (Conditional Lyapunov Dissipation)**.** *Let $V_t(e_t)$ be the ResEKF Lyapunov candidate, as defined in Definition 7. Under the conditions of Lemma L.22, the conditional expectation of $V_t(e_t)$ satisfies:*

$$\mathbb{E}[V_t(e_t)|e_{t-1}] - V_{t-1}(e_{t-1}) \leq -\underline{\alpha}V_{t-1}(e_{t-1}) + \frac{3\bar{\pi}}{\alpha-\underline{\alpha}}(\zeta^2 + \rho + \upsilon^2) \qquad \forall\, 1 \leq t \leq \tau \tag{165}$$

where $0 < \underline{\alpha} < \alpha < 1$ and $\bar{\pi}$ are defined in Lemmas L.15 and L.12, respectively, and ζ, ρ and υ are the respective bounds for the linearisation residue, expected squared noise, and neural feedback (Lemmas L.17–L.21).

Proof. Applying conditional expectation on both sides of the implicit Lyapunov dissipation inequality from Lemma L.15 yields:

$$\mathbb{E}[V_t(e_t)|e_{t-1}] \leq (1-\underline{\alpha})V_{t-1}(e_{t-1}) + \left(\frac{\bar{\pi}}{\alpha-\underline{\alpha}}\right)\mathbb{E}[\|q_t\|_2^2|e_{t-1}] \qquad \forall\, 1 \leq t \leq \tau \tag{166}$$

where $q_t := s_t + r_t + u_t$. Substituting Eq. (163) and subtracting $V_{t-1}(e_{t-1})$ from both sides completes the proof. □

Lemma L.24 (Supermartingale Descent)**.** *Let ϵ_{sup} be the estimation error bound defined in Hypothesis 1. Let the total disturbance bound be denoted by $\mathcal{W}$, so that:*

$$\mathcal{W} := \frac{3\bar{\pi}}{\alpha-\underline{\alpha}}(\zeta^2 + \rho + \upsilon^2) \tag{167}$$

where $0 < \underline{\alpha} < \alpha < 1$ and $\bar{\pi}$ are defined in Lemmas L.15 and L.12, and ζ, ρ and υ are as defined in Lemmas L.17–L.21. Further, let the conditions of Lemma L.23 hold. Provided $\mathcal{W}$ satisfies:

$$\textit{(Supermartingale Descent Condition)} \qquad \mathcal{W} \leq \underline{\alpha}\underline{\pi}\xi^2\epsilon_{sup}^2 \qquad \forall\, \xi \in (0,1) \tag{168}$$

where $\underline{\pi}$ is as defined in Lemma L.12, the conditional expectation of the Lyapunov candidate $V_t(e_t)$ defined in Definition 7 satisfies:

$$\mathbb{E}[V_t(e_t)|e_{t-1}] - V_{t-1}(e_{t-1}) \leq 0 \qquad \forall\, t \in \{n \in \{1,\ldots,\tau\} : \|e_{n-1}\|_2 \geq \xi\epsilon_{sup}\} \tag{169}$$

Proof. From Lemma L.23, we have:

$$\mathbb{E}[V_t(e_t)|e_{t-1}] - V_{t-1}(e_{t-1}) \leq -\underline{\alpha} V_{t-1}(e_{t-1}) + \mathcal{W} \qquad \forall\, 1 \leq t \leq \tau \tag{170}$$

By Lemma L.13, the Lyapunov function is bounded from below by the squared Euclidean norm of the estimation error, yielding $V_{t-1}(e_{t-1}) \geq \underline{\pi}\|e_{t-1}\|_2^2$ for all $1 \leq t \leq \tau$. Thus, whenever $\xi\epsilon_{\text{sup}} \leq \|e_{n-1}\|_2 \leq \epsilon_{\text{sup}}$ holds for any given time step $1 \leq n \leq \tau$ and any $0 < \xi < 1$, it follows that:

$$V_{n-1}(e_{n-1}) \geq \underline{\pi}\xi^2\epsilon_{\text{sup}}^2 \implies -\underline{\alpha} V_{n-1}(e_{n-1}) \leq -\underline{\alpha\pi}\xi^2\epsilon_{\text{sup}}^2 \tag{171}$$

Substituting Eq. (171) into Eq. (170) yields:

$$\mathbb{E}[V_n(e_n)|e_{n-1}] - V_{n-1}(e_{n-1}) \leq -\underline{\alpha\pi}\xi^2\epsilon_{\text{sup}}^2 + \mathcal{W} \tag{172}$$

Applying the assumed condition $\mathcal{W} \leq \underline{\alpha\pi}\xi^2\epsilon_{\text{sup}}^2$ (Eq. (168)), concludes the proof. □

Proposition P.4 (Cubic Inequality)**.** *Let $a, b, c \in \mathbb{R}_{>0}$. The inequality*

$$ax + bx^3 \leq c \tag{173}$$

subject to $x > 0$ holds if and only if $x \leq \mathcal{R}(a, b, c)$, with:

$$\mathcal{R}(a,b,c) = \left(\sqrt[3]{\frac{c}{2b} + \sqrt{\frac{c^2}{4b^2} + \frac{a^3}{27b^3}}} + \sqrt[3]{\frac{c}{2b} - \sqrt{\frac{c^2}{4b^2} + \frac{a^3}{27b^3}}} \right) > 0 \tag{174}$$

Proof. Define the function $f : \mathbb{R}_{>0} \to \mathbb{R}$ as $f(x) := bx^3 + ax$. The first derivative is $f'(x) = 3bx^2 + a$. Because $a, b > 0$, $f'(x) > 0$ for all $x \geq 0$. Thus, $f(x)$ is strictly monotonically increasing on $[0, \infty)$.

Since $f(0) = 0$ and $\lim_{x\to\infty} f(x) = \infty$, there exists a unique real root $\mathcal{R}(a, b, c) > 0$ such that $f(\mathcal{R}(a, b, c)) = c$. Consequently, $f(x) \leq c$ holds for all $x \in (0, \mathcal{R}(a, b, c)]$.

Dividing the equality $f(\mathcal{R}(a, b, c)) = c$ by b and let $p := \frac{a}{b}$ and $q := -\frac{c}{b}$ yields:

$$x^3 + px + q = 0 \tag{175}$$

Observe that the quantity Δ defined below is positive since $a, b, c > 0$:

$$\Delta := \frac{q^2}{4} + \frac{p^3}{27} = \frac{c^2}{4b^2} + \frac{a^3}{27b^3} \tag{176}$$

Thus, we can apply Cardano's formula to obtain the unique real root:

$$\mathcal{R}(a,b,c) = \sqrt[3]{-\frac{q}{2} + \sqrt{\Delta}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\Delta}} \tag{177}$$

Substituting p, q, and Δ yields the stated formulation. □

Lemma L.25 (Sufficient Conditions for Stability)**.** *Let ϵ_{sup} be the estimation error bound defined in Hypothesis 1. Under the conditions of Lemma L.23, there exist finite, strictly positive scalars $\bar{\epsilon}$, $\bar{\epsilon}_v$, $\bar{c}$, and $\bar{\sigma}$ such that if the filter and its nominal model-based model have ϵ_{sup}, ϵ_v (the trajectory drift constant), $c_{\max}$ (the filter's norm-bound) and $\bar{s}$ (the measurement noise covariance's eigenvalue upper bound) satisfying:*

$$\epsilon_{sup} < \bar{\epsilon}, \qquad \epsilon_v < \bar{\epsilon}_v, \qquad c_{\max} < \bar{c} \qquad \textit{and} \qquad \bar{s} \leq \bar{\sigma} \tag{178}$$

then the Supermartingale Descent Condition, Eq. (168) holds.

Proof. Consider the linearisation residue bound, ζ from Eq. (157):

$$\zeta := (1 + \bar{k}\bar{H})c_\varphi \epsilon_{\text{sup}}^2 + \bar{k}c_\chi(\bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2 + \epsilon_v(1 + \bar{F}) + c_{\max})^2 \tag{179}$$

with $\bar{k}$ as defined in Lemma L.16, $\bar{F}$ and $\bar{H}$ as defined in Lemma L.2, c_χ and c_φ as defined in Lemma L.20 and $c_{\max}$ being ResEKF's norm-bound.

To isolate the parameters within the squared polynomial, we apply Jensen's inequality $(\sum_{i=1}^n x_i)^m \leq n^{m-1}(\sum_{i=1}^n x_i^m)$ by setting $x_1 := \bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2$, $x_2 := \epsilon_v(1 + \bar{F})$, $x_3 := c_{\max}$, $n = 3$, $m = 2$. This yields:

$$(\bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2 + \epsilon_v(1 + \bar{F}) + c_{\max})^2 \leq 3(\bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2)^2 + 3(1 + \bar{F})^2\epsilon_v^2 + 3c_{\max}^2 \tag{180}$$

Substituting this into Eq. (179) provides the following upper bound, through which we isolate the dependency of ζ on ϵ_{sup}, ϵ_v and $c_{\max}$:

$$\zeta \leq \zeta_\epsilon + \zeta_v + \zeta_c \tag{181}$$

where the terms comprising the bound are defined as:

$$\zeta_\epsilon := (1 + \bar{k}\bar{H})c_\varphi \epsilon_{\text{sup}}^2 + 3\bar{k}c_\chi(\bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2)^2 \tag{182a}$$

$$\zeta_v := 3\bar{k}c_\chi(1 + \bar{F})^2\epsilon_v^2 \tag{182b}$$

$$\zeta_c := 3\bar{k}c_\chi c_{\max}^2 \tag{182c}$$

Repeatedly applying the same Jensen's inequality to ζ_ϵ yields:

$$\begin{aligned}
\zeta_\epsilon^2 &\leq 2(1 + \bar{k}\bar{H})^2 c_\varphi^2 \epsilon_{\text{sup}}^4 + 18\bar{k}^2 c_\chi^2(\bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2)^4 \\
&\leq 2(1 + \bar{k}\bar{H})^2 c_\varphi^2 \epsilon_{\text{sup}}^4 + 18\bar{k}^2 c_\chi^2(8\bar{F}^4\epsilon_{\text{sup}}^4 + 8c_\varphi^4 \epsilon_{\text{sup}}^8) \\
&= \underbrace{2((1 + \bar{k}\bar{H})^2 c_\varphi^2 + 72\bar{k}^2 c_\chi^2 \bar{F}^4)\epsilon_{\text{sup}}^4 + 144\bar{k}^2 c_\chi^2 c_\varphi^4 \epsilon_{\text{sup}}^8}_{:=\bar{\zeta}_\epsilon^2}
\end{aligned} \tag{183}$$

Now applying the same Jensen's inequality to ζ using Eq. (181) gives:

$$\zeta^2 \leq (\zeta_\epsilon + \zeta_v + \zeta_c)^2 \leq 3(\zeta_\epsilon^2 + \zeta_v^2 + \zeta_c^2) = 3\bar{\zeta}_\epsilon^2 + 3\zeta_v^2 + 3\zeta_c^2 \tag{184}$$

Substituting Eqs. (137), (158) and (184) into the total disturbance bound $\mathcal{W}$ (Eq. (167)) yields:

$$\begin{aligned}
\mathcal{W} &:= \frac{3\bar{\pi}}{\alpha - \underline{\alpha}}(\zeta^2 + \rho + \upsilon^2) \\
&\leq \frac{3\bar{\pi}}{\alpha - \underline{\alpha}}\left(3\bar{\zeta}_\epsilon^2 + 3\zeta_v^2 + 3\zeta_c^2 + (1 + \bar{k}\bar{H})^2\epsilon_v^2(1 + \bar{F})^2 + \bar{k}^2 d\bar{s} + c_{\max}^2(2 + \bar{k}\bar{H})^2\right) \\
&= \underbrace{\frac{9\bar{\pi}\bar{\zeta}_\epsilon^2}{\alpha - \underline{\alpha}}}_{:=T_\epsilon} + \underbrace{\frac{3\bar{\pi}(3\zeta_v^2 + (1 + \bar{k}\bar{H})^2\epsilon_v^2(1 + \bar{F})^2)}{\alpha - \underline{\alpha}}}_{:=T_v} + \underbrace{\frac{3\bar{\pi}(3\zeta_c^2 + c_{\max}^2(2 + \bar{k}\bar{H})^2)}{\alpha - \underline{\alpha}}}_{:=T_c} + \underbrace{\frac{3\bar{\pi}\bar{k}^2 d\bar{s}}{\alpha - \underline{\alpha}}}_{:=T_\Sigma}
\end{aligned} \tag{185}$$

where $\bar{\pi}$ is as defined in Lemma L.12, and $0 < \underline{\alpha} < \alpha < 1$ are as defined in Lemma L.15.

Now consider the following hypotheses involving the above-mentioned terms:

$$(\epsilon \textbf{ Hypothesis}) \qquad T_\epsilon \leq w_\epsilon \underline{\alpha\pi} \xi^2 \epsilon_{\text{sup}}^2 \tag{186a}$$

$$(\epsilon_v \textbf{ Hypothesis}) \qquad T_v \leq w_v \underline{\alpha\pi} \xi^2 \epsilon_{\text{sup}}^2 \tag{186b}$$

$$(c \textbf{ Hypothesis}) \qquad T_c \leq w_c \underline{\alpha\pi} \xi^2 \epsilon_{\text{sup}}^2 \tag{186c}$$

$$(\Sigma \textbf{ Hypothesis}) \qquad T_\Sigma \leq w_\Sigma \underline{\alpha\pi} \xi^2 \epsilon_{\text{sup}}^2 \tag{186d}$$

where $\underline{\pi}$ is as defined in Lemma L.12, ξ may be any scalar satisfying $0 < \xi < 1$ and $\{w_\epsilon, w_v, w_c, w_\Sigma\}$ may be any set of scalars satisfying $w_\epsilon, w_v, w_c, w_\Sigma \geq 0$ and $w_\epsilon + w_v + w_c + w_\Sigma = 1$.

Validating the ϵ Hypothesis: Using Eqs. (185) and (183) and the fact that $\epsilon_{\text{sup}} > 0$, we can observe that the ϵ Hypothesis (186a) holds if:

$$\begin{aligned} &\frac{9\bar{\pi}(2((1+\bar{k}\bar{H})^2 c_\varphi^2 + 72\bar{k}^2 c_\chi^2 \bar{F}^4)\epsilon_{\text{sup}}^4 + 144\bar{k}^2 c_\chi^2 c_\varphi^4 \epsilon_{\text{sup}}^8)}{\alpha - \underline{\alpha}} \leq w_\epsilon \underline{\alpha\pi} \xi^2 \epsilon_{\text{sup}}^2 \\ \iff &((1+\bar{k}\bar{H})^2 c_\varphi^2 + 72\bar{k}^2 c_\chi^2 \bar{F}^4)\epsilon_{\text{sup}}^2 + 72\bar{k}^2 c_\chi^2 c_\varphi^4 \epsilon_{\text{sup}}^6 \leq \frac{w_\epsilon \underline{\alpha}(\alpha - \underline{\alpha})\underline{\pi}\xi^2}{18\bar{\pi}} \end{aligned} \tag{187}$$

Applying Proposition P.4 to Eq. (187) with $a = (1+\bar{k}\bar{H})^2 c_\varphi^2 + 72\bar{k}^2 c_\chi^2 \bar{F}^4 > 0$, $b = 72\bar{k}^2 c_\chi^2 c_\varphi^4 > 0$, $c = \frac{w_\epsilon \underline{\alpha}(\alpha - \underline{\alpha})\underline{\pi}\xi^2}{18\bar{\pi}} > 0$ and $\epsilon_{\text{sup}} > 0$ yields:

$$0 < \epsilon_{\text{sup}}^2 \leq \mathcal{R}\left((1+\bar{k}\bar{H})^2 c_\varphi^2 + 72\bar{k}^2 c_\chi^2 \bar{F}^4, 72\bar{k}^2 c_\chi^2 c_\varphi^4, \frac{w_\epsilon \underline{\alpha}(\alpha - \underline{\alpha})\underline{\pi}\xi^2}{18\bar{\pi}}\right) \tag{188}$$

Thus, the ϵ Hypothesis (186a) holds if $\epsilon_{\text{sup}} \leq \tilde{\epsilon}$, where $\tilde{\epsilon}$ is given by:

$$\tilde{\epsilon} := \sqrt{\mathcal{R}\left((1+\bar{k}\bar{H})^2 c_\varphi^2 + 72\bar{k}^2 c_\chi^2 \bar{F}^4, 72\bar{k}^2 c_\chi^2 c_\varphi^4, \frac{w_\epsilon \underline{\alpha}(\alpha - \underline{\alpha})\underline{\pi}\xi^2}{18\bar{\pi}}\right)} \tag{189}$$

Validating the ϵ_v Hypothesis: Using Eqs. (185) and (182b), we can observe that the ϵ_v Hypothesis (186b) holds if:

$$\begin{aligned} &\frac{3\bar{\pi}(27(\bar{k}c_\chi(1+\bar{F})^2\epsilon_v^2)^2 + (1+\bar{k}\bar{H})^2\epsilon_v^2(1+\bar{F})^2)}{\alpha - \underline{\alpha}} \leq w_v \underline{\alpha\pi} \xi^2 \epsilon_{\text{sup}}^2 \\ \iff &27\bar{k}^2 c_\chi^2 (1+\bar{F})^4 \epsilon_v^4 + (1+\bar{k}\bar{H})^2(1+\bar{F})^2 \epsilon_v^2 \leq \frac{w_v \underline{\alpha}(\alpha - \underline{\alpha})\underline{\pi}\xi^2 \epsilon_{\text{sup}}^2}{3\bar{\pi}} \end{aligned} \tag{190}$$

This is a quadratic inequality in ϵ_v^2, which, since $\epsilon_v > 0$ by definition, is equivalent to:

$$\epsilon_v^2 \leq \frac{-B_v + \sqrt{B_v^2 + 4A_v C_v}}{2A_v} \tag{191}$$

where the strictly positive scalars A_v, B_v and C_v are defined as:

$$A_v := 27\bar{k}^2 c_\chi^2 (1+\bar{F})^4 \tag{192a}$$

$$B_v := (1+\bar{k}\bar{H})^2 (1+\bar{F})^2 \tag{192b}$$

$$C_v := \frac{w_v \underline{\alpha}(\alpha - \underline{\alpha})\underline{\pi}\xi^2 \epsilon_{\text{sup}}^2}{3\bar{\pi}} \tag{192c}$$

Thus, the ϵ_v Hypothesis (186b) holds if $\epsilon_v \leq \tilde{\epsilon}_v$, where $\tilde{\epsilon}_v$ is given by:

$$\tilde{\epsilon}_v := \sqrt{\frac{-B_v + \sqrt{B_v^2 + 4A_vC_v}}{2A_v}} \tag{193}$$

Validating the c Hypothesis: Using Eqs. (185) and (182c), we can observe that the c Hypothesis (186c) holds if:

$$\begin{aligned} &\frac{3\bar{\pi}(27\bar{k}^2c_\chi^2c_{\max}^4 + (2+\bar{k}\bar{H})^2c_{\max}^2)}{\alpha - \underline{\alpha}} \leq w_c\underline{\alpha}\underline{\pi}\xi^2\epsilon_{\text{sup}}^2 \\ \iff &27\bar{k}^2c_\chi^2c_{\max}^4 + (2+\bar{k}\bar{H})^2c_{\max}^2 \leq \frac{w_c\underline{\alpha}(\alpha-\underline{\alpha})\underline{\pi}\xi^2\epsilon_{\text{sup}}^2}{3\bar{\pi}} \end{aligned} \tag{194}$$

As with ϵ_v, the above is a quadratic inequality in $c_{\max}^2$, which, since $c_{\max} > 0$ by definition, is equivalent to:

$$c_{\max}^2 \leq \frac{-B_c + \sqrt{B_c^2 + 4A_cC_c}}{2A_c} \tag{195}$$

where the strictly positive scalars A_c, B_c and C_c are defined as:

$$A_c := 27\bar{k}^2c_\chi^2 \tag{196a}$$

$$B_c := (2+\bar{k}\bar{H})^2 \tag{196b}$$

$$C_c := \frac{w_c\underline{\alpha}(\alpha-\underline{\alpha})\underline{\pi}\xi^2\epsilon_{\text{sup}}^2}{3\bar{\pi}} \tag{196c}$$

Thus, the c Hypothesis (186c) holds if $c_{\max} \leq \tilde{c}$, where $\tilde{c}$ is given by:

$$\tilde{c} := \sqrt{\frac{-B_c + \sqrt{B_c^2 + 4A_cC_c}}{2A_c}} \tag{197}$$

Satisfying the Covariance Boundedness Condition: Notice that one of the conditions of the lemma (through its dependence on Lemma L.23, which ultimately depends on Lemma L.11), is the Covariance Boundedness Condition (Eq. (61)).

To satisfy this, let ϑ_ϵ, ϑ_v and ϑ_c be any scalars satisfying $\vartheta_\epsilon, \vartheta_v, \vartheta_c > 0$ and $\vartheta_\epsilon + \vartheta_v + \vartheta_c = 1$. Then consider the following hypotheses involving terms that are all strictly positive:

$$\mathcal{L}(s)\epsilon_{\text{sup}} < \vartheta_\epsilon\mathcal{M}(s), \quad \mathcal{L}(s)\epsilon_v < \vartheta_v\mathcal{M}(s), \quad \text{and} \quad \mathcal{C}(s)c_{\max} < \vartheta_c\mathcal{M}(s) \tag{198}$$

If valid, these hypotheses lead to the satisfaction of the Covariance Boundedness Condition, since $\mathcal{L}(s)(\epsilon_{\text{sup}} + \epsilon_v) + \mathcal{C}(s)c_{\max} < (\vartheta_\epsilon + \vartheta_v + \vartheta_c)\mathcal{M}(s) = \mathcal{M}(s)$. Thus we have the following criteria for satisfying the Covariance Boundedness Condition:

$$\epsilon_{\text{sup}} < \vartheta_\epsilon\frac{\mathcal{M}(s)}{\mathcal{L}(s)}, \quad \epsilon_v < \vartheta_v\frac{\mathcal{M}(s)}{\mathcal{L}(s)}, \quad \text{and} \quad c_{\max} < \vartheta_c\frac{\mathcal{M}(s)}{\mathcal{C}(s)} \tag{199}$$

Satisfying the Linearisation Boundedness Conditions: Similarly, two other conditions on the Lemma, through its dependence on Lemma L.20 are the Transition and Measurement Linearisation Boundedness Conditions, Eqs. (138) and (145). The former is an atomic condition that $\epsilon_{\text{sup}} \leq \epsilon_{\mathcal{X}}$,

whereas the latter is a compound condition on $\epsilon_{\text{sup}}, \epsilon_v$ and $c_{\max}$. To handle the compound condition, let $\digamma_\epsilon, \digamma_v$ and $\digamma_c$ be any scalars satisfying $\digamma_\epsilon, \digamma_v, \digamma_c > 0$ and $\digamma_\epsilon + \digamma_v + \digamma_c = 1$. Then consider the following hypotheses involving terms that are all strictly positive:

$$\bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2 \leq \digamma_\epsilon \epsilon_{\mathcal{X}}, \quad \epsilon_v(1+\bar{F}) \leq \digamma_v \epsilon_{\mathcal{X}}, \quad \text{and} \quad c_{\max} \leq \digamma_c \epsilon_{\mathcal{X}} \tag{200}$$

If valid, these hypotheses lead to the satisfaction of the Measurement Linearisation Boundedness Condition, since $\bar{F}\epsilon_{\text{sup}} + c_\varphi \epsilon_{\text{sup}}^2 + \epsilon_v(1+\bar{F}) + c_{\max} \leq (\digamma_\epsilon + \digamma_v + \digamma_c)\epsilon_{\mathcal{X}}$. Adding on the atomic Transition Linearisation Boundedness Condition $\epsilon_{\text{sup}} \leq \epsilon_{\mathcal{X}}$, we have the following criteria for satisfying both the Linearisation Boundedness Conditions:

$$\epsilon_{\text{sup}} \leq \check{\epsilon}, \quad \epsilon_v \leq \frac{\digamma_v \epsilon_{\mathcal{X}}}{1+\bar{F}}, \quad \text{and} \quad c_{\max} \leq \digamma_c \epsilon_{\mathcal{X}} \tag{201}$$

where:

$$\check{\epsilon} = \min\left(\epsilon_{\mathcal{X}}, \frac{\sqrt{\bar{F}^2 + 4c_\varphi \digamma_\epsilon \epsilon_{\mathcal{X}}} - \bar{F}}{2c_\varphi}\right) \tag{202}$$

Sufficient Conditions on $\epsilon_{\text{sup}}, \epsilon_v$ and $c_{\max}$: Observe that the ϵ, ϵ_v and c Hypotheses (Eqs. (186a)–(186c)) are valid and the Covariance Boundedness as well as both the Linear Boundedness Conditions are simultaneously satisfied if $\epsilon_{\text{sup}} < \bar{\epsilon}$, $\epsilon_v < \bar{\epsilon}_v$ and $c_{\max} < \bar{c}$ with the right-hand sides defined as:

$$\bar{\epsilon} := \min\left(\tilde{\epsilon}, \vartheta_\epsilon \frac{\mathcal{M}(s)}{\mathcal{L}(s)}, \check{\epsilon}\right), \ \bar{\epsilon}_v := \min\left(\tilde{\epsilon}_v, \vartheta_v \frac{\mathcal{M}(s)}{\mathcal{L}(s)}, \frac{\digamma_v \epsilon_{\mathcal{X}}}{1+\bar{F}}\right) \text{ and } \bar{c} := \min\left(\tilde{c}, \vartheta_c \frac{\mathcal{M}(s)}{\mathcal{C}(s)}, \digamma_c \epsilon_{\mathcal{X}}\right) \tag{203}$$

Validating the Σ Hypothesis: Finally, using Eq. (185), we can observe that the Σ Hypothesis (186d) holds if:

$$\frac{3\bar{\pi}\bar{k}^2 d\bar{s}}{\alpha - \underline{\alpha}} \leq w_\Sigma \underline{\alpha}\underline{\pi}\xi^2 \epsilon_{\text{sup}}^2 \tag{204}$$

Thus, the Σ Hypothesis (186d) holds if $\bar{s} \leq \bar{\sigma}$, where $\bar{\sigma}$ is given by:

$$\bar{\sigma} := \frac{w_\Sigma \underline{\alpha}(\alpha - \underline{\alpha})\underline{\pi}\xi^2 \epsilon_{\text{sup}}^2}{3\bar{\pi}\bar{k}^2 d} \tag{205}$$

Conclusion: Thus, provided the system satisfies the bounds $\epsilon_{\text{sup}} < \bar{\epsilon}$, $\epsilon_v < \bar{\epsilon}_v$, $c_{\max} < \bar{c}$, and $\bar{s} \leq \bar{\sigma}$, the hypotheses in Eqs. (186a) through (186d) are valid and the Covariance Boundedness Condition (Eq. (61)) is simultaneously satisfied. Substituting the above mentioned valid hypotheses back into the total disturbance bound defined in Eq. (185) yields:

$$\mathcal{W} \leq (w_\epsilon + w_v + w_c + w_\Sigma)\underline{\alpha}\underline{\pi}\xi^2 \epsilon_{\text{sup}}^2 \tag{206}$$

Because the set of scalars $\{w_\epsilon, w_v, w_c, w_\Sigma\}$ form a simplex such that $w_\epsilon + w_v + w_c + w_\Sigma = 1$, this simplifies exactly to $\mathcal{W} \leq \underline{\alpha}\underline{\pi}\xi^2 \epsilon_{\text{sup}}^2$. Therefore, the Supermartingale Descent Condition in Eq. (168) holds, concluding the proof. □

I.3 Final Proof

Proof. The proof proceeds by mathematical induction. The base case for the induction is that the initial error satisfies $\|e_0\|_2 < \bar{\epsilon}$, but this is a express condition of the theorem, and thus holds.

Under the inductive hypothesis, for time $1 \leq t < \tau$, Lemmas L.1–L.25 hold. However, whereas the theorem only requires the initial error to satisfy $\|e_0\|_2 < \bar{\epsilon}$, Lemma L.25 requires $\epsilon_{\text{sup}} < \bar{\epsilon}$ in order for Eq. (169) to be a supermartingale. To resolve this, we proceed as Reif et al. (1999) do. They too require only the 'smallness' of initial errors to impose a uniform bound such as ϵ_{sup} on the infinite-horizon error process (i.e., for all time). Specifically:

1. In Reif et al. (1999, p. 721-22), they show that linearisation residues (LR) are bounded under the assumption that the filter estimate ($\hat{z}$) is bounded; our inductive hypothesis upto time $\tau - 1$, which, due to Lemma L.3 and other conditions of the theorem, allow us to do bound linearisation residues via Lemmas L.18–L.20.

2. Then in Reif et al. (1999, p. 722), they proceed as follows: "we have to show that the boundedness with probability one of ζ_n and z_n implies the desired bounds on A_n, C_n and P_n. Then we obtain from Theorem 3.1 the boundedness of ζ_{n+1}, i.e., for the next time step. Repeating this procedure we get bounds on and therefore on $A_{n+1}, C_{n+1}, P_{n+1}$ and therefore on ζ_{n+2}". In their work, z_n is the state at time n, ζ_n, the filter's estimation error, A_n, C_n the transition and measurement Jacobian matrices and P_n the state covariance estimate. We show that provided $e_{\tau-1}$ is bounded (inductive hypothesis), Jacobian matrices F_τ and H_τ are bounded as are the covariance estimates $P_{\tau|\tau}$ via Lemmas L.4 and L.12.

3. Then within Reif et al. (1999, Theorem 3.1), they first show that the Lyapunov candidate is bounded; we show this via Lemma L.13.

4. They then bound the conditional expectation of the forward difference of the Lyapunov function ('CELF') in Reif et al. (1999, Eq. (77)); we perform this via Lemma L.23.

5. They then note that the above two bounds (Lyapunov candidate and CELF bound) are conditional on the error ζ_n at any given time n being not merely bounded but *uniformly* small Reif et al. (1999, Eqs. (78)-(79)). Thus, having assumed only the smallness of the initial error, they contend with the same problem as we mention above. To resolve this, they proceed as follows:

   - They show that if (1) ϵ, which importantly is their initial error threshold, and (2) the maximum eigenvalue of the measurement noise covariance δ are both small Reif et al. (1999, Eqs. (78) and (81)), the CELF becomes non-positive Reif et al. (1999, Eq. (80)) in the annulus $\tilde{\epsilon} \leq \|\zeta\|_2 \leq \epsilon$.
   - This result applies to systems with no process noise. However, they remark (Reif et al., 1999, Pg. 722) that that an additional condition on the process drift bound ϵ_w (which we denote by ϵ_v) makes Reif et al. (1999, Eq. (80)) valid for systems with process noise.
   - Then, invoking the stopping time theorem, Gard (1988, Theorem 5.2), they imply probability-one uniform boundedness of the infinite-horizon stochastic error process, i.e., $\zeta_n \leq \epsilon$, where again, ϵ is the initial error threshold, for all time $n \geq 0$. Crucially, this is essential for the CELF to become non-positive, i.e. a supermartingale.
   - Having satisfied all the prerequisites of Reif et al. (1999), we also invoke the stopping time theorem to enable us to use Lemma L.25 under smallness of the initial error e_0, the trajectory drift constant ϵ_v, the eigenvalue bound on the measurement noise covariances $\bar{s}$, and, additionally, the norm-bound $c_{\max}$ to establish a probability-one uniform bound ϵ_{sup} on the ResEKF stochastic error process e_t over the infinite horizon.

We thus successfully close the induction, granting us time-independent validity of all the Lemmas L.1–L.25. To derive explicit algebraic forms of the functions β, γ and δ in Eq. (3), we proceed as follows. From Eq. (161), the expected Lyapunov dissipation is bounded by:

$$\mathbb{E}[V_t(e_t)] \leq (1 - \underline{\alpha})\mathbb{E}[V_{t-1}(e_{t-1})] + \mathcal{W} \tag{207}$$

Unrolling this recursive inequality from $t = 1$ to an arbitrary time t and using $\mathbb{E}[V_0(e_0)] \leq V_0(e_0)$ yields the geometric series:

$$\begin{aligned}\mathbb{E}[V_t(e_t)] &\leq (1-\underline{\alpha})^t V_0(e_0) + \sum_{i=0}^{t-1}(1-\underline{\alpha})^i \mathcal{W} \\ &\leq (1-\underline{\alpha})^t V_0(e_0) + \frac{1-(1-\underline{\alpha})^t}{\underline{\alpha}}\mathcal{W} \\ &\leq (1-\underline{\alpha})^t V_0(e_0) + \frac{\mathcal{W}}{\underline{\alpha}} \end{aligned} \tag{208}$$

From Lemma L.13, the Lyapunov function is bounded by the Rayleigh quotient $\underline{\pi}\|e_t\|_2^2 \leq V_t(e_t) \leq \bar{\pi}\|e_t\|_2^2$. Taking the expectation and substituting these bounds into the unrolled inequality provides the mean-square error bound:

$$\underline{\pi}\mathbb{E}[\|e_t\|_2^2] \leq (1-\underline{\alpha})^t \bar{\pi}\|e_0\|_2^2 + \frac{\mathcal{W}}{\underline{\alpha}} \tag{209}$$

Dividing by $\underline{\pi}$ and recalling the decomposition of $\mathcal{W} \leq T_\epsilon + T_v + T_c + T_\Sigma$ from Eq. (185), we obtain the final bound of Eq. (3):

$$\mathbb{E}[\|e_t\|_2^2] \leq \underbrace{\frac{\bar{\pi}}{\underline{\pi}}(1-\underline{\alpha})^t\|e_0\|_2^2}_{:=\beta(\|e_0\|_2^2,t)} + \underbrace{\frac{T_c}{\underline{\alpha}\underline{\pi}}}_{:=\gamma(c_{\max})} + \underbrace{\frac{T_\epsilon + T_v + T_\Sigma}{\underline{\alpha}\underline{\pi}}}_{:=\delta(c_{\max})} \tag{210}$$

Notice that $\beta(\|e_0\|_2^2, t)$ is a Class $\mathcal{KL}$ function since it is continuous, strictly increasing with respect to the initial squared error norm $\|e_0\|_2^2$ for any given t, with $\beta(0,t) = 0$, and exponentially and monotonically decays to zero as $t \to \infty$ (due to $0 < \underline{\alpha} < 1$), for any given $\|e_0\|_2^2$. Similarly, $\gamma(c_{\max})$ is a Class $\mathcal{K}$ function; substituting T_c from Eqs. (185) and (182c) yields $\gamma(c_{\max}) = \frac{3\bar{\pi}}{\underline{\alpha}\underline{\pi}(\alpha-\underline{\alpha})}(27\bar{k}^2 c_\chi^2 c_{\max}^4 + (2+\bar{k}\bar{H})^2 c_{\max}^2)$, which is continuous, strictly increasing for $c_{\max} \geq 0$, and satisfies $\gamma(0) = 0$. This concludes the proof.

□

I.4 Further Remarks

Remark 3 (Necessity of Norm-Bounding and Sufficiency of its Smallness)**.** *Theorem 1 establishes the existence of an admissible class of neural non-linearities that may improve filtering under model misspecification without compromising EKF stability Reif et al. (1999). ResEKF's region of attraction is characterised by a fundamental trade-off between the systemic upper bounds ($\bar{\epsilon}$, $\bar{\epsilon}_v$ and $\bar{\sigma}$), and that on the filter's norm-bound ($\bar{c}$), as shown in the proof of Lemma L.25. While ResEKF's neural feedback enables compensation for nominal model misspecifications, it necessarily consumes a portion of the purely model-based EKF's stability margin, concretising the necessity for norm-bounding learned residuals in order to guarantee stability. Since the theorem's only additional condition is the smallness of the norm-bound $c_{\max}$, our sufficiency claim also stands proven.*

Remark 4 (Conservative Bounds)**.** *The constraints $\bar{\epsilon}, \bar{\epsilon}_v, \bar{c}$ and $\bar{\sigma}$ of Theorem 1 are derived via worst-case Lyapunov analysis and may be, as Reif et al. (1999) remark, very conservative. In practice, ResEKFs may tolerate larger initial errors and system disturbances than Theorem 1 suggests.*

Remark 5 (Practical Relevance: Instability Diagnosis and Remedial Actions)**.** *Our simulated experiments §4.2–§4.3 underscore the possibility of closing the model misspecification gaps. However, perhaps in the same way as we must content ourselves with non-exact inference in nonlinear systems, in the real-world, we must rely on Theorem 1's non-destabilisation implication: BRC's neural backbones cannot endogenously disrupt EKF stability when their norm-bound $c_{\max}$ is controlled. This uniquely affords practical diagnostic and remedial utility absent in other hybrid filters. Should ResEKF diverge in a real-world context, one can evaluate the underlying EKF: if this EKF remains stable, the initial divergence stems from large neural feedback and ResEKF can be stabilised simply*

by reducing $c_{\max}$—precisely the remedy other hybrid filters lack. Otherwise, the instability is exogenous, necessitating classical remedies that practitioners already have at their disposal, such as reducing initial error, tuning noise covariances, collecting higher-fidelity measurements, or refining the nominal model itself. Indeed, although the formal theorem precludes model misspecification, all our empirical evaluations feature misspecifications where the nominal filters remained stable and BRC-augmented variations with controlled norm-bounds successfully preserved nominal model stability, while competing hybrid filters could not always.

Remark 6 (Diagnostic vs. Prospective Utility)**.** *The conditions of Theorem 1 depend on system parameters (such as the eigenvalue of measurement noise) that in practice may not, in general, be verifiable in advance. This is a well-known issue affecting nonlinear stability theorems in general (see e.g. Barrau & Bonnabel (2017)). They are, however, useful in diagnostic instability when it does occur.*

Remark 7 (Joint Optimisation)**.** *If noise covariances and initial states ($Q_t, \Sigma_t, \mu_{0|0}, P_{0|0}$) must be learned jointly with neural parameters, we recommend pre-tuning the underlying EKF to a stable configuration. This provides a robust initialisation for joint end-to-end optimisation, provided $c_{\max}$ is appropriately bounded. However, if Σ is also learned, setting $c_{\max}$ too small, which mutes neural corrections, risks divergence as the joint optimiser may inflate Σ_t to compensate for the aforementioned muting.*

Remark 8 (Function Classes and Arguments)**.** *References to Class $\mathcal{KL}$ and $\mathcal{K}$ are in the proof of Theorem 1 are not essential; however, they are common in stability literature and definitions of these function classes can be found, for instance, in (Sontag, 2008). We also note that whereas the three functions β, γ and δ depend on the system's own constant parameters, in our presentation we highlight their respective dependence on e_0 and $c_{\max}$.*

J Stability Guarantee for EKF Extensions

J.1 Measurement Noise Covariance Inflation Extensions

Under linearisation, the model-based filters RORF (Piché et al., 2012) and WoLF (Duran-Martin et al., 2024) operate exactly as EKFs, except they inflate the EKF's measurement noise covariance Σ_t adaptively. Natively, the EKF allows for adaptive 'tuning' of noise covariances, but the stochastic stability of (Reif et al., 1999) is conditional on the maximum eigenvalue of the measurement noise covariance being small ($\bar{s} < \bar{\sigma}$); i.e., the EKF cannot tolerate unbounded covariance inflation if it must be stable.

Thus, provided the covariance inflation mechanisms of RORF and WoLF are bounded so that resulting inflated measurement noise covariances satisfy the stability margin constraint, they shall remain stable under standard regularity conditions. Practically, this can be achieved by clamping the scalar factor inflating the covariances. The argument for the preservation of nominal stability for BRC-augmented versions of the stabilised RORF and WoLF filters is identical to the ResEKF case.

J.2 State Covariance Inflation Extensions

EKF extensions such as AREKF (Xiong et al., 2009) (as well as predecessors such as REKF (Einicke & White, 1999)) operate by adaptively inflating the state covariance estimates themselves. AREKF, for instance, inflates the *a priori* state covariance estimate. Once again, the stochastic stability of (Reif et al., 1999) is conditional on covariance estimates being bounded (see Lemma L.11).

However, unlike in the case of RORF (Piché et al., 2012) and WoLF (Duran-Martin et al., 2024), where there was a clear, controllable lever to stabilise the filter (the clamping mentioned in §J.1), in this case, nominal filter stability *assumes* state covariance boundedness—this is explicit in the authors' own papers. Under this assumption, these EKF extensions are also stable. Since BRC's objective is to preserve nominal filter stability under the nominal filter's own regularity conditions and assumptions, our argument for stability would remain identical to the ResEKF case.

J.3 ENSEMBLE OF EKFS

We can see the Gaussian Sum Filter (GSF) (Alspach & Sorenson, 1972) as an ensemble of EKFs. The GSF is stable if each of its ensemble (mixture) components—each an EKF—are. Similarly, the BRC-augmented GSF (see Appendix A.1.2 and Algorithms 9 and 10) is stable if each ResEKF component is.