Bias-Variance Analysis of Multi-Step Loss Functions for Dynamical System Identification

Fabien Lionti
Université de Côte d'Azur - INRIA
2004 Rte des Lucioles, 06902 Valbonne
fabien.lionti@inria.fr

Nicolas Gutowski
Université d'Angers - LERIA
2 Bd de Lavoisier, 49000 Angers
nicolas.gutowski@univ-angers.fr

Sébastien Aubin
Direction Générale de l'Armement
Route de Laval - 49245 Avrillé
sebastien.aubin@intradef.gouv.fr

Philippe Martinet
Université de Côte d'Azur - INRIA
2004 Rte des Lucioles, 06902 Valbonne
philippe.martinet@inria.fr

***Abstract*— System identification is a fundamental task in understanding and modeling dynamical systems, with extensive applications in engineering. Traditional statistical estimators for system identification rely on loss functions based on single-step predictions of model state variables. However, these approaches often lack robustness and reliability in real-world scenarios characterized by noisy and imperfect data. Recent advancements have introduced multi-step loss functions for autoregressive neural network predictions, leading to significant improvements in system identification performance. These loss functions are optimized via gradient descent, leveraging backpropagation through the numerically integrated neural network architecture. Despite their potential, the statistical and mathematical properties of these gradient estimators, such as bias, variance and robustness, remain underexplored. This paper examines the statistical and mathematical characteristics of multi-step loss function estimators in the context of dynamical system identification. We provide a theoretical foundation for the bias-variance decomposition of these loss functions, enabling the separation of error contributions from disturbances and deterministic model parameterization. Theoretical insights are validated and extended through empirical analysis, allowing an exploration of the bias-variance decomposition dynamics across the training phase. Our results demonstrate both theoretically and practically the influence of the contractive properties of the underlying dynamical system and the autoregressive prediction horizon on training stability. By bridging the theoretical and practical gap in the exploration of multi-step loss functions, this work contributes to the understanding and development of more robust and reliable methods for dynamical system identification involving gradient descent.**

***Index Terms*— Dynamical System Identification, Multi-step Loss Functions, Bias-Variance Decomposition, Prediction Horizon, Gradient Stability**

I. Introduction

System identification corresponds to the process of deriving mathematical models from observed data, playing a pivotal role in understanding and predicting the behavior of dynamical systems. Its applications span a wide range of fields, including control systems, predictive modeling, state observation, and anomaly detection, all of which rely heavily on the accuracy and reliability of these models [1].

In practical scenarios, system identification is often challenged by the presence of noise, outliers, and disturbances in the data. These imperfections significantly impact the performance of traditional statistical estimators, often leading to reduced accuracy and reliability under such conditions.

The effectiveness of system identification methods depends critically on several factors: the choice of the model class, the information provided by prior knowledge or available data, the statistical estimation process, and the validation methods used to ensure the generalization capability of the identified model [2]. System identification using neural network model classes has recently gained renewed attention, driven by the conceptual linkage between deep residual networks and continuous dynamical systems [3]. These methods utilize automatic differentiation frameworks to optimize the numerical integration of neural network predictions across entire autoregressive trajectories, employing multi-step loss function. They have been explored in both discrete settings, leveraging the connection between residual architectures and discrete integration methods [4], and in continuous frameworks [5], where neural networks are represented as continuous layer depth and trained using adjoint sensitivity methods. These approaches enable more stable training processes and enhances the generalization capability of the identified models with respect to one-step ahead model fitting over various type of dynamical systems [6], [7], [8], [9].

The mathematical characterization of statistical estimators [10] in system identification context such as unbiasedness, efficiency, consistency, and robustness are of interests as it allows a better understanding of the estimation process, opening door for improvement or use appropriated statistical method for specific cases. However, while these properties have been extensively studied for traditional single-step loss functions [1], their behavior in the context of multi-step loss functions remains less explored in system identification context using neural network, particularly in case of gradient descent optimization method.

In this direction, this paper explores the statistical characterization of multi-step loss function estimator through bias-variance decomposition in the context neural network model optimization using gradient descent methods. The

bias-variance decomposition for multi-step loss functions enables the separation of deterministic errors, caused by modeling inaccuracies, from variance errors arising from disturbances in data acquisition allowing to asses deterministic from stochastic gradient contribution, and gaining insight in convergence dynamic. The paper provides both theoretical and empirical justifications for the superior generalization performance of multi-step loss functions over single-step loss functions, particularly in long-horizon simulations of neural network architectures trained on noisy observations. Additionally, results highlight the statistical property dependence of this loss on contractive and non-contractive properties of the underlying dynamical system. These findings open opportunities to enhance training stability by leveraging knowledge of system-specific properties, offering pathways to more robust and reliable system identification methods.

II. Related work

The behavior of multi-step loss functions is of increasing interests in many engineering fields for their capability to identify systems with long horizon simulation capabilities. In robotic context [11] introduces a multi-step loss function for training predictive models in model-based reinforcement learning context. This approach addresses the issue of compounding errors when simulating trajectories over long horizons, particularly in noisy environments. By proposing a weighted multi-step Mean Squared Error (MSE) objective, they demonstrate its advantages in both linear and nonlinear dynamical systems. Closely related, [12] presents a robust predictive control method leveraging multi-step prediction models identified from data using a set-membership approach. By modeling systems with multi-step predictors that account for bounded uncertainties, the proposed algorithm ensures input/output constraint satisfaction, recursive feasibility, and robust convergence. The authors validate their approach with simulations, demonstrating improved precision and effectiveness compared to traditional methods relying on single-step models. In the context of anomaly detection, [13] develops a neural network-based framework for dynamical systems, leveraging robustness bounds on neural network estimators under input perturbations. By propagating auto-regressively ellipsoidal confidence bounds through a neural network, the authors define a robust prediction bound that distinguishes normal behavior from anomalies. The framework is applied to both linear and nonlinear systems, showcasing its effectiveness in fault detection and quantifying the impact of training on noisy versus noiseless data. In the context of Single Source of Error state space approach, [14] showed that the main benefits of multi-step estimator is it acts as a shrinkage effect on the model parameters induced by increasing the number of model steps during the estimation process. Implying reduced variance and increase robustness properties in the context of time series modeling. Closely related, [15] introduces a robust nonlinear system identification framework using feedforward multilayer neural networks and radial basis function networks. Novel identification algorithms ensure the persistency of excitation condition, enabling accurate modeling of system dynamics under noise and disturbances. Evaluated with L_1, L_2, and H_∞ cost criteria, the proposed methods demonstrate robust performance and effectiveness in identifying nonlinear systems, even in challenging conditions.

In a similar direction, this paper proposes a mathematical framework to characterize the behavior of gradient descent estimators derived from multi-step loss functions in the context of nonlinear system identification using neural network model classes. The framework relies on the bias-variance decomposition of the multi-step loss, isolating noise from the modeling bias contribution to the final gradient estimate. Complementary experiments on nonlinear dynamical systems illustrate the theoretical results, providing further insight into the estimator's behavior depending on the contractive properties of the observed dynamical system.

III. Proposed Method

A. Dynamical System Description

We consider the following discrete-time dynamical system f:

$$x_{t+1} = f(x_t, u_t, \theta), \tag{1}$$

where:

- $x_t \in \mathbb{R}^m$ represents the state of the system,
- $u_t \in \mathbb{R}^n$is the control inputs,
- θ represents the set of parameters characterizing the dynamics f.

B. Dataset and Observations

We define a dataset D comprising N trajectory pairs, where each trajectory spans T time steps. The dataset is defined as:

$$D = \left\{ \{(x_0^i, u_0^i), (x_1^i, u_1^i), \ldots, (x_T^i, u_T^i)\}_{i=1}^N \right\}.$$

Each trajectory is generated recursively by the dynamics f starting from the initial state $\bar{x}_0$, parameterized by the true parameters θ^*, as:

$$\bar{x}_{t+1} = f(\bar{x}_t, u_t, \theta^*), \tag{2}$$

A Gaussian noise ϵ is added at each step t, resulting in the observed state:

$$x_t = \bar{x}_t + \epsilon, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \Sigma)$ represents measurement noise with a mean zero and covariance Σ.

C. Prediction Model

For a given parameterization θ, we define the predicted trajectory starting from $x_0^{(i)}$ as:

$$\left(\hat{x}_1^{(i)}, \hat{x}_2^{(i)}, \ldots, \hat{x}_T^{(i)}\right),$$

where the states are recursively generated using the control inputs $\{u_t^{(i)}\}_{t=0}^T$ from the dataset:

$$\hat{x}_{t+1}^{(i)} = f(\hat{x}_t^{(i)}, u_t^{(i)}, \theta). \tag{4}$$

We define the distribution $Q(\hat{x}_1, \ldots, \hat{x}_T \mid \bar{x}_0)$ describes the uncertainty in the trajectory induced by the Gaussian uncertainty in the initial state x_0 and its deterministic propagation through the dynamics f. Formally:

$$Q(\hat{x}_1, \ldots, \hat{x}_T \mid \bar{x}_0) = R(x_0 \mid \bar{x}_0) \times \prod_{t=1}^{T} \delta\big(\hat{x}_t - f(\hat{x}_{t-1}, u_t, \theta)\big). \tag{5}$$

- The Dirac delta function $\delta(\cdot)$ enforces the deterministic relationship $\hat{x}_t = f(\hat{x}_{t-1}, u_t, \theta)$. If this condition is not met, the probability becomes zero.
- The term $R(x_0 \mid \bar{x}_0)$ represents the Gaussian uncertainty related to ϵ perturbations in the initial state: $x_0 \sim \mathcal{N}(\bar{x}_0, \Sigma)$.
- The uncertainty in x_0 propagates deterministically through f, defining the distribution $Q(\hat{x}_1, \ldots, \hat{x}_T \mid \bar{x}_0)$.

D. Autoregressive Loss

To fit the model, we define the multi-step loss $L(\theta)$ as the average Mean Squared Error (MSE) over N trajectory predictions of T time-step:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \|x_t^{(i)} - \hat{x}_t^{(i)}\|^2, \tag{6}$$

where $x_t^{(i)}$ and $\hat{x}_t^{(i)}$ denote the true and predicted states at time t for trajectory i. For a datasets D containing a sufficiently large number of trajectory, where multiple similar trajectories are available under comparable initial conditions perturbed by ϵ, the loss can be reformulated as an expectation over two key distributions:

- $P(x_0, u_t)$: The distribution of initial states and control inputs under the true dynamics parameterized by θ^*. This distribution arises naturally from the observed data. It captures the statistical variability in the initial conditions and control inputs sampled during system observations.
- $Q(\hat{x}_t \mid x_0)$: The distribution of predicted trajectories generated by the model parameterized by θ. its sensitivity to initial conditions x_0. It encapsulates the uncertainty inherent in the model predictions for close trajectory corresponding to perturbed initial state.

The separation into these two distributions allows to distinguish between variability introduced by the observed data (through $P(x_0, u_t)$) and the model's ability to replicate the observed trajectories (through $Q(\hat{x}_t \mid x_0)$). This distinction is crucial in understanding how the model handles noise, variability, and long-term prediction accuracy.

Formally:

$$L(\theta) = \mathbb{E}_{x_0 \sim P}\left[\mathbb{E}_{\hat{x}_t \sim Q}\left[\frac{1}{T}\sum_{t=1}^{T} \|x_t - \hat{x}_t\|^2\right]\right]. \tag{7}$$

E. Uncertainty Propagation and Mean Trajectory

The mean trajectory of the system, μ_t, is the expected value of the system's dynamics under noise ϵ:

$$\mu_{t+1} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Sigma)}\left[f(x_t + \epsilon, u_t, \theta)\right]. \tag{8}$$

For small perturbations ϵ, the dynamics $f(x_t, u_t, \theta)$ can be approximated near the mean trajectory using a first-order Taylor expansion:

$$f(x_t, u_t, \theta) \approx f(\mu_t, u_t, \theta) + J_f(\mu_t)(x_t - \mu_t), \tag{9}$$

where $J_f(\mu_t)$ is the Jacobian matrix of f with respect to the state x_t, evaluated at μ_t:

$$J_f(\mu_t) = \left.\frac{\partial f}{\partial x_t}\right|_{x_t = \mu_t}. \tag{10}$$

The deviations from the mean trajectory are defined as $\delta x_t = x_t - \mu_t$. These deviations propagate approximately as for small ϵ:

$$\delta x_{t+1} \approx J_f(\mu_t)\delta x_t \tag{11}$$

Under these assumptions, the covariance propagation reduces to:

$$\Sigma_{t+1} = \mathbb{E}_{x_t \sim Q}\left[(\delta x_{t+1})(\delta x_{t+1})^T\right], \tag{12}$$

$$\Sigma_{t+1} = J_f(\mu_t)\Sigma_t J_f(\mu_t)^T. \tag{13}$$

F. Link Between Q and Σ_t

- **Distribution of Predicted States** $Q(\hat{x}_t \mid x_0)$ is the distribution of predicted states $\hat{x}_t$ under the model parameterized by θ, starting from the initial state x_0. Q captures the full uncertainty in the predicted trajectory, including the initial uncertainty in x_0 and its propagation through the dynamics.
- **Covariance of Deviations** Σ_t is the covariance matrix of deviations $\delta x_t = x_t - \mu_t$, where x_t are the states in Q and μ_t is the mean trajectory. Σ_t quantifies the spread (uncertainty) of the states x_t around the mean trajectory μ_t.
- **Connection:** Q defines the full probabilistic distribution of the predicted trajectory with $\Sigma_0 = \Sigma$ the covariance matrix of the initial perturbation ϵ applied on $\bar{x}_0$, while Σ_t provides a summary of Q in terms of its second-order moments (covariance). Both Q and Σ_t evolve according to the same dynamics f. The mean trajectory μ_t governs the central tendency of Q, while the Jacobian $J_f(\mu_t)$ determines how uncertainty (and thus Σ_t) propagates through the system.

G. Impact of System Stability on Σ_t

The stability of the system f depends on the eigenvalues of the Jacobian matrix $J_f(\mu_t)$:

- **Contracting Systems:** If all eigenvalues of $J_f(\mu_t)$ have magnitudes less than 1, the covariance Σ_t diminishes over time, leading to convergence to a deterministic trajectory.

- **Non-Contracting Systems:** If any eigenvalue has magnitude greater than 1, Σ_t may grow, resulting in divergent trajectories.

The covariance propagation over T steps (Figure 1) is given by :

$$\Sigma_T = \prod_{k=0}^{T-1} J_f(\mu_k)\Sigma_0 \prod_{k=0}^{T-1} J_f(\mu_k)^T. \tag{14}$$

For contracting systems:

$$\Sigma_T \to 0 \quad \text{as } T \to \infty. \tag{15}$$

H. Trace of the Covariance Matrix

The trace of the covariance matrix, denoted as $\text{Trace}(\Sigma_T)$ measures the total variance of the system trajectories, distributed across all principal directions of the uncertainty ellipsoid. The trace of Σ_t at a specific time step t can be expressed using the dot product as:

$$\text{Trace}(\Sigma_t) = \mathbb{E}_{\hat{x}_t \sim Q}\left[(\hat{x}_t - \mu_t)^\top (\hat{x}_t - \mu_t)\right], \tag{16}$$

which is equivalent to the expectation of the squared Euclidean norm of the deviations between the trajectory and its mean:

$$\text{Trace}(\Sigma_t) = \mathbb{E}_{\hat{x}_t \sim Q}\left[\|\hat{x}_t - \mu_t\|^2\right]. \tag{17}$$

This formulation shows that the trace of the covariance matrix Σ_t represents the total variance of the system at time t, distributed across all state dimensions. It establishes a quantitative connection between the amplitude of Σ_t and the empirical variance observed in the system's trajectories starting from initial state $\bar{x}_0$ perturbed by ϵ.

I. Bias-variance decomposition

To assess the impact of random perturbations ϵ in the dataset D on the estimation of the parameterization θ, we first reformulate the loss function using a bias-variance decomposition. This reformulation separates the contributions of random perturbations (*variance*) from those arising due to parameterization errors (*bias*).

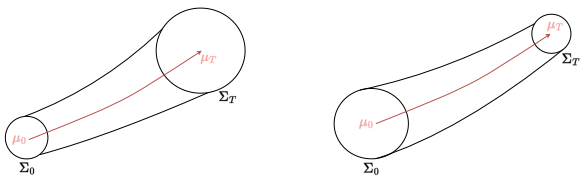

Fig. 1: The left graph shows the expansion of the uncertainty Σ over time, while the right graph illustrates the contraction of the uncertainty, which is related to the eigenvalue properties of the Jacobian matrix of the dynamical system.

Theorem 1 (Bias-Variance Decomposition)**.** *The multi-step loss decomposes as:*

$$L(\theta) = \mathit{Bias}_T^2 + \mathit{Var}_T, \tag{18}$$

where:

$$\mathit{Bias}_T^2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\left[\|x_t - \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t]\|^2\right], \tag{19}$$

$$\mathit{Var}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\left[\mathit{Trace}(\Sigma_t)\right]. \tag{20}$$

Proof: Starting from 7 the multi-step loss formulation:

$$L(\theta) = \mathbb{E}_{x_t \sim P}\left[\mathbb{E}_{\hat{x}_t \sim Q}\left[\frac{1}{T}\sum_{t=1}^{T}\|x_t - \hat{x}_t\|^2\right]\right]. \tag{21}$$

Expanding the expectation by leveraging it's linearity property:

$$L(\theta) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\left[\mathbb{E}_{\hat{x}_t \sim Q}\left[\|x_t - \hat{x}_t\|^2\right]\right]. \tag{22}$$

Decomposing the squared norm:

$$L(\theta) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\Bigg[\mathbb{E}_{\hat{x}_t \sim Q}\Big[\|x_t - \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t] + \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t] - \hat{x}_t\|^2\Big]\Bigg]. \tag{23}$$

Expanding the squared term, we get:

$$\begin{aligned} L(\theta) = {} & \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\left[\mathbb{E}_{\hat{x}_t \sim Q}\left[\|x_t - \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t]\|^2\right]\right] \\ & + \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\left[\mathbb{E}_{\hat{x}_t \sim Q}\left[\|\hat{x}_t - \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t]\|^2\right]\right] \\ & + \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\left[\mathbb{E}_{\hat{x}_t \sim Q}\left[(x_t - \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t])^\top (\hat{x}_t - \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t])\right]\right]. \end{aligned} \tag{24}$$

The terms in the bias-variance decomposition are:

- **Bias squared (Bias_T^2)**:

$$\text{Bias}_T^2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim P}\left[\mathbb{E}_{\hat{x}_t \sim Q}\left[\|x_t - \mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t]\|^2\right]\right]. \tag{25}$$

This term represents the systematic error due to the deviation of the mean predictions $\mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t]$ from the true values x_t. It quantifies the parametrization error introduced by the model's inability to perfectly capture the true dynamics. As the bias term only depends on the mean prediction $\mathbb{E}_{\hat{x}_t \sim Q}[\hat{x}_t]$, not on the variability of $\hat{x}_t$ around its mean. Since this variability is irrelevant to the bias, the inner expectation over Q becomes unnecessary for this term.

Thus, the bias simplifies to 19:

$$\text{Bias}_T^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x_t\sim P}\left[\|x_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t]\|^2\right].$$

- **Variance (Var_T)**:

$$\text{Var}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x_t\sim P}\left[\mathbb{E}_{\hat{x}_t\sim Q}\left[\|\hat{x}_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t]\|^2\right]\right]. \tag{26}$$

This term measures the average variability of the model's predictions $\hat{x}_t$ around their mean $\mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t]$, aggregated across all trajectories $x_t \sim P$. The vector $(\hat{x}_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t])$ is referred to as the **variance vector**. This term is related to the trace of the covariance matrix Σ_T 17, simplifying to 20:

$$\text{Var}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x_t\sim P}\left[\text{Trace}(\Sigma_t)\right],$$

which represents the expected norm of the uncertainty ellipsoid caused by perturbations in the initial conditions or stochastic dynamics over all initial conditions sampled from P.

- **Cross-Term**:

$$\text{Cross-Term} = \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}_{x_t\sim P}\Bigg[\mathbb{E}_{\hat{x}_t\sim Q}\big[(x_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t])^\top (\mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t] - \hat{x}_t)\big]\Bigg]. \tag{27}$$

This term represents the interaction between the bias vector $(x_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t])$ and the variance vector $(\hat{x}_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t])$. Under the assumption that these two vectors are uncorrelated or independent, the cross-term vanishes:

$$\mathbb{E}_{x_t\sim P}\Bigg[\mathbb{E}_{\hat{x}_t\sim Q}\big[(x_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t])^\top (\mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t] - \hat{x}_t)\big]\Bigg] = 0. \tag{28}$$

It implies a situation where the model does not overfit the noise ϵ during training. The model is effectively separating the true signal (captured in the bias) from the noise (captured in the variance). This ensures that the predictions remain generalizable and not overly influenced by specific random perturbations in the data.

Under this hypothesis, the loss simplifies to 18:

$$L(\theta) = \text{Bias}_T^2 + \text{Var}_T.$$

Corollary 1 (Effect of Recursive Steps on Variance)**.** *For contracting systems, where the norm of the Jacobian satisfies $\|J_f(\mu_t)\| < 1$, the variance Σ_t at each time step t decreases monotonically.*

Consequences:

1) *The total variance satisfies:*

$$\textit{Trace}(\Sigma_{T+1}) < \textit{Trace}(\Sigma_T). \tag{29}$$

and equivalently:

$$\textit{Var}_{T+1} < \textit{Var}_T, \tag{30}$$

2) *As $T \to \infty$, the time-averaged variance converges to zero:*

$$\textit{Trace}(\Sigma_T) \to 0.$$

Subsequently, for sufficiently long T and dynamical system with contractive properties, the loss function $L(\theta)$ 18 tend to only represents the bias component related to parametrization error $\theta \neq \theta^*$.

J. Impact on Gradient Descent Optimization

In order to optimize the parametrization of f, we consider the gradient descent method:

$$\theta_{k+1} = \theta_k - \eta\nabla L(\theta_k), \tag{31}$$

where k corresponds to the gradient descent step, and η is the learning rate.

From Theorem 1, the gradient $\nabla L(\theta)$ is given by:

$$\nabla L(\theta) = \nabla\text{Bias}_T^2 + \nabla\text{Var}_T, \tag{32}$$

$$= \nabla\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x_t\sim P}\left[\|x_t - \mathbb{E}_{\hat{x}_t\sim Q}[\hat{x}_t]\|^2\right] + \nabla\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{x_t\sim P}\left[\text{Trace}(\Sigma_t)\right]. \tag{33}$$

Corollary 2 (Expected Improvement in the Loss Function)**.** *The expected improvement in the loss function $L(\theta)$ at each gradient descent step is given by:*

$$L(\theta_k) - L(\theta_{k+1}) \approx \eta\|\nabla L(\theta_k)\|^2, \tag{34}$$

where the squared norm of the gradient can be expanded as:

$$\|\nabla L(\theta_k)\|^2 = \|\nabla \textit{Bias}_T^2\|^2 + 2\nabla \textit{Bias}_T^2 \cdot \nabla \textit{Var}_T + \|\nabla \textit{Var}_T\|^2. \tag{35}$$

Substituting this expansion into the loss improvement equation yields:

$$L(\theta_k) - L(\theta_{k+1}) \approx \eta\Big(\|\nabla \textit{Bias}_T^2\|^2 + 2\nabla \textit{Bias}_T^2 \cdot \nabla \textit{Var}_T + \|\nabla \textit{Var}_T\|^2\Big). \tag{36}$$

Implications for Contractive Dynamical Systems

- For large T, variance terms ($2\nabla\text{Bias}_T^2 \cdot \nabla\text{Var}_T$, $\|\nabla\text{Var}_T\|^2$) and diminish, and loss improvement is dominated by $\|\nabla\text{Bias}_T^2\|^2$.
- Small T amplifies gradient variance components, reducing expected loss improvement.

Implications for Non-Contractive Dynamical Systems

- For large T, variance terms may grow due to trajectory divergence, overshadowing bias reduction.
- Small T mitigates variance divergence, stabilizing gradients but potentially limiting accuracy.

Theorem 1 and its corollaries highlight the interplay between bias and variance in gradient descent optimization. These results underscore the importance of leveraging the system's dynamic properties and carefully managing variance through T to achieve effective and robust parameter optimization.

IV. Experimental Setting

A. Empirical Study Design

In this section, we outline the training setup used to empirically study the evolution of the bias-variance decomposition (Theorem 1) across multiple training epochs. To account for variability due to initialization, the experiments are conducted using five different initial parameterizations θ. All experiments are performed with a sampling period of $\Delta_t = 0.05$ seconds, and the number of assessed autoregressive steps T is set to $T \in \{1, 4, 8, 16\}$. The dynamical system $\hat{x}_{t+1} = f(\hat{x}_t, u_t, \theta)$ is represented by a fully connected multi-layer perceptron (MLP) with a single hidden layer of eight neurons. Training is carried out using the Adam optimizer [16] with a learning rate of $\eta = 0.001$. To investigate the dynamics of contractive and partially contractive systems, we employed two distinct dynamical systems to generate the dataset D: the forced Duffing oscillator and the Lorenz system. These systems were specifically chosen to demonstrate contrasting contractivity properties, providing complementary perspectives on the behavior of dynamical systems under different regimes. For both systems, 100 trajectories of 20 seconds each were generated using the forward Euler method with a fixed sampling period of Δ_t. Initial conditions were uniformly sampled from $[-5, 5]$, ensuring a diverse representation of the state space.

The **forced Duffing oscillator** is a nonlinear dynamical system with parameters specifically tuned to exhibit *contractive properties*. Its state-space representation is:

$$\ddot{x} + \delta\,\dot{x} + \alpha\,x + \beta\,x^3 = F\,\cos(\omega t),$$

where x is the displacement, $\dot{x}$ the velocity, and $\ddot{x}$ the acceleration. The system parameters include: $\delta = 0.5$: moderate damping coefficient, ensuring energy dissipation while allowing oscillations. $\alpha = 1.0$: linear stiffness, providing a restoring force. $\beta = 0.1$: small nonlinear stiffness, contributing stabilizing effects for large displacements. $F = 1.0$: forcing amplitude, introducing significant external periodic energy. $\omega = 1.0$: forcing frequency, governing the oscillation rate of the periodic input. This parameter set ensures that the system exhibits strong damping and bounded oscillatory behavior, though the increased forcing amplitude ($F = 1.0$) introduces greater variability in trajectories. While the damping and restoring forces dominate, the periodic energy injection prevents strict global contractivity, though local contractivity is retained near periodic attractors.

To contrast the forced Duffing oscillator, we employed the **Lorenz system**, a well-known dynamical system that exhibits *partially contractive properties*. Its dynamics are governed by:

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z) - y, \quad \dot{z} = xy - \beta z,$$

where σ, ρ, and β are the system parameters.

The Lorenz system's behavior is highly dependent on the region of the state space: In some regions, particularly near stable fixed points, trajectories exhibit contractive behavior, as nearby points converge due to the system's attractors. However, non-contractivity dominates in regions near bifurcation points or within the chaotic Lorenz attractor. Specifically, in the "wings" of the attractor, trajectories repeatedly diverge and converge as they spiral, creating a complex interplay of local contraction and divergence. For this study, we selected: $\sigma = 10, \quad \rho = 14, \quad \beta = \frac{8}{3}$.. With these parameters, the Lorenz system transitions out of the classical chaotic regime, exhibiting complex, yet partially convergent dynamics. The reduced ρ value moderates the chaotic behavior, increasing the likelihood of transient contractive regions before divergence dominates. This configuration highlights the Lorenz system's sensitivity to initial conditions and its partially contractive nature, in contrast to the Duffing oscillator's more consistent behavior. In both case, observations are generated with an ϵ noise corresponding to a diagonal covariance matrix Σ with standard deviation of 0.05 over each state variable.

B. Empirical Bias-Variance Computation

The exact computation of the bias-variance decomposition requires the dataset D to include repeated versions of the same trajectory with different perturbations. These repetitions are necessary to estimate Var_T and $Bias_T^2$. To validate the conclusions of the proposed theorem and its corollaries, we use Algorithm 1. This algorithm computes $Bias_T^2$ and Var_T over N deterministic trajectories $\bar{x}_{0:T}$, generated from $P(x_0, u_T)$ and perturbed K times by adding noise $\epsilon \sim N(0, \Sigma)$. Bias and variance are tracked for each θ_k parametrization, with gradient descent updating the parameters based on only one perturbed version of each of the N trajectories. [1]

V. Results

The Figure 2 illustrates the temporal evolution of the Root Mean Squared Error (RMSE) across trajectories generated over a ten-second period for both the Duffing and Lorenz systems. For the Duffing system, the results demonstrate improved generalization capability as the T parameter increases. This improvement is accompanied by a reduction in prediction variability, visualized through the uncertainty

[1] Due to anonymization restrictions, we are unable to provide our GitHub link at this stage. It will be included in the camera-ready version upon acceptance.

Algorithm 1 Empirical Bias and Variance Computation

for $(\bar{x}_0, \ldots, \bar{x}_T), (u_0, \ldots, u_T) \in D$ **do**
  1. Generate Perturbed Trajectories:
  for $j = 1 \to K$ **do**
    Sample noise: $\epsilon \sim \mathcal{N}(0, \Sigma)$
    $x_{0j} \leftarrow \bar{x}_0 + \epsilon$, $\hat{x}_{0j} \leftarrow x_{0j}$
    for $t = 1 \to T$ **do**
      Predict: $\hat{x}_{tj} \leftarrow f(\hat{x}_{t-1j}, u_{t-1}, \theta)$
  2. Compute Bias and Variance over all K
  for $t = 1 \to T$ **do**
    $E[\hat{x}_t]_i \leftarrow \frac{1}{K} \sum_{j=1}^{K} \hat{x}_{tj}$
    $\text{Var}[x_t]_i \leftarrow \frac{1}{K} \sum_{j=1}^{K} (E[\hat{x}_t]_i - \hat{x}_{tj})^2$
3. Compute Aggregate Metrics over D**:**
$Var_T = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \|\text{Var}[x_t]_i\|^2$
$Bias_T^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \|x_t^i - E[\hat{x}_t]_i\|^2$

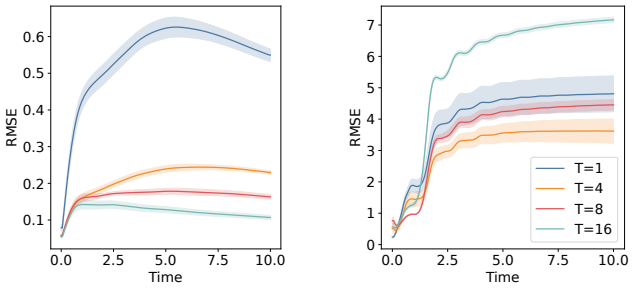

Fig. 2: Temporal evolution of RMSE $\pm 0.05 std$ for autoregressive predictions over a 10-second horizon. The left subfigure illustrates the forced Duffing oscillator, the right subfigure depicts the Lorenz system.

bands corresponding to $\pm 0.5 \sigma_{RMSE}$. These trends highlight the benefits of leveraging larger T values in systems exhibiting contractive properties.

For the Lorenz dynamical system, increasing T does not consistently lead to a reduction in RMSE across all prediction horizons. For short-term predictions, RMSE is lower when training with smaller T values, while for longer horizons, a tradeoff emerges where an intermediate T provides optimal performance. The Lorenz system exhibits locally strong non-contractive behavior, making it highly sensitive to initial conditions. This sensitivity leads to an irreducible error introduced by increasing the number of autoregressive steps, which persists during training. Consequently, extending T highlights a fundamental limitation: the tradeoff reflects a balance between the destabilizing effect of small T, where noise fluctuations can cause the model to diverge over time, and the irreducible error from initial condition sensitivity, which results in diverging trajectories during training.

The Figure 3 illustrates the evolution of the variance Var_T during each epoch of the training process. For larger T, the initial parametrization θ_0 of the neural network fails to align with the underlying contractive property of the Duffing system, as the model has not yet learned to

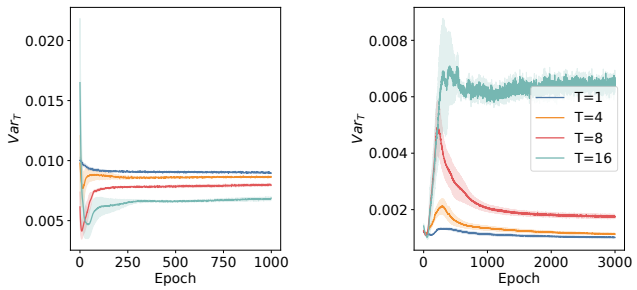

Fig. 3: Evolution of $Var_T \pm 0.5 std$ across training epochs for different values of the autoregressive step T. The left figure illustrates the Duffing system, while the right figure represents the Lorenz system.

reproduce the system's underlying behavior, resulting in higher initial Var_T values. However, as training progresses and the parametrization of f quickly converges toward an approximation of the observed system dynamics, Var_T gradually asymptotically converges toward decreasing values ordered by increasing T. We argue that this asymptotic phase corresponds to a situation where the θ parametrization is close to a minimum of the loss $L(\theta)$, but the presence of gradient noise associated with Var_T prevents precise convergence to this minimum, resulting in a "noise floor." In accordance with Corollary 1, increasing T reduces sensitivity to this noise, allowing the model to approach the minimum more closely by mitigating the "noise floor" effect. This improved parameter identification facilitates a reduction in error during long-horizon simulations, as illustrated in Figure 2. In the case of the Lorenz system, due to the presence of strongly non-contractive regions in the phase space, we observe the opposite scenario. An increase in the asymptotic value of Var_T during training is observed with increasing T, which can be attributed to the model learning the dynamics of the underlying system that exhibits non-contractive properties. Despite this increase in Var_T, the results observed in Figure 2 still benefit from longer T.

To further analyze the performance gains observed with increasing T in the Duffing and Lorenz case, the Figure 4 study the ratio $\frac{\text{Bias}_T^2}{L(\theta)}$, which represents the proportion of the loss $L(\theta)$ attributed to parametrization error independently of initial noise perturbations. This ratio provides insights into the signal-to-noise ratio's contribution to the final computed gradient. A higher ratio indicates that the model is more influenced by the systematic error arising from its parametrization, while a lower ratio suggests that gradient updates are increasingly dominated by noise fluctuations.

In the Duffing case, as T increases, we observe a systematic increase in $\frac{\text{Bias}_T^2}{L(\theta)}$, reflecting improved parametrization alignment with the underlying dynamics. This increase indicates that the model's gradient updates are less perturbed by noise throughout the training phase, enabling more stable convergence toward a loss minimum that corresponds to a better long-term representation of the underlying dynamical

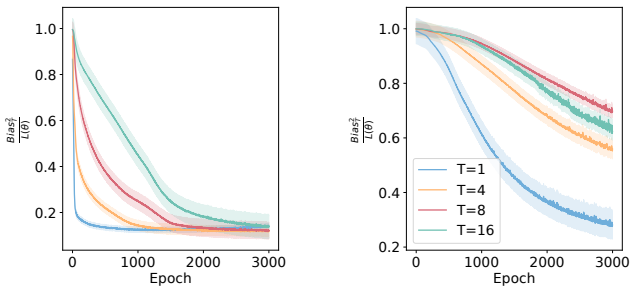

Fig. 4: Evolution of the $Bias_T^2$ contribution to the loss function $L(\theta)$ over training epochs for the Duffing (left) and the Lorenz system (right) with uncertainty band $\pm 0.5 std$.

system. By contrast, in the Lorenz case, the ratio exhibits a less pronounced increase with increasing T, revealing the emergence of a trade-off and highlighting the challenge of noise amplification due to the system's non-contractive regions. These findings underscore the differential impact of T on gradient stability and parametrization accuracy between the two systems, explaining the RMSE evolution over long simulation horizons.

VI. Limitations

While the proposed framework provides insights into the link between performance improvement with longer training horizons and the underlying contractive properties of dynamical systems, several limitations of this work must be acknowledged. 1) Assumptions on Noise and Dynamics: The analysis assumes Gaussian noise with a known covariance matrix and linearizable dynamics near the mean trajectory. Although these assumptions simplify the theoretical framework, they may not fully capture the behavior of highly nonlinear systems or account for scenarios requiring higher-order statistics for uncertainty propagation. 2) Experimental Scope: To illustrate the theoretical results, experiments were conducted on two contrasting systems: the forced Duffing oscillator, exhibiting strong contractive properties, and the Lorenz system, which locally exhibits strong non-contractive properties. However, real-world scenarios may involve more nuanced systems, falling between these extremes. In such cases, increasing the horizon can still provide value, even for moderately non-contractive systems, as it results from a trade-off between bias and variance, despite the variance diverging over time.

VII. Conclusion and Future Work

In this study, we investigated the statistical and mathematical properties of multi-step loss function estimators in the context of dynamical system identification. By leveraging a bias-variance decomposition framework, we demonstrated how multi-step training improves the predictive performance of neural network models, particularly through the reduction of variance during the convergence phase in the case of contractive dynamical systems. These findings provide a deeper theoretical foundation for understanding the behavior of multi-step estimators.

The empirical results align with theoretical insights, demonstrating that increasing the number of autoregressive steps during training significantly improves long-term prediction accuracy. This increase enhances gradient stability by reducing perturbation effects, ultimately leading to improved convergence properties. Future work could explore alternative formulations of the multi-step loss function using the provided statistical indicators to enhance training stability. This could involve accounting for local system properties, such as systems with locally non-contractive patterns, to adapt the autoregressive horizon locally. Further extensions could consider higher-order effects of uncertainty propagation. In summary, this work bridges theoretical advancements and practical applications in system identification, contributing to the development of more robust and reliable models.

References

[1] L. Ljung, *System identification (2nd ed.): theory for the user*. USA: Prentice Hall PTR, 1999.

[2] ——, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.

[3] W. Weinan, "A proposal on machine learning via dynamical systems"," *Communications in Mathematics and Statistics*, vol. 5, no. 1, pp. 1–11, 2017.

[4] C. Qiu, A. Bendickson, J. Kalyanapu, and J. Yan, "Accuracy and architecture studies of residual neural network method for ordinary differential equations," *J. Sci. Comput.*, vol. 95, no. 2, 2023.

[5] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 6571–6583.

[6] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Multistep neural networks for data-driven discovery of nonlinear dynamical systems," *arXiv preprint arXiv:1801.01236*, 2018.

[7] D. Masti and A. Bemporad, "Learning nonlinear state–space models using autoencoders," *Automatica*, vol. 129, p. 109666, 2021.

[8] S. S. Rangapuram, M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski, "Deep state space models for time series forecasting," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associatdam: A method for stochasties, Inc., 2018.

[9] D. Gedon, N. Wahlström, T. B. Schön, and L. Ljung, "Deep state space models for nonlinear system identification," *IFAC-PapersOnLine*, vol. 54, no. 7, pp. 481–486, 2021, 19th IFAC Symposium on System Identification SYSID 2021.

[10] E. Lehmann and G. Casella, *Theory of Point Estimation*, ser. Springer Texts in Statistics. Springer New York, 2003.

[11] A. Benechehab, A. Thomas, G. Paolo, M. Filippone, and B. Kegl, "A multi-step loss function for robust learning of the dynamics in model-based reinforcement learning," *ArXiV*, 2024.

[12] E. Terzi, M. Farina, L. Fagiano, and R. Scattolini, "Robust predictive control with data-based multi-step prediction models," in *2018 European Control Conference (ECC)*, 2018, pp. 1710–1715.

[13] N. Hashemi, M. Fazlyab, and J. Ruths, "Performance bounds for neural network estimators: Applications in fault detection," *2021 American Control Conference (ACC)*, pp. 3260–3266, 2021.

[14] I. Svetunkov, N. Kourentzes, and R. Killick, "Multi-step estimators and shrinkage effect in time series models," *Computational Statistics*, vol. 39, no. 3, pp. 1203–1239, 2024.

[15] S. Lu and T. Basar, "Robust nonlinear system identification using neural network models," in *Proceedings of 1995 34th IEEE Conference on Decision and Control*, vol. 2, 1995, pp. 1840–1845 vol.2.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015.