PAC-Bayesian Error Bound, via Rényi Divergence, for a Class of Linear Time-Invariant State-Space Models

Deividas Eringis [1] John Leth [1] Zheng-Hua Tan [1] Rafal Wisniewski [1] Mihály Petreczky [2]

Abstract

In this paper we derive a PAC-Bayesian error bound for a class of stochastic dynamical systems with inputs, namely, for linear time-invariant stochastic state-space models (stochastic LTI systems for short). This class of systems is widely used in control engineering and econometrics, in particular, they represent a special case of recurrent neural networks. In this paper we 1) formalize the learning problem for stochastic LTI systems with inputs, 2) derive a PAC-Bayesian error bound for such systems, and 3) discuss various consequences of this error bound.

1. Introduction

The Probably Approximately Correct (PAC)-Bayesian learning theory is an important tool for analysing theoretical properties of machine learning algorithm, see (Guedj, 2019; Alquier, 2021; Zhang, 2006; Grünwald, 2012; Alquier et al., 2016; Germain et al., 2016; Sheth and Khardon, 2017). In this paper we will present PAC-Bayesian error bounds for time-series supervised learning, with quadratic loss. More precisely, we consider time-series realised by stochastic linear time-invariant (LTI) systems in state-space form with (stochastic) inputs and unbounded (sub-Gaussian) noise. The proposed bound is based on Rényi divergence and converge to zero as the number of data points N goes to infinity with the rate $O(\frac{1}{\sqrt{N}})$, for a fixed posterior.

Motivation for studying PAC-Bayesian bounds for dynamical systems in state-space form: While there is a wealth of literature on PAC-Bayesian bounds for static models (Alquier, 2021; Guedj, 2019), much less is known for dynamical systems. Moreover, existing results on dynamical systems often concentrate on autoregressive models. However, for various learning problems involving sequential, online or time-series data dynamical systems in state-space form represent a more general and more expressive hypothesis class, see Section 2 for a more detailed discussion. In a nutshell, the hidden states in state-space models encode a growing window of past observations, allowing for potentially more accurate predictions. These favorable properties of state-space models are well-known in econometrics (Durbin and Koopman, 2012) and control theory (Ljung, 1999). Recently, they were also explored in machine learning, in the form of structured state-space models (SSMs), due to their relatively low inference complexity and often high performance (Gu et al., 2021; 2023; Gu and Dao, 2023; Li et al., 2021; Wang and Xue, 2023). Note that recurrent neural networks (RNNs) and their various versions (LSTM, etc.) are all dynamical systems in state-space form. PAC-Bayesian bounds for state-space representations have the potential for providing theoretical guarantees for existing learning algorithms and potentially deriving new ones, see Remark 2.5.

Motivation for LTI systems: The motivation for working with PAC-Bayesian bounds for LTI systems is as follows. First, LTI systems are among the simplest class of dynamical systems in state-space form with partially observed state. In turn, state-space representations with partially observed states 1) are widely used for modeling physical systems, 2) they contain various standard models such as recurrent neural networks (RNN), or structured state-space models (SSMs), 3) are more general than autoregressive models, in particular, in contrast to autoregressive models, state-space models allow integrating a growing number of past observations for generating predictions. In particular,, PAC-Bayesian bounds for LTI models could help to derive such bounds for more general classes of systems. For example, LTI systems are special cases of RNNs and they are a building blocks of SSMs.

Moreover, despite their simplicity, LTI systems are widely used in econometrics and control theory, due to their ability to model various physical processes, see (Ljung, 1999; Pillonetto et al., 2022). In particular, learning of LTI systems is an active research topic, both on its own and as a intermediate step in LQG reinforcement learning (Lale et al., 2020;

[1] Department of Electronic Systems, Aalborg University, Denmark [2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France. Correspondence to: Mihaly Petreczky <mihaly.petreczky@centralelille.fr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Simchowitz, 2021; Simchowitz and Foster, 2020; Hazan et al., 2018). In particular, finite-sample bounds for LTI systems have received significant attention (Lale et al., 2020; Oymak and Ozay, 2022; Simchowitz, 2021; Simchowitz et al., 2019; Sarkar et al., 2021; Tsiamis and Pappas, 2019; Hazan et al., 2018; Lee and Lamperski, 2019). However, the cited papers provide error-bounds for specific learning algorithms. Moreover, they concentrate on parameter estimation error, and not on the generalisation gap (i.e. the difference between the generalisation and empirical loss). PAC-Bayesian bounds also represent a finite-sample bound for learning LTI systems. However, in contrast to the results cited above, the PAC-Bayesian error bounds of this paper, characterize the generalization gap for a wide variety of learning algorithms. The PAC-Bayesian bound can be used for bounding the parameter estimation error, see Remark 4.4.

Related work: Error bounds for certain classes of nonlinear dynamical systems were also derived in (Sattar et al., 2022; Sattar and Oymak, 2022; Blanke and Lelarge, 2023; Foster and Simchowitz, 2020; Mania et al., 2022; Sayedana et al., 2022; Shi et al., 2022; Roy et al., 2021; Ziemann et al., 2022; Ziemann and Tu, 2022; Li et al., 2023), but they assume full state observation and they provide an error bound for a specific learning algorithm. In contrast, we consider models with unobserved (hidden) states.

The literature on LTI systems (Ljung, 1999) traditionally focused on statistical consistency, with a few papers (Campi and Weyer, 2002; Vidyasagar and Karandikar, 2006) on PAC bounds. However, the latter papers are applicable only for bounded signals. PAC bounds for RNNs (which contain LTI systems as a special class) were developed in (Koiran and Sontag, 1998; Sontag, 1998; Chen et al., 2020) using VC dimension, and in (Wei and Ma, 2019; Akpinar et al., 2020; Joukovsky et al., 2021; Chen et al., 2020) using Rademacher complexity, and in (Zhang et al., 2018) using a PAC-Bayesian bounds. However, all the cited papers assume noiseless models, a fixed number of time-steps, that the training data are i.i.d sampled time-series, and the signals are bounded. Moreover, several papers (Koiran and Sontag, 1998; Sontag, 1998; Hanson et al., 2021) assume Lipschitz loss functions, while we use quadratic loss function.

PAC-Bayesian error bounds for dynamical systems were considered in (Alquier and Wintenberger, 2012; Alquier et al., 2013; Haussmann et al., 2021; Banerjee et al., 2021; Shalaeva et al., 2020), but for different learning problems: (Alquier and Wintenberger, 2012; Alquier et al., 2013; Shalaeva et al., 2020) considers autoregressive models, (Haussmann et al., 2021) considers stochastic differential equations, (Banerjee et al., 2021) Markov-chains with observed states. PAC-Bayesian bounds were proposed for (super-)martingales (Haddouche and Guedj, 2022b; Seldin et al., 2012), however they are not applicable to LTI, as the empirical loss for LTI systems is not a super martingale. PAC-Bayesian bounds were also derived for online learning (Haddouche and Guedj, 2022a), but again, those bounds are not applicable to LTI systems. In (Eringis et al., 2023a; 2021) PAC-Bayes bounds based on KL-divergence were developed for LTI systems with unbounded signals However, those bound do not converge to zero. In contrast, in this paper we use Rényi divergence, and the error bound converges to zero with $O(\frac{1}{\sqrt{N}})$, at least when the Rényi divergence between the prior and posterior is bounded, e.g., the posterior is fixed. In (Eringis et al., 2023c;a) KL-based PAC-Bayesian bounds were developed for state-space systems, in particular for LTI systems, but all the signals were assumed to be bounded. In contrast, we consider unbounded signals.

Motivation for using Rényi instead of KL divergence: In this paper we follow the approach (Bégin et al., 2016; Alquier and Guedj, 2018) and we use Rényi divergence between the prior and posterior to formulate PAC-Bayesian bounds. In contrast, most of the existing literature considers PAC-Bayesian bounds which use Kuehlback-Leibner (KL) divergence to express this difference. The use of KL divergence leads to tighter bounds. However, the use of Rényi divergence allowed us to avoid certain technical difficulties which arise for LTI systems with unbounded signals and quadratic loss function. At the same time, the latter assumptions are standard in applications of LTI systems in econometrics and control (Durbin and Koopman, 2012; Ljung, 1999), and hence cannot be ignored. The difficulty of using KL bounds for dynamical systems with unbounded signals and quadratic loss were also apparent in (Shalaeva et al., 2020; Eringis et al., 2021; 2023a), where the derived KL-based bounds do not converge to zero as the number of data points $N \to \infty$. For a more detailed discussion on these technical difficulties see Section B, Appendix.

Further challenges related to state-space systems: As it was mentioned above, the hidden states of LTI systems potentially encode information on unbounded number of past inputs. This requires adjusting the standard definition of generalisation loss and prevents the application of PAC-Bayesian bounds for autoregressive models to LTI systems. We refer to Section 2, for a more detailed discussion.

Main contribution and novelty: The main contribution of the paper is that **(1)** it proposes a PAC-Bayesian bound, i.e., a bound on the generalisation gap **(2)** the bound is $O(\frac{1}{\sqrt{N}})$ in terms of the length N of the time-series, **(3)** the bound is valid for a potentially large class of learning algorithms, **(4)** the bound applies to stochastic LTI systems in state-space form, with unbounded signals and hidden states, and for learning from a single time series. To the best of our knowledge, *there are no other comparable PAC bounds in the literature for this learning problem.*

Outline of the paper: In Section 2 we present the informal problem formulation and the main result. In Section 3 we state the formal assumptions. In Section 4 we state the main results of the paper. A numerical example is presented in Section 5. Some of the proofs and further numerical examples are presented in Appendix.

2. Informal statement of the result

Notation and terminology: To enhance readability we occasionally use $\triangleq$ to denote ”defined by”. Let $\mathbf{F}$ denote a σ-algebra on the set Ω and let $\mathbf{P}$ be a probability measure on $\mathbf{F}$. Unless otherwise stated all probabilistic considerations will be with respect to the probability space $(\Omega, \mathbf{F}, \mathbf{P})$, and we let $\mathbf{E}(\mathrm{z})$ denote expectation of the stochastic variable z. We use bold face letters to indicate stochastic variables/processes. Each euclidean space is associated with the topology generated by the 2-norm $\|\cdot\|_2$, and the Borel σ-algebra generated by the open sets. The induced matrix 2-norm is also denoted $\|\cdot\|_2$. We will call a square matrix a Schur matrix, if all its eigenvalues are inside the unit disk. We denote by I_m the $m \times m$ identity matrix.

Features and labels: Let us fix two stochastic processes: the label process $\mathbf{y}(t) \in \mathbb{R}^{n_y}$, and input process $\mathbf{u}(t) \in \mathbb{R}^{n_u}$, which share a time axis $t \in \mathbb{Z}$, i.e., $\mathbf{y}(t), \mathbf{u}(t)$ are random vectors on $(\Omega, \mathbf{F}, \mathbf{P})$. We want to predict the values of $\mathbf{y}(t)$ based on the past and present values of the inputs $\{\mathbf{u}(s)\}_{s \le t}$ and possibly the past values of $\{\mathbf{y}(s)\}_{s<t}$. We will aim at learning predictors based on the finite sample $\mathcal{D} = \{\mathbf{y}(t)(\omega), \mathbf{w}(t)(\omega)\}_{t=0}^{N-1}$, $\omega \in \Omega$.

Predictors (parameterized linear time-invariant (LTI) systems): In this paper we consider the class of predictors $\mathcal{F}$ such that the following holds. The goal of each predictor $f \in \mathcal{F}$, is to estimate the current label $\mathbf{y}(t)$ based on available features $\mathbf{w}(s) \in \mathbb{W}$, $s \in \mathbb{Z}$ where $\mathbf{w}(s)$ is either:

- $\mathbf{w}(s) = \mathbf{u}(s)$, $\mathbb{W} = \mathbb{R}^{n_u}$ in which case the to be learnt models will predict the current label $\mathbf{y}(t)$ based on the past and current inputs $\{\mathbf{u}(s)\}_{s=t_0}^{t}$, or
- $\mathbf{w}(s) = [\mathbf{y}^T(s),\ \mathbf{u}^T(s)]^T$, $\mathbb{W} = \mathbb{R}^{n_y+n_u}$, in which case, the to be learnt models will use past labels $\{\mathbf{y}(s)\}_{s=t_0}^{t-1}$ in addition to the past and current inputs $\{\mathbf{u}(s)\}_{s=t_0}^{t}$ to predict the current label $\mathbf{y}(t)$.

To unify notation we write the input-output data as $\{\mathbf{y}(i), \mathbf{w}(i)\}_{i \in I}$ for any $I \subseteq \mathbb{Z}$. Clearly, this notation contains redundant information, in the case of feature $\mathbf{w}(s) = [\mathbf{y}^T(s),\ \mathbf{u}^T(s)]^T$.

We will assume that the elements of $\mathcal{F}$ are parametrized by a parameter set Θ, where Θ is a subset of an Eucledian space, and that the elements of $\mathcal{F}$ arise as outputs of linear time-invariant state-space systems driven by $\mathbf{w}$, i.e., every $f \in \mathcal{F}$ is of the form $f = f_\theta$ for some $\theta \in \Theta$, and there exist matrices $\hat{A}(\theta), \hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta)$ of suitable dimensions such that for every sequence $\{\mathbf{w}(k)\}_{k=t_0}^{t}$, $t_0 \le t$

$$\hat{\mathbf{x}}(t+1) = \hat{A}(\theta)\hat{\mathbf{x}}(t) + \hat{B}(\theta)\mathbf{w}(t), \quad \hat{\mathbf{x}}(t_0) = 0 \tag{1a}$$

$$f_\theta(\{\mathbf{w}(k)\}_{k=t_0}^{t}) = \hat{C}(\theta)\hat{\mathbf{x}}(t) + \hat{D}(\theta)\mathbf{w}(t). \tag{1b}$$

If feature $\mathbf{w}(s) = [\mathbf{y}^T(s),\ \mathbf{u}^T(s)]^T$, $\mathbb{W} = \mathbb{R}^{n_y+n_u}$, then we assume that $\hat{D}(\theta)\mathbf{w}(t)$ depends on $\mathbf{u}(t)$ only, i.e., the first n_y columns of $\hat{D}(\theta)$ are zero.

Predictors of the form (1) are used in an online manner, i.e., as soon as $\mathbf{w}(t)$ becomes available, we compute the estimate $f_\theta(\{\mathbf{w}(k)\}_{k=t_0}^{t})$ of $\mathbf{y}(t)$. Intuitively, t_0 represents the starting time from which observations become available. Without loss of generality, during the deployment and learning of predictors, we can assume that $t_0 = 0$. However, for some theoretical considerations, $t_0 \neq 0$ will also be useful.

The celebrated (stationary) Kalman-filters are predictors of the type (1)[1]. Note that the number of data points used to predict $\mathbf{y}(t)$ increases with t, and hence it is unbounded. That is, while we would like to learn a predictor from a sequence of inputs of length N, we deploy the learnt predictor on inputs of arbitrary length during the deployment. This is in contrast to auto-regressive predictors, where only a finite portion of the past values of $\mathbf{w}$ is used to predict $\mathbf{y}(t)$.

Empirical and generalisation losses: In order to quantify the goodness of each predictor $f \in \mathcal{F}$, we shall define two quantities: empirical loss, and generalisation loss. For any predictor $f \in \mathcal{F}$, and data set $\{\mathbf{y}(i), \mathbf{w}(i)\}_{i=0}^{N-1}$ the *empirical loss* of f is defined in the usual manner, i.e.,

$$\hat{\mathcal{L}}_N(f) \triangleq \frac{1}{N}\sum_{i=0}^{N-1} \|f(\{\mathbf{w}(s)\}_{s=0}^{i}) - \mathbf{y}(i)\|_2^2. \tag{2}$$

Note that we use the quadratic loss function, which is a standard choice for regression in general, and time-series prediction in particular. Note that, $\hat{\mathcal{L}}_N(f)$ is a random variable w.r.t. the probability space $(\Omega, \mathbf{F}, \mathbf{P})$.

In classical PAC-Bayesian literature, one defines generalisation loss as $\mathbf{E}[\|f(\{\mathbf{w}(s)\}_{s=0}^{t}) - \mathbf{y}(t)\|_2^2]$. For time-series prediction that choice is not suitable, as it captures the prediction error at time instance t, when only the last $t+1$ inputs are used for prediction. However, during the deployment of predictors, both t and the number of past inputs used for prediction increases. Moreover, as t increases, the prediction error tends to decrease, as more and more data points are used for prediction. This phenomenon can be observed not only for predictors of the form (1), but for more general ones, for instance in non-stationary Kalman-filters[2]

[1] Note that Kalman-filters are not learnt from data, but they are constructed from a known model of the data generator. Nevertheless, they represent a useful analogy.

[2] In non-stationary Kalman-filters, in contrast to the stationary

(Lindquist and Picci, 2015). The latter are known to generate the smallest possible prediction error. For this reason, the standard practice in learning dynamical systems (Ljung, 1999) is to define the *generalisation loss* as the limit:

$$\mathcal{L}(f) \triangleq \lim_{t\to+\infty} \mathbf{E}[\|f(\{\mathbf{w}(k)\}_{k=0}^{t}) - \mathbf{y}(t)\|_2^2] \tag{3}$$

That is, the generalisation loss captures the long-term, steady-state prediction error obtained during the deployment of the predictor on an increasing number of past inputs [3]. If $\mathcal{L}(f)$ is small as $t \to \infty$, then with high probability the prediction error $\|f(\{\mathbf{w}(s)\}_{s=0}^{t}) - \mathbf{y}(t)\|_2^2$ will be small. The existence of the limit on the right-hand side of (3) is a standard result (Hannan and Deistler, 1988; Lindquist and Picci, 2015) and it will be recalled in Lemma 3.4.

Data generators (noisy LTI systems): In addition to specifying the class of predictors, we need assumptions on the data, i.e., we assume that $\mathbf{y}(t)$ is the output of a noisy LTI state-space representation driven by the input $\mathbf{u}$, i.e.,

$$\begin{aligned}\hat{\mathbf{x}}_g(t+1) &= A_0\hat{\mathbf{x}}_g(t) + B_0\mathbf{u}(t) + K_0\mathbf{e}^s(t)\\ \mathbf{y}(t) &= C_0\hat{\mathbf{x}}_g(t) + D_0\mathbf{u}(t) + \mathbf{e}^s(t)\end{aligned} \tag{4}$$

where A_0, B_0, K_0, C_0, D_0 are suitable matrices and $\mathbf{e}^s(t)$ is a zero-mean i.i.d. process. Moreover, $\mathbf{u}$ is generated by a noisy LTI state-space representation, possibly driven by the output process $\mathbf{y}$. That is, we allow $\mathbf{u}(t)$ to depend on past outputs $\{\mathbf{y}(s)\}_{s<t}$. That is, in the terminology of system identification, we consider both the *open-loop* and the *close-loop settings*. In the formal statement of the main result, we will use an assumption which is more general than the existence of (4). This will be done in order to keep the result more general and the statements more streamlined.

This assumption can be viewed as a realizability assumption, as data generator (4) gives rise to the predictor f_{true}

$$\begin{aligned}&\hat{\mathbf{x}}_g(t+1) = \hat{A}_0\hat{\mathbf{x}}_g(t) + \hat{B}_0\mathbf{w}(t),\ \hat{\mathbf{x}}_g(0) = 0\\ &f_{\text{true}}(\{\mathbf{w}(s)\}_{s=0}^{t}) = \hat{C}_0\hat{\mathbf{x}}_g(t) + \hat{D}_0\mathbf{w}(t)\end{aligned} \tag{5}$$

where $\hat{A}_0 = A_0, \hat{B}_0 = B_0, \hat{C}_0 = C_0$ if $\mathbb{W} = \mathbb{R}^{n_u}$ and $\hat{A}_0 = A_0 - K_0C_0$, $\hat{B}_0 = \begin{bmatrix} K_0 & B_0 - K_0D_0\end{bmatrix}$, $\hat{C}_0 = C_0$, $\hat{D}_0 = D_0$ if $\mathbb{W} = \mathbb{R}^{n_u+n_y}$. When $\mathbf{w} = [\mathbf{y}^T(s),\ \mathbf{u}^T(s)]^T$, the predictor f_{true} corresponds to stationary Kalman-filter, and it is optimal in the sense that the generalisation loss $\mathcal{L}(f_{\text{true}})$ is the smallest possible among all the predictors f which arises via a LTI state-space representation (1).

The learning problem: We can then formulate the learning problem considered in this paper as follows.

Problem 2.1 (Learning problem). Find a parameter $\theta_\star$ from the sampled data $\mathcal{D} = \{\mathbf{y}(t)(\omega), \mathbf{w}(t)(\omega)\}_{t=0}^{N-1}$ of the random variables $\{\mathbf{y}(t), \mathbf{w}(t)\}_{t=0}^{N-1}$ such that $\mathcal{L}(\theta_\star)$ is as small as possible. The corresponding optimal predictor $f_{\theta_\star}$ is denoted by $f_\star$.

In particular, if f_{true} corresponds to the element of $\mathcal{F}$ parameterized by the parameter θ_{true}, and the correspondence $\theta \mapsto f_\theta$ is one-to-one (the parameterization is *identifiable*)[4], then $\theta_\star = \theta_{\text{true}}$. That is, *finding the predictor with the smallest generalisation loss* amounts to learning the parameters of the data generator, i.e. *our learning problem is consistent with the standard system identification problem.*

PAC-Bayesian approach, main result: We use the Bayesian perspective, i.e., we start with a prior density π on the hypothesis class $\mathcal{F}$. We then use the sampled data to refine the prior density into a posterior density $\hat{\rho}$, such that the average generalisation loss according to $\hat{\rho}$ is minimal.

Since $\mathcal{F}$ is a class of functions defined on sequences of arbitrary length, the definition of a probability density on $\mathcal{F}$ is not trivial. To circumvent this problem, we use the correspondence between the parameter values from Θ and the elements of $\mathcal{F}$, and we will use densities defined on Θ instead of densities defined on $\mathcal{F}$. Since Θ is a subset of an Eucledian space, densities on Θ can be defined in a classical way. We will need to take expectations of a function g on $\mathcal{F}$ w.r.t. the probability distribution induced by a density ρ on Θ. To this end, we can identify g with the function $\theta \mapsto g(f_\theta)$ on Θ, and take the expectation of the latter function w.r.t. the probability distribution induced by ρ. By a slight abuse of notation, we we will denote the latter expectation by $E_{f\sim\rho}g(f)$. The latter expectation exists under some mild assumptions. We defer the formal definition to Definition 3.5.

With this notation in mind, the posterior $\rho^{\mathcal{D}}$ computed from data $\mathcal{D}$, should be such that $E_{f\sim\rho^{\mathcal{D}}}\mathcal{L}(f)$ is small. Then, following standard practices in PAC-Bayesian learning, problem 2.1 can be solved by either taking $\theta_\star$ such that $f_{\theta_\star}$ is the the mean model according to $\rho^{\mathcal{D}}$, or sampling θ_* randomly from $\rho^{\mathcal{D}}$, or taking the most likely parameter according to $\rho^{\mathcal{D}}$, i.e., $\theta_\star = \arg\max_{\theta\in\Theta} \rho^{\mathcal{D}}(\theta)$.

Since $\mathcal{L}(f)$ is unknown, we could try to minimize the average empirical loss $E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)$ instead. However, this would not allow us to integrate the prior π, and could cause over-fitting. Instead, we derive an upper bound

$$E_{f\sim\rho}\mathcal{L}(f) \le E_{f\sim\rho}\hat{\mathcal{L}}_N(f) + r_N(\rho,\pi)$$

and we choose $\rho^{\mathcal{D}}$ by minimizing $E_{f\sim\rho}\hat{\mathcal{L}}_N(f) + r_N(\rho,\pi)$ over all densities ρ.

one, the matrix $B(\theta)$ may depend on time, so it is not realizable by an LTI system.

[3] In (Ljung, 1999) instead of letting $t \to +\infty$, t is fixed and $k \to -\infty$, which is equivalent, see Lemma 3.4.

[4] Identifiability is a standard assumption in parametric system identification (Ljung, 1999).

This calls for deriving upper bounds for the generalisation gap $E_{f\sim\rho}\mathcal{L}(f) - E_{f\sim\rho}\hat{\mathcal{L}}_N(f)$. In this paper we derive such an error bound. To state the main result, for any two densities ρ and π, we denote by

$$\bar{\mathcal{D}}_2(\rho\|\pi) \triangleq \left(E_{\theta\sim\pi}\left(\frac{\rho(\theta)}{\pi(\theta)}\right)^2\right)^{\frac{1}{2}} \tag{6}$$

the 2 Rényi divergence between ρ and π. Here, $E_{\theta\sim\pi}$ denotes the expectation with respect to the probability measure on Θ induced by π. Let $\mathcal{M}_\pi$ be the family of all densities on Θ, the induced probability measure of which is absolutely continuous w.r.t. the probability measure induced by π.

Theorem 2.1 (Main result, informal)**.** *Let π be any density on the parameter set Θ and let $\delta \in (0,1]$ be arbitrary. Then the following inequality holds with probability at least $1-2\delta$ over the data*

$$\forall\rho\in\mathcal{M}_\pi : E_{f\sim\rho}|\mathcal{L}(f) - \hat{\mathcal{L}}_N(f)| \le r_N(\rho,\pi) \tag{7}$$

$$r_N(\rho,\pi) \triangleq \frac{K}{\sqrt{\delta N}}\bar{\mathcal{D}}_2(\rho\|\pi)\left[G_1 + \frac{4}{\sqrt{N}}G_2\right] \tag{8}$$

where K, G_1 and G_2 are constants which depend on the prior π and the hypothesis class $\mathcal{F}$. In particular, with probability at least $1-2\delta$ over the data,

$$\forall\rho\in\mathcal{M}_\pi : E_{f\sim\rho}\mathcal{L}(f) \le E_{f\sim\rho}\hat{\mathcal{L}}_N(f) + r_N(\rho,\pi) \tag{9}$$

The formal counterpart of Theorem 2.1 is presented in Theorem 4.1.

Remark 2.2 ($O(\frac{1}{\sqrt{N}})$ bound). The bound $r_N(\rho,\pi)$ converges to zero as $N\to\infty$ at rate $O(\frac{1}{\sqrt{N}})$ for fixed ρ and π. That is, for large enough N, it will give a non-trivial guarantee on the generalization loss. For data dependent posteriors the asymptotic behavior of $r_N(\rho,\pi)$ depends on that of the Rényi divergence between the posterior and the prior. The latter could grow with N. However, for reasonable posteriors this is not the case, as they tend to concentrate around the best model, see Section C, Appendix. A formal analysis of the asymptotic behavior of the Rényi divergence remains a topic of future research.

Remark 2.3 (Dependence on δ). As it is customary for bounds based on Rényi divergence (Bégin et al., 2016) $r_N(\rho,\pi)$ depends on $\frac{1}{\sqrt{\delta}}$. This is in contrast to the dependence on $\ln(\frac{1}{\delta})$ of other PAC bounds. This more conservative dependence is the price to pay for Rényi bounds being easier to derive, see Section B, Appendix.

Remark 2.4 (Comparison with prior work). The convergence rate $O(\frac{1}{\sqrt{N}})$ of Theorem 4.1 is comparable with the results of (Alquier and Guedj, 2018) for auto-regressive models. However, it holds for the state-space case. It is also comparable with the finite-sample bounds (Lale et al., 2020; Simchowitz et al., 2019; Sarkar et al., 2021; Tsiamis and Pappas, 2019; Hazan et al., 2018), which apply in a more restricted setting, see Remark 4.4 for mode details.

Remark 2.5 (Application for learning). Various learning algorithms can be derived by finding a data-dependent posterior $\rho^{\mathcal{D}}$ such that $E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)+r_N(\rho^{\mathcal{D}},\pi)$ is small. The term r_N can be viewed as a regularisation term. Interpreting priors or prior dependent expressions as regularisation terms was explored for LTI systems in the book (Pillonetto et al., 2022). For an explicit expression for the minimizer of the right-hand side of (8) see (Alquier and Guedj, 2018).

Once a posterior $\rho^{\mathcal{D}}$ is obtained, there are several standard ways to use it for choosing a parameter $\theta_\star$. Without claiming completeness, we mention the following two possibilities.

For instance, we could take $\theta_\star$ such that $f_{\theta_\star}$ is the mean of f_θ according to $\rho^{\mathcal{D}}$, i.e.

$$\mathbb{E}_{\theta\sim\rho^{\mathcal{D}}} f_\theta(\{\mathbf{w}(s)\}_{s=0}^t) = f_{\theta_\star}(\{\mathbf{w}(s)\}_{s=0}^t) \tag{10}$$

for all $t\ge 0$, see Remark A.8 of Appendix for conditions when this is the case. Then with probability $1-2\delta$,

$$\mathcal{L}(f_{\theta_\star}) \le E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f) + r_N(\rho^{\mathcal{D}},\pi), \tag{11}$$

see Lemma A.7, Appendix for the proof of (11).

Alternatively, we can sample $\theta_\star$ from the posterior $\rho^{\mathcal{D}}$. We can formulate a naive counterpart of (8), as follows: for any $\delta\in(0,0.5)$, $\delta_1\in(0,1)$ with a probability $(1-2\delta)(1-\delta_1)$ over all samples $\theta_\star$ drawn from ρ^θ and over the data

$$\mathcal{L}(f_{\theta_\star}) \le \hat{\mathcal{L}}_N(f_{\theta_\star}) + \frac{1}{\delta_1}r_N(\rho^{\mathcal{D}},\pi) \tag{12}$$

for the formal proof of (12) see Lemma A.6. That is, with a high probability we have an upper bound on the generalisation loss $\mathcal{L}(f_{\theta_\star})$ in terms of the empirical loss $\hat{\mathcal{L}}_N(f_{\theta_\star})$. The derivation of single draw bounds which are less conservative than (12) is not entirely trivial (Hellström et al., 2023), and thus remains a topic of future research.

Remark 2.6 (Relationship with parameter estimation). In system identification literature, one is often interested in the parameter estimation error, i.e., the difference between the learnt model f and the true model f_{true} from (5). In contrast, Theorem 4.1 bounds the the generalisation gap, i.e., the difference $\mathcal{L}(f) - \hat{\mathcal{L}}_N(f)$ between the generalisation loss and the empirical loss $\hat{\mathcal{L}}_N(f)$. The latter depends not only on how far the model f is from the true one f_{true}, but also on the various statistical properties of f and f_{true}. In fact, even if $f = f_{\mathrm{true}}$, the generalisation and empirical losses are not zero, and the generalisation gap need not be zero. However, the PAC-Bayesian bound can be used to bound the parameter estimation error, see Remark 4.4.

3. Formal problem formulation

Assumption 3.1 (Predictors). Given integers n_θ and $\hat{n}$, a compact set $\Theta \subset \mathbb{R}^{n_\theta}$, and a tuple $\Sigma(\theta) \triangleq$

$(\hat{A}(\theta), \hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta))$ of continuous matrix functions $\hat{A} : \mathbb{R}^{n_\theta} \to \mathbb{R}^{\hat{n}\times\hat{n}}, \hat{B} : \mathbb{R}^{n_\theta} \to \mathbb{R}^{\hat{n}\times n_w}, \hat{C} : \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_y\times\hat{n}}, \hat{D} : \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_y\times n_w}$, with $\hat{A}(\theta)$ Schur, and[5] $\hat{D}(\theta) = \begin{bmatrix}0 & \hat{D}_u(\theta)\end{bmatrix}$ for all $\theta \in \Theta$. The class of predictors $\mathcal{F}$ is then given by

$$\mathcal{F} = \left\{ f_\theta : \bigcup_{k=1}^{\infty} \mathbb{W}^k \to \mathbb{R}^{n_y} \mid \theta \in \Theta \right\} \tag{13}$$

with f_θ determined by (1).

That is, predictors f_θ will be functions realised by stable LTI system. We denote a predictor by f when θ is clear from the context. Note that, the predictor f_θ can be identified with either $\Sigma(\theta)$ or simply the parameter θ. Next, we state the formal assumptions on the data generator.

Assumption 3.2 (Data generator). Label $\mathbf{y}(t)$ and input $\mathbf{u}(t)$ are generated by a LTI state-space representation

$$\mathbf{x}_g(t+1) = A_g\mathbf{x}_g(t) + K_g\mathbf{e}_g(t) \tag{14a}$$

$$\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix} = C_g\mathbf{x}_g(t) + \mathbf{e}_g(t) \tag{14b}$$

with $\mathbf{e}_g(t)$ a zero mean, sub-Gaussian, i.i.d. process, and $A_g \in \mathbb{R}^{n\times n}, K_g \in \mathbb{R}^{n\times m}, C_g \in \mathbb{R}^{m\times n}$, $n > 0$, $m = n_y + n_u \geq 2$, and $\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t)$ all stationary, mean square integrable, zero-mean, and A_g and $A_g - K_gC_g$ are Schur matrices, and $\mathbf{e}_g(t)$ is uncorrelated with $\mathbf{x}_g(t-k)$, $k \geq 0$.

From standard theory on stochastic LTI systems (Lindquist and Picci, 2015), it follows that $\mathbf{e}_g(t)$ is the innovation process of $[\mathbf{y}^T(t)\ \mathbf{u}^T(t)]^T$ and (14) is in the so called forward innovation form, see (Lindquist and Picci, 2015) for the definition of these concepts.

Remark 3.3 (Existence of (4) implies Assumption 3.2). Assume that (4) holds, with Schur matrices A_0 and $A_0 - K_0C_0$, $\mathbf{e}^s(t)$ a zero-mean i.i.d. process, and $\hat{\mathbf{x}}_g(t)$ stationary. In the open-loop case, when $\mathbf{u}(t)$ is an ARMA process such that there is no feedback[6] from $\mathbf{y}(t)$ to $\mathbf{u}(t)$, by (Lindquist and Picci, 2015) Assumption 3.2 holds, and by (Eringis et al., 2023b) the matrices A_g, K_g, C_g can be computed from those of (4) and from the ARMA representation of $\mathbf{u}(t)$. In addition, when $\mathbf{e}^s(t) = 0$ then (4) can be viewed as an optimal predictor of $\mathbf{y}(t)$. When $\mathbf{u}(t)$ is generated by an LTI systems driven by $\mathbf{y}(t)$, i.e.,

$$\mathbf{x}_\mathbf{u}(t+1) = A_u\mathbf{x}_\mathbf{u}(t) + K_u\mathbf{y}(t), \quad \mathbf{u}(t) = C_u\mathbf{x}_\mathbf{u}(t) \tag{15}$$

i.e., in the closed-loop case, then again (14) can be obtained by adding (15) to (4), and letting $\mathbf{x}_g(t) = [\hat{\mathbf{x}}_g^T(t)\ , \mathbf{x}_\mathbf{u}^T(t)]^T$.

Recall the definitions of the empirical and generalisation loss from (2) and (3). In the sequel we will need the following interpretation of the generalisation loss.

Lemma 3.4 ((Hannan and Deistler, 1988)**).** *The limit* $\hat{\mathbf{y}}_f(t) \triangleq \lim_{s\to-\infty} f(\{\mathbf{w}(k)\}_{k=s}^t)$ *exists in the mean-square sense, the process* $\hat{\mathbf{y}}_f(t)$ *is stationary, and*

$$\begin{aligned}\mathbf{E}[\|\hat{\mathbf{y}}_f(t) - \mathbf{y}(t)\|_2^2] &= \lim_{\tau\to+\infty} \mathbf{E}[\|f(\{\mathbf{w}(k)\}_{k=0}^{\tau}) - \mathbf{y}(\tau)\|_2^2] \\ &= \lim_{s\to-\infty} \mathbf{E}[\|f(\{\mathbf{w}(k)\}_{k=s}^{t}) - \mathbf{y}(t)\|_2^2]\end{aligned}$$

In particular, the limit of the right-hand side of (3) *exists and* $\mathcal{L}(f) = \mathbf{E}[\|\hat{\mathbf{y}}_f(t) - \mathbf{y}(t)\|_2^2$.

Intuitively, $\hat{\mathbf{y}}_f(t)$ can be interpreted as the prediction of $\mathbf{y}(t)$ generated by the predictor f based on all (infinite) past and present values of the features. The generalisation loss is then the variance of the difference $\hat{\mathbf{y}}_f(t) - \mathbf{y}(t)$. Note that the latter does not depend on t, as $\mathbf{y}(t)$ and $\hat{\mathbf{y}}_f(t)$ are both stationary processes. Note also that $\lim_{t\to\infty}(\hat{\mathbf{y}}_f(t) - f(\{\mathbf{w}(k)\}_{k=0}^t) = 0$, since $\hat{A}(\theta)$ is Schur. Hence for large t, $\hat{\mathbf{y}}_f(t)$ is an approximation of the output of the predictor f.

Next, we formalize the notion of an expectation of a function on $\mathcal{F}$ w.r.t. to a density on the parameter space Θ.

Definition 3.5. Let B_Θ be the σ-algebra of Lebesque-measurable subsets of the parameter set $\Theta \subset \mathbb{R}^{n_\theta}$, and m denote the Lebesque measure on $\mathbb{R}^{n_\theta}$. With the identification $\theta \leftrightarrow f_\theta = f$ in mind we then define

$$\underset{f\sim\rho}{E}\, g(f) \triangleq \int_{\theta\in\Theta} \rho(\theta)g(f_\theta)dm(\theta) \tag{16}$$

where ρ is a probability density function on the measure space (Θ, B_θ, m), and $g : \mathcal{F} \to \mathbb{R}$ is such that $\Theta \ni \theta \mapsto g(f_\theta)$ is measurable and absolutely integrable map. For a probability density π on (Θ, B_θ, m), denote by $\mathcal{M}_\pi$ the family of probability densities on (Θ, B_θ, m) whose probability measure is absolutely continuous w.r.t. that of π.

4. Main Results

In this section we present the formal counterpart of Theorem 2.1. To this end, for each $f = f_\theta \in \mathcal{F}$ and corresponding $\Sigma(\theta) = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ choose $\hat{M} = M(\theta) > 1$ and $\hat{\gamma} = \hat{\gamma}(\theta) \in [0, 1)$, such that $\|\hat{A}^k\|_2 \leq \hat{M}\hat{\gamma}$, and such that the functions $\theta \mapsto M(\theta), \theta \mapsto \hat{\gamma}(\theta)$ are continuous. Define

$$g_p(f) \triangleq \frac{\hat{M}\|\hat{C}\|_2\|\hat{B}\|_2}{1-\hat{\gamma}} \tag{17}$$

$$G_p(f) \triangleq g_p(f)\left(1 + \|\hat{D}\|_2 + g_p(f)\right)\frac{1}{1-\hat{\gamma}} \tag{18}$$

We can think of $G_p(f)$ as an upper-bound on the difference between the infinite past and the finite past response of the

[5] This assumption is necessary, since otherwise we would be using the components of $\mathbf{y}(t)$ to predict $\mathbf{y}(t)$, which is not meaningful.

[6] see Definition 17.1.1. of (Lindquist and Picci, 2015) for the definition of a feedback-free process.

predictor, i.e., between the responses when the predictor was started at time zero with zero initial state or it was already in stead-state at time zero. In other words, $G_p(f)$ measures the robustness of the predictors w.r.t. state disturbances.

Next, we define the norm $G_e(f)$ of the following LTI system which is driven by the noise of the data generator and whose output is the infinite horizon prediction error:

$$\begin{aligned}\tilde{\mathbf{x}}(t+1) &= A_e\tilde{\mathbf{x}}(t) + K_e\mathbf{e}_g(t),\\ \mathbf{y}(t) - \hat{\mathbf{y}}_f(t) &= C_e\tilde{\mathbf{x}}(t) + D_e\mathbf{e}_g(t)\end{aligned} \tag{19}$$

where A_e, K_c, C_e, D_e are defined as follows: Let A_g, K_g, C_g be the matrices of the data generator from (14), then $D_e = I - \hat{D}_w$, $C_e = \begin{bmatrix} C_1 - \hat{D}C_w & -\hat{C} \end{bmatrix}$, and

$$A_e = \begin{bmatrix} A_g & 0 \\ \hat{B}C_w & \hat{A} \end{bmatrix}, \; K_e = \begin{bmatrix} K_g \\ \hat{B}_w \end{bmatrix}$$

where $C_g = \begin{bmatrix} C_1^T & C_2^T \end{bmatrix}^T$ and C_1 has n_y rows and C_2 has n_u rows; and $(C_w, \hat{B}_w, \hat{D}_w) = (C_2, \begin{bmatrix} 0 & \hat{B} \end{bmatrix}, \begin{bmatrix} 0 & \hat{D} \end{bmatrix})$ if $\mathbf{w} = \mathbf{u}$, and $(C_w, \hat{B}_w, \hat{D}_w) = (C_g, \hat{B}, \hat{D})$, if $\mathbf{w} = [\mathbf{y}^T \; \mathbf{u}^T]^T$. Then we define

$$G_e(f) \triangleq \|D_e\|_2 + \sum_{k=0}^{\infty} \|C_e A_e^k K_e\|_2 \tag{20}$$

We can think of $G_e(f)$ as the ℓ_1-norm (Chellaboina et al., 1999) of the error system (19), i.e. the distance between the predictor f and the data generator. In particular, the smaller $G_e(f)$ is, the closer f is to the optimal predictor (5).

Let $\mathbf{e}_g$ be the sub-Gaussian noise process of the data generator (14), and let $\mu_{max}(Q_e) > 0$ be such that for any $t \in \mathbb{Z}$, there exists a Gaussian $\mathbf{z}(t) \sim \mathcal{N}(0, I_{n_y+n_u})$ for which

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r] \le \mu_{\max}(Q_e)^{\frac{r}{2}} \mathbf{E}[\|\mathbf{z}(t)\|_2^r], \quad 0 < r \in \mathbb{N}, \tag{21}$$

If $\mathbf{e}_g \sim \mathcal{N}(0, Q_e)$, $Q_e > 0$ is Gaussian, then $\mu_{max}(Q_e)$ can be taken as the maximal eigenvalue of the covariance $Q_e = \mathbf{E}[\mathbf{e}_g(t)\mathbf{e}_g^T(t)]$, and $\mathbf{z}(t) = Q_e^{-1/2}\mathbf{e}_g(t)$, see for instance Eringis et al. (2023a, Lemma A.1). For the general case, an explicit formula for $\mu_{max}(Q_e)$ is shown in Lemma A.9, Appendix. In both cases, $\mu_{max}(Q_e)\sqrt{n_u + n_y}$ is an upper bound on the variance of $\mathbf{e}_g$.

We can now state the following PAC-Bayesian bound using Rényi divergence.

Theorem 4.1 (Main result, formal version of Theorem 2.1). *Under Assumptions 3.1-3.2 it follows that for any density π on Θ and any $\delta \in (0, 0.5)$,*

$$\begin{aligned}&\mathbf{P}\Big(\Big\{\omega \in \Omega \mid \forall \rho \in \mathcal{M}_\pi : \\ &\mathbb{E}_{f\sim\rho}\Big|\mathcal{L}(f) - \hat{\mathcal{L}}_N(f)(\omega)\Big| \le r_N(\rho, \pi)\Big\}\Big) > 1 - 2\delta\end{aligned} \tag{22}$$

where $r_N(\rho, \pi)$ is as in (8), $\mathcal{M}_\pi$ is as in Definition 3.5, and K, G_1 and G_2 are defined as follows

$$K \triangleq \mu_{\max}(Q_e)\sqrt{(n_u + n_y + 1)!} \tag{23}$$

$$G_1 \triangleq 6n_y \left(E_{f\sim\pi} G_e^4(f)\right)^{\frac{1}{2}} \tag{24}$$

$$G_2 \triangleq \|\Sigma_{gen}\|_{\ell_1}^2 \left(E_{f\sim\pi} G_p^2(f)\right)^{\frac{1}{2}} \tag{25}$$

$$\|\Sigma_{gen}\|_{\ell_1} = \|I_{n_y+n_u}\|_2 + \sum_{k=1}^{\infty} \|C_g A_g^{k-1} K_g\|_2 \tag{26}$$

where $G_p(f)$ is as in (18), $G_e(f)$ is as in (20), and $\mu_{max}(Q_e)$ is as in (21).

Discussion on the constants The constant $\|\Sigma_{gen}\|_{\ell_1}$ is the induced ℓ_1-norm of the generating system (14) (Chellaboina et al., 1999). It exists due to stability of (14): the larger this constant is, the more sensitive the data generator is to the driving noise $\mathbf{e}_g$. The constant $\mu_{max}(Q_e)$ in an upper bound on the level of noise in the data. The term K depends only on the noise covariance of the data generator. The term G_1 is the average difference between the data generator and the predictors, where the average is computed w.r.t. the prior π on the predictors. Intuitively, KG_1 is related to the average (w.r.t. π) generalisation loss. The term G_2 describes the average (w.r.t. π) robustness (ℓ_1-gain) of the predictor class multiplied by the ℓ_1-gain of the predictor. Intuitively, KG_2 bounds the gap between the empirical loss obtained using finitely many past data and infinitely many past data.

Remark 4.2 (Computing G_1 and G_2). It may seem that the knowledge of the generator system might be necessary to evaluate r_N^R. In fact, it is sufficient to have an upper bound $\mathcal{C}$ on $\|\Sigma_{gen}\|_{\ell_1}$, as $G_e(f) \le \|\Sigma_{gen}\|_{\ell_1} + G_p(f)\|\Sigma_{gen}\|_{\ell_1} \le \mathcal{C}(1 + G_p(f))$. Furthermore, if we have an upper bound $G_p = \sup_{\theta\in\mathcal{F}} G_p(f)$, then we can take $G_1 = 6n_y(\mathcal{C} + G_p\mathcal{C})^2$ and $G_2 = \mathcal{C}^2 G_p$.

Proof. As the first step, for every $f \in \mathcal{F}$ let us define the infinite past *empirical loss* $V_N(f)$ as

$$V_N(f) = \frac{1}{N}\sum_{t=0}^{N-1} \|\mathbf{y}(t) - \hat{\mathbf{y}}_f(t)\|_2^2 \tag{27}$$

where $\hat{\mathbf{y}}_f(t)$ is defined in Lemma 3.4. That is, $V_N(f)$ can be viewed as the empirical loss of the predictor f if it has been run for an infinite amount of time in the past. In particular, $\mathbf{E}[V_N(f)] = \mathcal{L}(f)$, i.e., this empirical loss has the usual property that its expectation is the generalisation loss. In system identification literature (Ljung, 1999), empirical loss is defined as $V_N(f)$, and the transient behavior is ignored.

The proof relies on the Rényi's change of measure from Theorem 8 of (Bégin et al., 2016). Let us call a function Z defined on $\Omega \times \mathcal{F}$ a random function if the function $(\omega, \theta) \mapsto$

$Z(\omega, f_\theta)$ is measurable w.r.t. to the Cartesian product $\mathbf{F} \times B_\Theta$ of σ-algebras. If Z is a random function, then $Z(f) : \Omega \ni \omega \mapsto Z(\omega, f)$ is a random variable. Moreover, if Z takes values in $[0, +\infty)$, then for any density ρ on Θ, $\omega \mapsto E_{f\sim\rho} Z(f)(\omega)$ is also a random variable, and $\theta \mapsto \mathbf{E}[Z(f_\theta)]$ is a measurable function w.r.t. B_Θ.

Lemma 4.3. *Let X,Y be random functions defined on $\Omega \times \mathcal{F}$ and $\delta \in [0, 1)$, then it follows that*

$$\mathbf{P}\Big(\Big\{\omega \mid \forall \rho \in \mathcal{M}_\pi :\ \mathbb{E}_{f\sim\rho} |X(f) - Y(f)| (\omega) \le \frac{\bar{\mathcal{D}}_2(\rho\|\pi) \left(\mathbb{E}_{f\sim\pi}\mathbf{E}\left[(X(f) - Y(f))^2\right]\right)^{\frac{1}{2}}}{\delta^{\frac{1}{2}}}\Big\}\Big) \ge 1 - \delta.$$

The proof of Lemma 4.3 is presented in Appendix. Using Lemma 4.3 with $X(f, \omega) = \mathcal{L}(f)$, $Y(f, \omega) = V_N(f, \omega)$, it follows that with probability $1 - \delta$,

$$\forall \rho \in \mathcal{M}_\pi : \mathbb{E}_{f\sim\rho} |\mathcal{L}(f) - V_N(f)| \le \delta^{-\frac{1}{2}} \bar{\mathcal{D}}_2(\rho\|\pi) \left(\mathbb{E}_{f\sim\pi}\mathbf{E}\left[|\mathcal{L}(f) - V_N(f)|^2\right]\right)^{\frac{1}{2}} \tag{28}$$

and by applying Lemma 4.3 with $X(f, \omega) = V_N(f, \omega)$ $Y(f, \omega) = \mathcal{L}_N(f, \omega)$, it follows that with probability $1 - \delta$,

$$\forall \rho \in \mathcal{M}_\pi : \mathbb{E}_{f\sim\rho} \left|V_N(f) - \hat{\mathcal{L}}_N(f)\right| \le \delta^{-\frac{1}{2}} \bar{\mathcal{D}}_2(\rho\|\pi) \left(\mathbb{E}_{f\sim\pi}\mathbf{E}\left[\left|V_N(f) - \hat{\mathcal{L}}_N(f)\right|^2\right]\right)^{\frac{1}{2}}. \tag{29}$$

We prove in Corollary A.5, Appendix, that

$$\mathbf{E}\left[|\mathcal{L}(f) - V_N(f)|^2\right] \le \left(\frac{6n_yK}{\sqrt{N}}\right)^2 (G_e(f))^4, \tag{30}$$

and we prove in Lemma A.2, Appendix, that

$$\mathbf{E}\left[\left|V_N(f) - \hat{\mathcal{L}}_N(f)\right|^2\right] \le \left(\frac{4K\|\Sigma_{gen}\|_{\ell_1}^2}{N}\right)^2 G_p^2(f) \tag{31}$$

By substituting the upper bound (30) into (28), and the upper bound (31) into (29) and applying the union bound, we obtain the statement of the theorem.

The inequalities (30) and (31) are the key to the proof of the theorem. Their proofs rely on structural properties of LTI systems and control theory. □

Remark 4.4 (Bound on parameter estimation error). The bound from Theorem 7 can used for parameter estimation too. To this end, let $\theta_{true} \in \Theta$ be such that predictor $\Sigma(\theta_{true})$ equals the predictor (5) corresponding to the data generator. Let $\rho^{\mathcal{D}}$ be a data-dependent posterior, and assume that $\theta_\star$ either satisfies (10), or $\theta_\star$ is randomly sampled from $\rho^{\mathcal{D}}$. From (11)–(12) respectively, we can then derive bounds in high-probability on the H_2 distance between the predictor $\Sigma(\theta_\star)$ and the predictor $\Sigma(\theta_{true})$, see (84)–(85) in Appendix. Assuming that the LTI systems $\Sigma(\theta)$ are all minimal, we can use Theorem 5.2, Lemma 5.1 of (Oymak and Ozay, 2022) to derive an error bound (in high-probability) on the difference between the matrices of $\Sigma(\theta_\star)$ and that of $\Sigma(\theta_{true})$, see (92)-(94) of the supplementary material for more details.

These bounds involve either the average the empirical error $\mathbb{E}_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)$ or the empirical error $\hat{\mathcal{L}}_N(f_{\theta_\star})$. They apply to any learning algorithm which can be presented as sampling/take average from a posterior, see (Pillonetto et al., 2022) for examples. This is in contrast to (Oymak and Ozay, 2022; Lale et al., 2020; Simchowitz and Foster, 2020), where the upper bounds for parameter estimation error were formulated for a specific learning algorithm. On the downside, the bounds (84)-(85) and (92)-(94) depend on the empirical loss for the learnt posterior, which can be large if the learning algorithm is not good.

Remark 4.5 (Extension to r-Rényi divergence). Theorem 4.1 can be extended to hold for general r-Rényi divergence,i.e., (7) holds with $\delta^{-\frac{1}{2}}\bar{\mathcal{D}}_2(\rho\|\pi)$ replaced by $\delta^{-\frac{1}{r}}\bar{\mathcal{D}}_r(\rho\|\pi)$, where $\bar{\mathcal{D}}_r(\rho\|\pi)$ is the r-Rényi divergence (Bégin et al., 2016) between ρ and π with a suitable choice of K, G_1, G_2. However, it is known that $\bar{\mathcal{D}}_2(\rho\|\pi) \le \bar{\mathcal{D}}_r(\rho\|\pi)$ for all $r > 2$, which means that extensions of (7) to general r-Rényi divergence might lead to looser bounds. For this reason, in this paper we will not pursue this extension.

5. Numerical Example

For the purposes of this numerical example, we shall generate the data via sampling $\mathbf{e}_g(t)$ and propagating it through the generator system The data is generated by (14), such that $n_u = n_y = 1$, $\mathbf{e}_g(t) \sim \mathcal{N}(0, Q_e)$

$$A_g = \begin{bmatrix} 0.16 & -0.30 \\ 0 & -0.05 \end{bmatrix}, \ K_g = \begin{bmatrix} 0.33 & -0.75 \\ 0 & -0.09 \end{bmatrix},$$
$$C_g = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \ Q_e = \begin{bmatrix} 0.054 & 0.018 \\ 0.018 & 0.248 \end{bmatrix}$$

We choose predictors with two states, i.e., $\hat{n} = 2$, and we parameterise all entries in $\Sigma(\theta) = (\hat{A}(\theta), \hat{B}(\theta), \hat{C}(\theta), \hat{D}(\theta))$. In Figure 1, results of a numerical simulation are shown. We make the following choices **(1)** the prior distribution is chosen as $\pi(\theta) = \mathcal{N}(0, 0.02I)$, **(2)** posterior is chosen as the Gibbs distribution $\rho(\theta) = Z^{-1}\pi(\theta)\exp\{-\lambda\hat{\mathcal{L}}_N(f_\theta)\}$, where $Z = E_{f\sim\pi}\exp\{-\lambda\hat{\mathcal{L}}_N(f)\}$ is the normalisation constant and $\lambda = 0.0043$, for the sake of comparison with (Eringis et al., 2023a), and **(3)** $\delta = 0.1$, i.e. the bound holds with probability at least 0.8.

All quantities are approximated Markov-Chain Monte-Carlo simulation. Our bound converges to zero, i.e., $r_N(\rho, \delta) +$

$E_{f\sim\rho}\hat{\mathcal{L}}_N(f)$ converges to $E_{f\sim\rho}\mathcal{L}(f)$. This is in contrast to he bound proposed in (Eringis et al., 2023a) for a similar example, which converged to a non-zero constant. However, our bound remains fairly conservative. Optimizing it by a better choice of priors and parametrizations remains a topic of future research. More details on this example and an example with real-life data are presented in Appendix C.

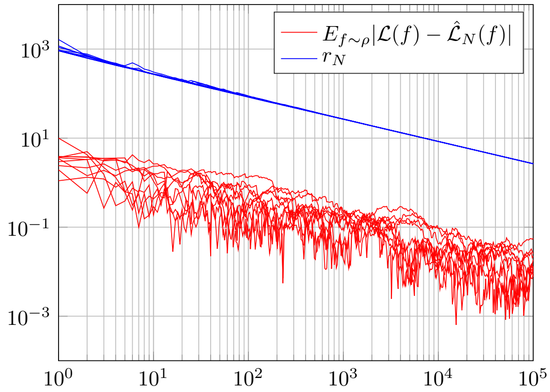

Figure 1. Results of a synthetic example, the case of $\mathbf{w} = \mathbf{u}$, 10 different realisations of data, $r_N = r_N(\rho, \pi)$

6. Conclusion

In this paper we have derived an alternative to KL divergence based PAC-Bayesian bounds for LTI systems. The error bound converges to 0 as the number of samples N grows, with a rate of convergence $O(\frac{1}{\sqrt{N}})$. Future research will be directed towards extending these results to more general state-space representations and using the results of the paper for deriving oracle inequalities (Alquier, 2021).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Akpinar, N.-J., Kratzwald, B., and Feuerriegel, S. (2020). Sample complexity bounds for rnns with application to combinatorial graph problems (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13745–13746.

Alquier, P. (2021). User-friendly introduction to pac-bayes bounds. *arXiv:2110.11216*.

Alquier, P. and Guedj, B. (2018). Simpler PAC-Bayesian Bounds for Hostile Data. *Machine Learning*, 107(5):887–902.

Alquier, P., Li, X., and Wintenberger, O. (2013). Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 1(2013):65–93.

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *JMLR*, 17(239):1–41.

Alquier, P. and Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883 – 913.

Banerjee, I., Rao, V. A., and Honnappa, H. (2021). Pac-bayes bounds on variational tempered posteriors for markov models. *Entropy*, 23(3).

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444. PMLR.

Blanke, M. and Lelarge, M. (2023). Flex: an adaptive exploration algorithm for nonlinear systems. *arXiv preprint arXiv:2304.13426*.

Campi, M. C. and Weyer, E. (2002). Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334.

Chellaboina, V., Haddad, W., Bernstein, D., and Wilson, D. (1999). Induced convolution operator norms for discrete-time linear systems. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*, volume 1, pages 487–492. IEEE.

Chen, M., Li, X., and Zhao, T. (2020). On generalization bounds of a family of recurrent neural networks. In *Proceedings of AISTATS 2020*, volume 108 of *PMLR*, pages 1233–1243.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. OUP Oxford.

Eringis, D., Leth, J., Tan, Z., Wisniewski, R., and Petreczky, M. (2023a). PAC-Bayesian bounds for learning LTI-ss systems with input from empirical loss. *arXiv preprint arXiv:2303.16816*.

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., Esfahan, A. F., and Petreczky, M. (2021). Pac-bayesian theory for stochastic lti systems. In *2021 60th IEEE CDC*, pages 6626–6633.

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., and Petreczky, M. (2023b). Explicit construction of the minimum error variance estimator for stochastic lti-ss systems. *Automatica*, 153:111018.

Eringis, D., Leth, J., Tan, Z.-H., Wisniewski, R., and Petreczky, M. (2023c). Pac-bayes generalisation bounds for dynamical systems including stable rnns. In *To appear in AAAI24*. ArXiv 2312.09793.

Foster, D. and Simchowitz, M. (2020). Logarithmic regret for adversarial online control. In *Proceedings of the 37th ICML*, volume 119 of *PMLR*, page 3211–3221. PMLR.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). Pac-bayesian theory meets bayesian inference. In *NIPS*, pages 1876–1884.

Grünwald, P. (2012). The safe Bayesian - learning the learning rate via the mixability gap. In *ALT*.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces.

Gu, A., Johnson, I., Goel, K., Saab, K. K., Dao, T., Rudra, A., and Re, C. (2021). Combining recurrent, convolutional, and continuous-time models with linear state space layers. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Gu, A., Johnson, I., Timalsina, A., Rudra, A., and Re, C. (2023). How to train your HIPPO: State space models with generalized orthogonal basis projections. In *International Conference on Learning Representations*.

Guedj, B. (2019). A Primer on PAC-Bayesian Learning. *arXiv:1901.05353*.

Haddouche, M. and Guedj, B. (2022a). Online pac-bayes learning. *Advances in Neural Information Processing Systems*, 35:25725–25738.

Haddouche, M. and Guedj, B. (2022b). Pac-bayes with unbounded losses through supermartingales. *arXiv preprint arXiv:2210.00928*.

Hannan, E. and Deistler, M. (1988). *The Statistical Theory of Linear Systems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Hanson, J., Raginsky, M., and Sontag, E. (2021). Learning recurrent neural net models of nonlinear systems. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *PMLR*, pages 425–435. PMLR.

Haussmann, M., Gerwinn, S., Look, A., Rakitsch, B., and Kandemir, M. (2021). Learning partially known stochastic dynamics with empirical pac bayes. *arXiv:2006.09914*.

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. (2018). Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Hellström, F., Durisi, G., Guedj, B., and Raginsky, M. (2023). Generalization bounds: Perspectives from information theory and pac-bayes.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A short note on concentration inequalities for random vectors with subgaussian norm.

Joukovsky, B., Mukherjee, T., Van Luong, H., and Deligiannis, N. (2021). Generalization error bounds for deep unfolding rnns. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *PMLR*, pages 1515–1524. PMLR.

Koiran, P. and Sontag, E. D. (1998). Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. (2020). Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888.

Lee, B. and Lamperski, A. (2019). Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. (2023). Transformers as algorithms: Generalization and stability in in-context learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, page 19565–19594. PMLR.

Li, Z., Han, J., E, W., and Li, Q. (2021). On the curse of memory in recurrent neural networks: Approximation and optimization analysis. In *International Conference on Learning Representations*.

Lindquist, A. and Picci, G. (2015). *Linear Stochastic Systems: A Geometric Approach to Modeling, Estimation and Identification*. Springer.

Ljung, L. (1999). *System Identification: Theory for the user (2nd Ed.)*. PTR Prentice Hall., Upper Saddle River, USA.

Mania, H., Jordan, M. I., and Recht, B. (2022). Active learning for nonlinear system identification with guarantees. *J. Mach. Learn. Res.*, 23:32–1.

Oymak, S. and Ozay, N. (2022). Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., and Ljung, L. (2022). *Regularized system identification: Learning dynamic models from data*. Springer Nature.

Roy, A., Balasubramanian, K., and Erdogdu, M. A. (2021). On empirical risk minimization with dependent and heavy-tailed data. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, page 8913–8926. Curran Associates, Inc.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. (2021). Finite time LTI system identification. *J. Mach. Learn. Res.*, 22:26:1–26:61.

Sattar, Y. and Oymak, S. (2022). Non-asymptotic and accurate learning of nonlinear dynamical systems. *The Journal of Machine Learning Research*, 23(1):6248–6296.

Sattar, Y., Oymak, S., and Ozay, N. (2022). Finite sample identification of bilinear dynamical systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6705–6711.

Sayedana, B., Afshari, M., Caines, P. E., and Mahajan, A. (2022). Consistency and rate of convergence of switched least squares system identification for autonomous markov jump linear systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6678–6685.

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. (2012). Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093.

Shalaeva, V., Esfahani, A. F., Germain, P., and Petreczky, M. (2020). Improved PAC-bayesian bounds for linear regression. *Proceedings of the AAAI Conference*, 34:5660–5667.

Sheth, R. and Khardon, R. (2017). Excess risk bounds for the bayes risk using variational inference in latent gaussian models. In *NIPS*, pages 5151–5161.

Shi, S., Mazhar, O., and De Schutter, B. (2022). Finite-sample analysis of identification of switched linear systems with arbitrary or restricted switching. *IEEE Control Systems Letters*, 7:121–126.

Simchowitz, M. (2021). *Statistical Complexity and Regret in Linear Control*. University of California, Berkeley.

Simchowitz, M., Boczar, R., and Recht, B. (2019). Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR.

Simchowitz, M. and Foster, D. (2020). Naive exploration is optimal for online lqr. In *Proceedings of the 37th ICML*, volume 119 of *PMLR*, page 8937–8948. PMLR.

Sontag, E. D. (1998). A learning result for continuous-time recurrent neural networks. *Systems & control letters*, 34(3):151–158.

Steele, J. M. (2004). *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press.

Tsiamis, A. and Pappas, G. J. (2019). Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654.

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices.

Vidyasagar, M. and Karandikar, R. L. (2006). A learning theory approach to system identification and stochastic adaptive control. *Probabilistic and randomized methods for design under uncertainty*, pages 265–302.

Wang, S. and Xue, B. (2023). State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wei, C. and Ma, T. (2019). Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, J., Lei, Q., and Dhillon, I. (2018). Stabilizing gradients for deep neural networks via efficient SVD parameterization. In *35th ICML*, volume 80 of *PMLR*, pages 5806–5814. PMLR.

Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Information Theory*, 52(4):1307–1321.

Ziemann, I. and Tu, S. (2022). Learning with little mixing. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, page 4626–4637. Curran Associates, Inc.

Ziemann, I. M., Sandberg, H., and Matni, N. (2022). Single trajectory nonparametric learning of nonlinear dynamics. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, page 3333–3364. PMLR.

A. Proofs

Throughout the proof we will use the following notation. For any predictor $f \in \mathcal{F}$ we define the random variable

$$\hat{\mathbf{y}}_f(t \mid t_0) \triangleq f(\{\mathbf{w}(k)\}_{k=t_0}^{t})$$

Intuitively, $\hat{\mathbf{y}}_f(t \mid t_0)$ is the output of (1) at time t, when the predictor was initialised at time t_0 instead of 0, i.e., the past $t - t_0 + 1$ values of $\mathbf{w}$ are used to predict $\mathbf{y}(t)$.

Proof of Lemma 4.3. Recall the Rényi change of measure (Bégin et al., 2016): for any function ϕ on $\mathcal{F}$ such that $\theta \mapsto \phi(f_\theta)$ is measurable,

$$\forall \rho \in \mathcal{M}_\pi : \; \mathbb{E}_{f\sim\rho}\phi(f) \leq \bar{\mathcal{D}}_2(\rho\|\pi)\left(\mathbb{E}_{f\sim\pi}\phi^2(f)\right)^{\frac{1}{2}}.$$

By choosing $\phi(f) \triangleq |X(f) - Y(f)|$

$$\forall \rho \in \mathcal{M}_\pi \; : \; \mathbb{E}_{f\sim\rho}|X(f) - Y(f)| \leq \bar{\mathcal{D}}_2(\rho\|\pi)\left(\mathbb{E}_{f\sim\pi}|X(f) - Y(f)|^2\right)^{\frac{1}{2}} \tag{32}$$

holds. Now by the Markov's inequality we also have

$$\mathbf{P}(\mathbb{E}_{f\sim\pi}|X(f) - Y(f)|^2 < \delta^{-1}\mathbf{E}[\mathbb{E}_{f\sim\pi}|X(f) - Y(f)|^2]) > 1 - \delta \tag{33}$$

Note that by Fubini's theorem $\mathbf{E}[\mathbb{E}_{f\sim\pi}|X(f) - Y(f)|^2] = \mathbb{E}_{f\sim\pi}\mathbf{E}[|X(f) - Y(f)|^2]$. Using this together with (33) we obtain from (32) the desired inequality. □

Lemma A.1. *Let $r \in \mathbb{N}$, $r > 0$, then with notation as above, the following holds.*

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^r\right] \leq \|\Sigma_{gen}\|_{\ell_1}^r G_r(\mathbf{e}_g) \tag{34}$$

with $m = n_u + n_y$, and

$$\|\Sigma_{gen}\|_{\ell_1} = \|I_{n_y+n_u}\|_2 + \sum_{k=1}^{\infty}\|C_g A_g^{k-1}K_g\|_2 \tag{35}$$

$$G_r(\mathbf{e}_g) = \begin{cases} 2^{\frac{r}{2}}\mu_{\max}(Q_e)^{\frac{r}{2}}(m + \frac{r}{2} - 1)!, & r \textit{ is even} \\ 2\mu_{\max}(Q_e)^{\frac{r}{2}}\sqrt{(m + r - 1)!}, & r \textit{ is odd} \end{cases} \tag{36}$$

Proof. The proof of Lemma A.1 is an extension of the proof of Eringis et al. (2023a, Lemma A10) to the case of sub-Gaussian signals. Note that from standard LTI theory (Ljung, 1999) it follows that $\begin{bmatrix}\mathbf{y}^T(t) & \mathbf{u}^T(t)\end{bmatrix}^T$ can be expressed as

$$\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix} = \sum_{k=1}^{\infty} C_g A_g^{k-1}K_g\mathbf{e}_g(t-k) + \mathbf{e}_g(t) = \sum_{k=0}^{\infty}\alpha_k(\mathbf{y},\mathbf{u})\mathbf{e}_g(t-k) \tag{37}$$

with $\mathbf{e}(t)$ stationary, where $\alpha_0(\mathbf{y},\mathbf{u})$ is the identity matrix and $\alpha_k(\mathbf{y},\mathbf{u}) = C_q A_g^{k-1}K_g$, $k > 0$. We can then apply Lemma A.14 to get

$$\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^r\right] \leq \left(\sum_{k=0}^{\infty}\|\alpha_k(\mathbf{y},\mathbf{u})\|_2\right)^r \mathbf{E}\left[\|\mathbf{e}_g(t)\|_2^r\right] \tag{38}$$

Let us denote $\|\Sigma_{gen}\|_{\ell_1} = \sum_{k=0}^{\infty}\|\alpha_k(\mathbf{y},\mathbf{u})\|_2$, the ℓ_1 norm of the generative system. Furthermore we can apply Lemma A.12 and Lemma A.13 to obtain,

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r] \leq G_r(\mathbf{e}_g)$$

$$G_r(\mathbf{e}_g) = \begin{cases} 2^{\frac{r}{2}}\mu_{\max}(Q_e)^{\frac{r}{2}}(m + \frac{r}{2} - 1)!, & r \text{ is even} \\ 2\mu_{\max}(Q_e)^{\frac{r}{2}}\sqrt{(m + r - 1)!}, & r \text{ is odd} \end{cases}$$

with $m = n_y + n_u$ we have the statement of the lemma. □

Lemma A.2. *For each $f \in \mathcal{F}$ choose $\hat{\gamma} \in [\hat{\gamma}^*, 1)$, and $\hat{M} > 1$ such that $\|\hat{A}^k\|_2 \leq \hat{M}\hat{\gamma}^k$ with $\hat{\gamma}^* = \hat{\gamma}^*(\hat{A})$ the spectral radius of $\hat{A}$. Recall the definition of $V_N(f)$ from (27). Then, the following holds*

$$\mathbf{E}[\Big|V_N(f) - \hat{\mathcal{L}}_N(f)\Big|^2] \leq \left(\frac{4}{N}G_{gen}\right)^2 G_p^2(f) \tag{39}$$

with

$$g_p(f) \triangleq \frac{\hat{M}\|\hat{C}\|_2\|\hat{B}\|_2}{1-\hat{\gamma}} \tag{40}$$

$$G_p(f) = g_p(f)\left(1 + \|\hat{D}\|_2 + g_p(f)\right)\frac{1}{1-\hat{\gamma}} \tag{41}$$

$$G_{gen} = \mu_{\max}(Q_e)\|\Sigma_{gen}\|_{\ell_1}^2\sqrt{(n_u+n_y+1)!} \tag{42}$$

Proof of Lemma A.2. The proof of Lemma A.2 is the extension of the proof of Eringis et al. (2023a, Lemma A12) to the case of sub-Gaussian signals and $r = 2$. The latter allows for a simplified proof. The proof goes as follows.

Define $\mathbf{z}_\infty(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_f(t)$, and $\mathbf{z}_{fin}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}_f(t|0)$

$$\mathbf{E}[\Big|V_N(f) - \hat{\mathcal{L}}_N(f)\Big|^2] = \mathbf{E}\left[\left|\frac{1}{N}\sum_{t=0}^{N-1}\|\mathbf{z}_\infty(t)\|_2^2 - \|\mathbf{z}_{fin}(t)\|_2^2\right|^2\right] \tag{43}$$

$$\leq \frac{1}{N^2}\sum_{t_1=0}^{N-1}\sum_{t_2=0}^{N-1}\mathbf{E}\left[\left(\|\mathbf{z}_\infty(t_1)\|_2^2 - \|\mathbf{z}_{fin}(t_1)\|_2^2\right)\left(\|\mathbf{z}_\infty(t_2)\|_2^2 - \|\mathbf{z}_{fin}(t_2)\|_2^2\right)\right] \tag{44}$$

applying Cauchy-Schwarz inequality we obtain

$$\mathbf{E}[\Big|V_N(f) - \hat{\mathcal{L}}_N(f)\Big|^2] \leq \frac{1}{N^2}\sum_{t_1=0}^{N-1}\sum_{t_2=0}^{N-1}\left(\sqrt{\mathbf{E}\left[(\|\mathbf{z}_\infty(t_1)\|_2^2 - \|\mathbf{z}_{fin}(t_1)\|_2^2)^2\right]}\cdot\sqrt{\mathbf{E}\left[(\|\mathbf{z}_\infty(t_2)\|_2^2 - \|\mathbf{z}_{fin}(t_2)\|_2^2)^2\right]}\right) \tag{45}$$

$$= \left(\frac{1}{N}\sum_{t=0}^{N-1}\sqrt{\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2^2 - \|\mathbf{z}_{fin}(t)\|_2^2)^2\right]}\right)^2 \tag{46}$$

For now let us focus on $\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2^2 - \|\mathbf{z}_{fin}(t)\|_2^2)^2\right]$, using the fact that $a^2 - b^2 = (a-b)(a+b)$

$$\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2^2 - \|\mathbf{z}_{fin}(t)\|_2^2)^2\right] = \mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2 - \|\mathbf{z}_{fin}(t)\|_2)^2(\|\mathbf{z}_\infty(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2)^2\right] \tag{47}$$

Applying Cauchy-Schwarz inequality again we get

$$\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2^2 - \|\mathbf{z}_{fin}(t)\|_2^2)^2\right] \leq \underbrace{\sqrt{\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2 - \|\mathbf{z}_{fin}(t)\|_2)^4\right]}}_{a}\underbrace{\sqrt{\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2)^4\right]}}_{b} \tag{48}$$

now we need to upper-bound two terms: a and b.

$$\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2 - \|\mathbf{z}_{fin}(t)\|_2)^4\right] = \mathbf{E}\left[\left(\Big|\|\mathbf{z}_\infty(t)\|_2 - \|\mathbf{z}_{fin}(t)\|_2\Big|\right)^4\right] \leq \mathbf{E}\left[(\|\mathbf{z}_\infty(t) - \mathbf{z}_{fin}(t)\|_2)^4\right] \tag{49}$$

By Lemma A.15 we get

$$\mathbf{E}[\|\mathbf{z}_\infty(t) - \mathbf{z}_{fin}(t)\|_2^4] \leq \left(\hat{M}\|\hat{C}\|_2\|\hat{B}\|_2\frac{\hat{\gamma}^t}{1-\hat{\gamma}}\right)^4\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\\mathbf{u}(t)\end{bmatrix}\right\|_2^4\right] \tag{50}$$

Now coming back to the term $\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2)^4\right]$
First notice that $(\|\mathbf{z}_\infty(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2)^2 \leq 2\|\mathbf{z}_\infty\|_2^2 + 2\|\mathbf{z}_{fin}(t)\|_2^2$, we can see this by first expanding

$$(\|\mathbf{z}_\infty(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2)^2 = \|\mathbf{z}_\infty(t)\|_2^2 + 2\|\mathbf{z}_\infty(t)\|_2\|\mathbf{z}_{fin}(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2^2 \tag{51}$$

and then apply the inequality of arithmetic and geometric means, we get $2\|\mathbf{z}_\infty(t)\|_2\|\mathbf{z}_{fin}(t)\|_2 \leq \|\mathbf{z}_\infty(t)\|_2^2 + \|\mathbf{z}_{fin}(t)\|_2^2$.
Now, similarly we have

$$(2\|\mathbf{z}_\infty(t)\|_2^2 + 2\|\mathbf{z}_{fin}(t)\|_2^2)^2 \leq 2(2\|\mathbf{z}_\infty(t)\|_2^2)^2 + 2(2\|\mathbf{z}_{fin}(t)\|_2^2)^2 = 8\|\mathbf{z}_\infty(t)\|_2^4 + 8\|\mathbf{z}_{fin}(t)\|_2^4 \tag{52}$$

With this, we have

$$\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2)^4\right] \leq \mathbf{E}\left[8\|\mathbf{z}_\infty(t)\|_2^4 + 8\|\mathbf{z}_{fin}(t)\|_2^4\right] = 8\mathbf{E}\left[\|\mathbf{z}_\infty(t)\|_2^4\right] + 8\mathbf{E}\left[\|\mathbf{z}_{fin}(t)\|_2^4\right] \tag{53}$$

By Lemma A.16 and Lemma A.17 we get

$$\mathbf{E}\left[(\|\mathbf{z}_\infty(t)\|_2 + \|\mathbf{z}_{fin}(t)\|_2)^4\right] \leq 16\left(1 + \|\hat{D}\|_2 + \frac{\hat{M}\|\hat{B}\|_2\|\hat{C}\|_2}{1-\hat{\gamma}}\right)^4 \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^4\right] \tag{54}$$

Defining $g_p \triangleq \hat{M}\|\hat{B}\|_2\|\hat{C}\|_2/(1-\hat{\gamma})$, taking (54) and (50) into (48) we get

$$\mathbf{E}\left[\left(\|\mathbf{z}_\infty(t)\|_2^2 - \|\mathbf{z}_{fin}(t)\|_2^2\right)^2\right] \leq 4\hat{\gamma}^{2t}g_p^2\left(1 + \|\hat{D}\|_2 + g_p\right)^2 \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^4\right] \tag{55}$$

Taking the above into (46), and for the sake of notation let

$$B_4 \triangleq \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^4\right]$$

we get

$$\begin{aligned}
&\mathbf{E}[\left|V_N(f) - \hat{\mathcal{L}}_N(f)\right|^2] && (56)\\
&\leq \left(\frac{1}{N}\sum_{t=0}^{N-1}\sqrt{4\hat{\gamma}^{2t}g_p^2\left(1 + \|\hat{D}\|_2 + g_p\right)^2 B_4}\right)^2 && (57)\\
&= \left(\frac{2}{N}g_p\left(1 + \|\hat{D}\|_2 + g_p\right)\sqrt{B_4}\sum_{t=0}^{N-1}\hat{\gamma}^t\right)^2 && (58)\\
&= \left(\frac{2}{N}g_p\left(1 + \|\hat{D}\|_2 + g_p\right)\sqrt{B_4}\frac{1-\hat{\gamma}^N}{1-\hat{\gamma}}\right)^2 && (59)\\
&= \left(\frac{2}{N}g_p\left(1 + \|\hat{D}\|_2 + g_p\right)\frac{1-\hat{\gamma}^N}{1-\hat{\gamma}}\right)^2 B_4 && (60)
\end{aligned}$$

By Lemma A.1 we get

$$\begin{aligned}
\mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^4\right] &\leq \|\Sigma_{gen}\|_{\ell_1}^4 G_4(\mathbf{e}_g) = \|\Sigma_{gen}\|_{\ell_1}^4 4\mu_{\max}(Q_e)^2(n_u + n_y + 1)!\\
&\leq \left(2\|\Sigma_{gen}\|_{\ell_1}^2\mu_{\max}(Q_e)\right)^2(n_u + n_y + 1)!
\end{aligned}$$

With this we obtain the statement of the lemma

$$\begin{aligned}
\mathbf{E}[\left|V_N(f) - \hat{\mathcal{L}}_N(f)\right|^2] &\leq \left(\frac{4g_p}{N}\left(1 + \|\hat{D}\|_2 + g_p\right)\frac{1-\hat{\gamma}^N}{1-\hat{\gamma}}\right)^2\left(\|\Sigma_{gen}\|_{\ell_1}^2\mu_{\max}(Q_e)\sqrt{(n_u + n_y + 1)!}\right)^2\\
&\leq \left(\frac{4g_p}{N}\left(1 + \|\hat{D}\|_2 + g_p\right)\frac{1}{1-\hat{\gamma}}\right)^2\left(\|\Sigma_{gen}\|_{\ell_1}^2\mu_{\max}(Q_e)\sqrt{(n_u + n_y + 1)!}\right)^2
\end{aligned}$$

□

Lemma A.3. *let $m = n_u + n_y$, then for $r \geq 2$, the quantity*

$$\begin{aligned}\sigma(r) &= \max\left\{(\mu_{\max}(Q_e)^r 4(m+r-1)!), (\mu_{\max}(Q_e)^r 3^r (m+r-1)!)\right\} \\ &= \mu_{\max}(Q_e)^r 3^r (m+r-1)!\end{aligned}$$

satisfies

$$\sigma(r) \geq \sup_{t,k,l} \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]$$

$$\mathbf{e}(t,k,j) = \begin{cases} Q_e - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k = j \\ -\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k \neq j \end{cases}$$

Proof. The proof of Lemma A.3 is an extension of the proof of Eringis et al. (2023a, Lemma A.16) to the case of sub-Gaussian processes. First let us take the case when $k \neq j$. Then

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] = \mathbf{E}[\| - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j)\|_2^r]$$

As $\mathbf{e}_g(t)$ is i.i.d. we have

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \mathbf{E}[\|\mathbf{e}_g(t-k)\|_2^r]\mathbf{E}[\|\mathbf{e}_g(t-j)\|_2^r]$$

and due to stationarity of $\mathbf{e}_g(t)$, we have $\mathbf{E}[\|\mathbf{e}_g(t-k)\|_2^r] = \mathbf{E}[\|\mathbf{e}_g(t-j)\|_2^r]$, therefore

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]^2$$

and again due to stationarity of $\mathbf{e}_g(t)$, the moments do not depend on t, and using Lemma A.13 we obtain

$$\sigma(r) \geq \mu_{\max}(Q_e)^r 4((m+r-1)!) \geq \mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r]^2$$

Now let us take the case when $k = j$. Then

$$\begin{aligned}\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] &= \mathbf{E}[\|Q_e - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-k)\|_2^r] \\ &\leq \mathbf{E}[(\|Q_e\|_2 + \|\mathbf{e}_g(t)\|_2^2)^r] \\ &= \mathbf{E}\left[\sum_{j=0}^{r} \binom{r}{j} \|Q_e\|_2^{r-j} \|\mathbf{e}_g(t)\|_2^{2j}\right] \\ &= \sum_{j=0}^{r} \binom{r}{j} \|Q_e\|_2^{r-j} \mathbf{E}\left[\|\mathbf{e}_g(t)\|_2^{2j}\right]\end{aligned}$$

As Q_e is a positive definite matrix,$\|Q_e\|_2 = \mu_{max}(Q_e)$, and hence

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \sum_{j=0}^{r} \binom{r}{j} \mu_{\max}(Q_e)^{r-j} \mathbf{E}\|\mathbf{e}_g(t)\|_2^{2j}]$$

using Lemma A.12 we obtain

$$\mathbf{E}[\|\mathbf{e}(t,k,l)\|_2^r] \leq \mu_{\max}(Q_e)^r \sum_{j=0}^{r} \binom{r}{j} 2^j (m+j-1)!.$$

Since for $j \leq r$, $(m+j-1)! \leq (m+r-1)!$, hence

$$\mathbf{E}\|\mathbf{e}(t,k,l)\|_2^{2r}] \leq \mu_{\max}(Q_e)^r (m+r-1)! \sum_{j=0}^{r} \binom{r}{j} 2^j$$

Notice $3^r = (1+2)^r = \sum_{j=0}^{r} \begin{pmatrix} r \\ j \end{pmatrix} 2^j$, hence

$$\mathbf{E}\|\mathbf{e}_g(t,k,l)\|_2^{2r}] \leq \mu_{\max}(Q_e)^r 3^r (m+r-1)!$$

Hence,

$$\sigma(r) = \max\left\{\mu_{\max}(Q_e)^r 4(m+r-1)!,\ \mu_{\max}(Q_e)^r 3^r(m+r-1)!\right\}.$$

As we are interested in moments higher or equal to two, i.e. $r \geq 2$, then

$$\sigma(r) = \mu_{\max}(Q_e)^r 3^r (m+r-1)!.$$

□

Lemma A.4. *With the notation as above, for any even $r > 0$ the raw moments are bounded*

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \leq \frac{n_y^r}{N}\sigma(r)4(r-1)G_e(f)^{2r} \tag{61}$$

Proof. The proof of Lemma A.4 is an extension of the proof of Eringis et al. (2023a, Lemma A14) to the case of sub-Gaussian processes and even r. For the sake of completeness, we present it below.

Firstly, let $\mathbf{y}_\nu(t), \hat{\mathbf{y}}_{f,\nu}(t), \hat{\mathbf{y}}_{f,\nu}(t|s) \in \mathbb{R}^1$ denote the ν'th component of $\mathbf{y}(t), \hat{\mathbf{y}}_f(t), \hat{\mathbf{y}}_f(t|s)$ respectively, then,

$$\begin{aligned}
\mathcal{L}_\nu(f) &\triangleq \mathbf{E}[(\hat{\mathbf{y}}_{f,\nu}(t) - \mathbf{y}_\nu(t))^2] && (62)\\
&= \lim_{s\to-\infty} \mathbf{E}[(\hat{\mathbf{y}}_{f,\nu}(t|s) - \mathbf{y}_\nu(t))^2] && (63)\\
V_{N,\nu}(f) &\triangleq \frac{1}{N}\sum_{t=0}^{N-1}(\hat{\mathbf{y}}_{f,\nu}(t) - \mathbf{y}_\nu(t))^2 && (64)
\end{aligned}$$

Since the prediction error is the output of the LTI system (19), it follows from standard LTI theory (Ljung, 1999), that the prediction error can be expressed as

$$(\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t)) = \sum_{k=0}^{\infty} \alpha_k \mathbf{e}_g(t-k)$$

with

$$\alpha_k = \alpha_k(\nu) = \begin{cases} D_{e_\nu}, & k = 0 \\ C_{e_\nu} A_e^{k-1} K_e, & k > 0 \end{cases}$$

where $D_{e_\nu} = \mathbf{1}_\nu D_e$, and $C_{e_\nu} = \mathbf{1}_\nu C_e$ denote the ν'th row of matrices D_e, C_e respectively. Then generalised loss $\mathcal{L}_\nu(f)$ for component ν is expressed as

$$\begin{aligned}
\mathcal{L}_\nu(f) &= \mathbf{E}[(\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t))^2]\\
&= \mathbf{E}\left[\left(\sum_{k=0}^{\infty}\alpha_k \mathbf{e}_g(t-k)\right)\left(\sum_{j=0}^{\infty}\alpha_j \mathbf{e}_g(t-j)\right)^T\right]\\
&= \sum_{k=0}^{\infty}\alpha_k Q_e \alpha_k^T,
\end{aligned}$$

Since $\mathbf{E}[\mathbf{e}(t-k)\mathbf{e}^T(t-j)] = 0$, for all $k \neq j$. Now the infinite horizon prediction loss is

$$\begin{aligned} V_{N,\nu}(f) &= \frac{1}{N}\sum_{t=0}^{N-1}(\mathbf{y}_\nu(t) - \hat{\mathbf{y}}_{f,\nu}(t))^2 \\ &= \frac{1}{N}\sum_{t=0}^{N-1}\left(\sum_{k=0}^{\infty}\alpha_k\mathbf{e}_g(t-k)\right)\left(\sum_{j=0}^{\infty}\alpha_j\mathbf{e}_g(t-j)\right)^T \\ &= \frac{1}{N}\sum_{t=0}^{N-1}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\alpha_k\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j)\alpha_j \end{aligned}$$

then

$$\mathcal{L}_\nu(f) - V_{N,\nu}(f) = \frac{1}{N}\sum_{t=0}^{N-1}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\alpha_k\mathbf{e}(t,k,j)\alpha_j^T$$

$$\mathbf{e}(t,k,j) = \begin{cases} Q_e - \mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k = j \\ -\mathbf{e}_g(t-k)\mathbf{e}_g^T(t-j), & k \neq j \end{cases}$$

For ease of notation let us define

$$\mathbf{z}(t,k,j) = \alpha_k\mathbf{e}(t,k,j)\alpha_j^T$$

then

$$\begin{aligned} &\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \\ &= \frac{1}{N^r}\sum_{t_1=0}^{N-1}\cdots\sum_{t_r=0}^{N-1}\sum_{k_1,j_1=0}^{\infty}\cdots\sum_{k_r,j_r=0}^{\infty}\mathbf{E}\left[\prod_{l=1}^{r}\mathbf{z}(t_l,k_l,j_l)\right] \end{aligned}$$

Note that, with i.i.d. innovation noise $\mathbf{e}_g(t)$, if

$$\begin{aligned} &t_r - k_r \notin \{t_i - k_i, t_i - j_i\}_{i=1}^{r-1} \\ &\quad\wedge t_r - j_r \notin \{t_i - k_i, t_i - j_i\}_{i=1}^{r-1} \end{aligned}$$

or similarly

$$\{t_r - k_r, t_r - j_r\} \cap \{t_i - k_i, t_i - j_i\}_{i=1}^{r-1} = \emptyset \tag{65}$$

then $\mathbf{z}(t_r,k_r,j_r)$ is independent of $\mathbf{z}(t_i,k_i,j_i)$. To see this, notice that $\mathbf{z}(t_i,k_i,j_i) = \alpha_{k_i}^T\mathbf{e}(t_i,k_i,j_i)\alpha_{j_i}^T$ and $\mathbf{e}(t_i,k_i,j_i) = \mathbf{E}[\mathbf{e}_g(t_i-k_i)\mathbf{e}_g^T(t_i-j_i)]-\mathbf{e}_g(t_i-k_i)\mathbf{e}_g^T(t_i-j_i)$, $i = 1,\ldots,r$. As the random variables $\mathbf{e}_g(s)$ occurring in $\{\mathbf{z}(t_l,k_l,j_l)\}_{l=1}^{r-1}$ are all different from those occurring in $\mathbf{z}(t_r,k_r,j_r)$, then due to $\mathbf{e}_g$ being and i.i.d. process, the variable $\mathbf{z}(t_r,k_r,j_r)$ is a function of random variables which are independent of those which define $\{\mathbf{z}(t_l,k_l,j_l)\}_{l=1}^{r-1}$, and hence $\mathbf{z}(t_r,k_r,j_r)$ itself is independent of $\{\mathbf{z}(t_l,k_l,j_l)\}_{l=1}^{r-1}$. Moreover, notice that $\mathbf{E}(\mathbf{z}(t_r,k_r,j_r)] = 0$. Indeed, from $\mathbf{e}(t_r,k_r,j_r) = \mathbf{E}[\mathbf{e}_g(t_r-k_r)\mathbf{e}_g^T(t_r-j_r)] - \mathbf{e}_g(t_r-k_r)\mathbf{e}_g^T(t_r-j_r)$ it follows that $\mathbf{e}(t_r,k_r,j_r)$ is the difference between the expected value of $\mathbf{e}_g(t_r-k_r)\mathbf{e}_g^T(t_r-j_r)$ and $\mathbf{e}_g(t_r-k_r)\mathbf{e}_g^T(t_r-j_r)$, hence its expectation is zero. That is, $\mathbf{E}[\mathbf{e}(t_r,k_r,j_r)] = 0$ and thus $\mathbf{E}[\mathbf{z}(t_r,k_r,j_r)] = \alpha_{k_r}^T\mathbf{E}[\mathbf{e}(t_r,k_r,j_r)]\alpha_{j_r}^T = 0$.

Hence, if (65), it holds that

$$\mathbf{E}\left[\prod_{l=1}^{r}z(t_l,k_l,j_l)\right] = \mathbf{E}\left[\prod_{l=1}^{r-1}\mathbf{z}(t_l,k_l,j_l)\right]\underbrace{\mathbf{E}[\mathbf{z}(t_r,k_r,j_r)]}_{=0} = 0. \tag{66}$$

Let us denote

$$\mathcal{Z} = \{t_i - k_i + k_r, t_i - j_i + k_r,\ t_i - k_i + j_r, t_i - j_i + j_r\}_{i=1}^{r-1}.$$

for any choice of $\{t_l, k_l, j_l\}_{l=1}^{r-1}, j_r, k_r$. Notice that $t_r \notin \mathcal{Z}$ implies (65). Then using (66), it follows that

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] = \frac{1}{N^r} \sum_{t_1=0}^{N-1} .. \sum_{t_{r-1}=0}^{N-1} \sum_{k_1,j_1=0}^{\infty} .. \sum_{k_r,j_r=0}^{\infty} \sum_{t_r \in \mathcal{Z}} \mathbf{E}\left[\prod_{l=1}^{r} \mathbf{z}(t_l, k_l, j_l)\right]. \tag{67}$$

Indeed, (66) differs from equation above only in the last sum, where instead of sum $\sum_{t_r=0}^{N-1}$ we take $\sum_{t_r \in \mathcal{Z}}$. However, as it was pointed out above, for all $t_r \notin \mathcal{Z}$ the summands $\mathbf{E}\left[\prod_{l=1}^{r} \mathbf{z}(t_l, k_l, j_l)\right]$ are zero.

Note that

$$\mathbf{E}\left[\prod_{l=1}^{r} \mathbf{z}(t_l, k_l, j_l)\right] \leq \left|\mathbf{E}\left[\prod_{l=1}^{r} \mathbf{z}(t_l, k_l, j_l)\right]\right| \leq \mathbf{E}\left[\prod_{l=1}^{r} |\mathbf{z}(t_l, k_l, j_l)|\right].$$

Let us focus on $|\mathbf{z}(t_i, k_i, j_i)|$: note that $|\mathbf{z}(t_l, k_l, j_l)| \leq \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \|\mathbf{e}(t_l, k_l, j_l)\|_2$, and thus

$$\mathbf{E}\left[\prod_{l=1}^{r} |\mathbf{z}(t_l, k_l, j_l)|\right] \leq \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \mathbf{E}\left[\prod_{l=1}^{r} \|\mathbf{e}(t_l, k_l, j_l)\|_2\right]$$

Then using Arithmetic Mean-Geometric Mean Inequality, (Steele, 2004) we have

$$\mathbf{E}\left[\prod_{l=1}^{r} \|\mathbf{e}(t_l, k_l, j_l)\|\right] \leq \frac{1}{r} \sum_{l=1}^{r} \mathbf{E}[\|\mathbf{e}(t_l, k_l, j_l)\|_2^r] \tag{68}$$

Now, let $\sigma(r)$, be such that the following holds

$$\sigma(r) \geq \sup_{t,k,l} \mathbf{E}[\|\mathbf{e}(t, k, l)\|_2^r], \tag{69}$$

see Lemma A.3. Then, $\frac{1}{r}\sum_{l=1}^{r} \mathbf{E}[\|\mathbf{e}(t_l, k_l, j_l)\|_2^r] \leq \sigma(r)$ and then from (68) it follows that

$$\mathbf{E}\left[\prod_{l=1}^{r} |\mathbf{e}(t_l, k_l, j_l)|\right] \leq \sigma(r) \tag{70}$$

Combining this with (67), it follows that

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_{r-1}=0}^{N-1} \sum_{k_1,j_1=0}^{\infty} \cdots \sum_{k_r,j_r=0}^{\infty} \sum_{t_r \in \mathcal{Z}} \sigma(r) \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \tag{71}$$

and the quantity $\sigma(r)\prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2$ does not depend on t_r. Moreover

$$\sum_{t_r \in \mathcal{Z}} \sigma(r) \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \leq \sigma(r) \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 |\mathcal{Z}|,$$

where $|\mathcal{Z}|$ is the cardinality of the set $\mathcal{Z}$. Note $|\mathcal{Z}| \leq 4(r-1)$, therefore

$$\sum_{t_r \in \mathcal{Z}} \sigma(r) \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \leq \sigma(r) \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 4(r-1),$$

Combining the latter inequality with (71), it follows that

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_{r-1}=0}^{N-1} \sigma(r) 4(r-1) \sum_{k_1,j_1=0}^{\infty} \cdots \sum_{k_r,j_r=0}^{\infty} \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \tag{72}$$

Now notice

$$\begin{aligned} G_{e,\nu}(f)^{2r} &= \left(\sum_{k=0}^{\infty} \|\alpha_k\|_2 \right)^{2r} = \left(\sum_{k,j=0}^{\infty} \|\alpha_k\|_2 \|\alpha_j\|_2 \right)^{r} \\ &= \sum_{k_1,j_1=0}^{\infty} \cdots \sum_{k_r,j_r=0}^{\infty} \prod_{l=1}^{r} \|\alpha_{k_l}\|_2 \|\alpha_{j_l}\|_2 \end{aligned}$$

therefore we obtain

$$\begin{aligned} &\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \\ &\quad \leq \frac{1}{N^r} \sum_{t_1=0}^{N-1} \cdots \sum_{t_{r-1}=0}^{N-1} \sigma(r) 4(r-1) G_{e,\nu}(f)^{2r} \\ &\quad \leq \frac{1}{N^r} N^{r-1} \sigma(r) 4(r-1) G_{e,\nu}(f)^{2r} \\ &\quad \leq \frac{1}{N} \sigma(r) 4(r-1) G_{e,\nu}(f)^{2r} \end{aligned}$$

and since

$$\|\alpha_k(\nu)\| = \begin{cases} \|\mathbf{1}_\nu D_e\|_2 \leq \|D_e\|_2, & k = 0 \\ \|\mathbf{1}_\nu C_e A_e^{k-1} K_e\|_2 \leq \|C_e A_e^{k-1} K_e\|_2, & k > 0 \end{cases}$$

then

$$G_{e,\nu} \leq G_e = \|D_e\|_2 + \sum_{k=1}^{\infty} \|C_e A_e^{k-1} K_e\|_2 \tag{73}$$

and since $2r > 1$ we obtain

$$\mathbf{E}[(\mathcal{L}_\nu(f) - V_{N,\nu}(f))^r] \leq \frac{1}{N} \sigma(r) 4(r-1) G_e(f)^{2r} \tag{74}$$

Now recall that

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] = \mathbf{E}\left[\left(\sum_{\nu=1}^{n_y} \mathcal{L}_\nu(f) - V_{N,\nu}(f) \right)^r \right] \tag{75}$$

$$= \sum_{\nu_1}^{n_y} \cdots \sum_{\nu_r}^{n_y} \mathbf{E}\left[\prod_{i=1}^{r} (\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f)) \right] \tag{76}$$

Then using Arithmetic Mean-Geometric Mean Inequality, (Steele, 2004), we get $\prod_{i=1}^{r} |\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f)| \leq \frac{1}{r} \sum_{i=1}^{r} (\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f))^r$, where we use the fact that r is even and hence $(\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f))^r = |\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f)|^r$, and thus

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \leq \sum_{\nu_1=1}^{n_y} \cdots \sum_{\nu_r=1}^{n_y} \frac{1}{r} \sum_{i=1}^{r} \mathbf{E}\left[(\mathcal{L}_{\nu_i}(f) - V_{N,\nu_i}(f))^r \right] \tag{77}$$

Now bringing (74) into the above we obtain

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^r] \leq \sum_{\nu_1=1}^{n_y} \cdots \sum_{\nu_r=1}^{n_y} \frac{1}{r} \sum_{i=1}^{r} \frac{1}{N} \sigma(r) 4(r-1) G_e(f)^{2r} = \frac{n_y^r}{N} \sigma(r) 4(r-1) G_e(f)^{2r} \tag{78}$$

□

Corollary A.5. *For $r = 2$, Lemma A.4 and Lemma A.3 imply*

$$\mathbf{E}[(\mathcal{L}(f) - V_N(f))^2] \leq \frac{1}{N} (n_u + n_y + 1)! \left(6 n_y \mu_{\max}(Q_e) G_e(f)^2 \right)^2 \tag{79}$$

A.1. Statement and proof of the mean and the single draw inequalities (11)-(12) from Remark 2.5

Let $\rho^{\mathcal{D}}$ be a function on $\Omega \times \Theta$ which is measurable w.r.t. $\mathbf{F} \times B_{\Theta}$ and such for any $\omega \in \Omega$, $\rho^{\mathcal{D}}(\omega) : \Theta \ni \theta \mapsto \rho^{\mathcal{D}}(\omega, \theta)$ is a probability density on Θ. We refer to ρ as a *random density on* Ω. Let B be a measurable subset of $\Theta \times \Omega$ and denote by $P_{\theta \sim \rho^{\mathcal{D}}}(B)$ the random variable $\omega \mapsto P_{\theta \sim \rho^{\mathcal{D}}(\omega)}(\{\theta \mid (\omega, \theta) \in B\})$, where $P_{\theta \sim \rho^{\mathcal{D}}(\omega)}$ is the probability measure induced by the density $\rho^{\mathcal{D}}(\omega)$. Define the probability measure $\mathbf{P} \times P_{\theta \sim \rho^{\mathcal{D}}}$ on $\mathbf{F} \times B_{\Theta}$ as follows:

$$\left(\mathbf{P} \times P_{\theta \sim \rho^{\mathcal{D}}}\right)(B) \triangleq \mathbf{E}[P_{\theta \sim \rho^{\mathcal{D}}}(B)]$$

Lemma A.6. *Consider the assumptions of Theorem 4.1 and the constants from (23)-(26). Assume that $\rho^{\mathcal{D}}$ is a random density such that for all $\omega \in \Omega$, $\rho^{\mathcal{D}}(\omega) \in \mathcal{M}_{\pi}$. Then*

$$\mathbf{P} \times P_{\theta \sim \rho^{\mathcal{D}}}\Big(\{(\omega, \theta) \mid |\mathcal{L}(f_{\theta}) - \hat{\mathcal{L}}_N(f_{\theta})|(\omega) \leq \frac{\bar{\mathcal{D}}_2(\rho^{\mathcal{D}}(\omega)\|\pi)]K}{\sqrt{\delta N}\delta_1}\Big[G_1 + \frac{4}{\sqrt{N}}G_2\Big]\Big) > (1-2\delta)(1-\delta_1)$$

Proof of Lemma A.6. From Theorem 4.1 it follows that (22) holds. Let B be the subset of all $(\theta, \omega) \in \Theta \times \Omega$, such that

$$|\mathcal{L}(f_{\theta}) - \hat{\mathcal{L}}(f_{\theta})|(\omega) \leq \frac{\bar{\mathcal{D}}_2(\rho^{\mathcal{D}}(\omega)\|\pi)]K}{\sqrt{\delta N}\delta_1}\Big[G_1 + \frac{4}{\sqrt{N}}G_2\Big]$$

Let B_1 be the set of all $\omega \in \Omega$, such that

$$|\mathcal{L}(f_{\theta}) - \hat{\mathcal{L}}(f_{\theta})|(\omega) \leq \frac{\bar{\mathcal{D}}_2(\rho^{\mathcal{D}}(\omega)\|\pi)]K}{\sqrt{\delta N}}\Big[G_1 + \frac{4}{\sqrt{N}}G_2\Big]$$

It follows that $\mathbf{P}(B_1) \geq 1 - 2\delta$. For every $\omega \in B_1$ let $B_2(\omega)$ be the set of all $\theta \in \Theta$, such that

$$|\mathcal{L}(f_{\theta}) - \hat{\mathcal{L}}(f_{\theta})|(\omega) \leq \frac{1}{\delta_1}\mathbb{E}_{f \sim \rho^{\mathcal{D}}}|\mathcal{L}(f_{\theta}) - \hat{\mathcal{L}}(f_{\theta})|(\omega)$$

It is easy to see that $\bar{B} = \bigcup_{\omega \in B_1}\{\omega\} \times B_2(\omega)$ is a subset of B. Hence, it is enough to show that $\left(\mathbf{P} \times P_{\theta \sim \rho^{\mathcal{D}}}\right)(\bar{B}) > (1-2\delta)(1-\delta_1)$. To this end, notice that $P_{\theta \sim \rho^{\mathcal{D}}}(\bar{B})(\omega)$ equals to $P_{\theta \sim \rho^{\mathcal{D}}}(B_2(\omega))\chi_{B_1}(\omega)$, where χ_{B_1} is the characteristic function of B_1. From Markov's inequality it follows that $P_{\theta \sim \rho^{\mathcal{D}}}(B_2(\omega)) > 1 - \delta_1$ and hence $P_{\theta \sim \rho^{\mathcal{D}}}(\bar{B})(\omega) > (1-\delta_1)\chi_{B_1}$. Hence, $\left(\mathbf{P} \times P_{\theta \sim \rho^{\mathcal{D}}}\right)(\bar{B}) \geq \mathbf{E}[\chi_{B_1}](1-\delta_1) \geq (1-2\delta)(1-\delta_1)$. □

Lemma A.7. *With the notation and assumptions of Remark 2.5, assume that $\theta_{\star}$ is a random variable, such that for any $\omega \in \Omega$, $t, t_0 \in \mathbb{Z}$, $t \geq t_0$,*

$$f_{\theta_{\star}(\omega)}(\{\mathbf{w}(s)\}_{s=t_0}^{t})(\omega) = \mathbb{E}_{f \sim \rho^{\mathcal{D}}(\omega)} f(\{\mathbf{w}(s)\}_{s=t_0}^{t})(\omega) \tag{80}$$

holds. Then for any $\delta \in (0, 0.5)$,

$$\mathbf{P}\Big(\{\omega \mid \mathcal{L}(f_{\theta_{\star}}) \leq E_{f \sim \rho^{\mathcal{D}}(\omega)}(\hat{\mathcal{L}}_N(f)(\omega) + r_N(\rho^{\mathcal{D}}(\omega), \pi)\}\Big) > 1 - 2\delta \tag{81}$$

Proof of Lemma A.7. Since for any $\omega \in \Omega$, $f_{\theta_{\star}(\omega)}(\{\mathbf{w}(s)(\omega)\}_{s=t_0}^{t}) = \mathbb{E}_{f \sim \rho^{\mathcal{D}}(\omega)} f(\{\mathbf{w}(s)\}_{s=t_0}^{t})(\omega)$, by Jensen's inequality it follows that

$$\begin{aligned}
&\mathbb{E}_{f \sim \rho^{\mathcal{D}}(\omega)}\|f(\{\mathbf{w}(s)\}_{s=t_0}^{t}) - \mathbf{y}(t)\|_2^2(\omega) \geq \\
&\|\mathbb{E}_{f \sim \rho^{\mathcal{D}}(\omega)} f(\{\mathbf{w}(s)\}_{s=t_0}^{t})(\omega) - \mathbf{y}(t)(\omega)\|_2^2 = \\
&\|f_{\theta_{\star}(\omega)}(\{\mathbf{w}(s)\}_{s=t_0}^{t})(\omega) - \mathbf{y}(t)(\omega)\|_2^2
\end{aligned}$$

Hence, $\mathbb{E}_{f \sim \rho^{\mathcal{D}}(\omega)}\hat{\mathcal{L}}_N(f)(\omega) \geq \hat{\mathcal{L}}_N(f_{\theta_{\star}(\omega)})(\omega)$, and by taking limits as $t \to +\infty$ and using (3) it follows that $\mathbb{E}_{f \sim \rho^{\mathcal{D}}(\omega)}\mathcal{L}(f)(\omega) \geq \mathcal{L}(f_{\theta_{\star}(\omega)})(\omega)$ for almost all $\omega \in \Omega$. The statement of the lemma follows now from (8), by choosing $\rho = \rho^{\mathcal{D}}(\omega)$ for every $\omega \in \Omega$ and using $\mathbb{E}_{f \sim \rho^{\mathcal{D}}(\omega)}\mathcal{L}(f)(\omega) \geq \mathcal{L}(f_{\theta_{\star}(\omega)})(\omega)$. □

Remark A.8. In order for (80) to hold, it is sufficient that

$$\hat{C}(\theta_{\star})\hat{A}(\theta_{\star})^k\hat{B}(\theta_{\star}) = \mathbf{E}_{\theta \sim \rho^{\mathcal{D}}(\omega)}\hat{C}(\theta)\hat{A}(\theta)^k\hat{B}(\theta)$$

for all $k \geq 0$, and $\hat{D}(\theta_{\star}) = \mathbf{E}_{\theta \sim \rho^{\mathcal{D}}(\omega)}D(\theta)$. This is the case, for instance, if the transfer function $H_{f_{\theta}}$ of the LTI system $(A(\theta), B(\theta), C(\theta), D(\theta))$ is linear in θ and $\theta_{\star}(\omega) = \mathbf{E}_{\theta \sim \rho^{\mathcal{D}}(\omega)}\theta$ for all ω.

Inequalities mentioned in Remark 4.4 and further inequalities

As the first step, lett $H_{f_{\mathrm{true}}}$ be the transfer function of the predictor f_{true} of the form (5) arising from the data generator (4). For any $f = f_\theta \in \mathcal{F}$, denote by H_f the transfer function of the corresponding LTI system $\Sigma(\theta)$ Let $\Phi_{\mathbf{w}}$ be the spectral density of the process $\mathbf{w}$ and assume that $\Phi_{\mathbf{w}}(iz) \geq m_{\mathbf{w}} I$ for all $z \in [-\pi, \pi]$. It can be shown (see below) that

$$\|H_{f_{\mathrm{true}}} - H_f\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}(\mathcal{L}(f) - \sigma_{\mathbf{e}^s}^2) \tag{82}$$

where $\sigma_{\mathbf{e}^s}^2$ is the variance of the noise process $\mathbf{e}^s$ and $\|\cdot\|_{H_2}$ denotes the H_2 norm of a transfer function. Note that $\mathcal{L}(f) \geq \sigma_{\mathbf{e}^s}^2$ for any $f \in \mathcal{F}$ and if $\mathbf{w} = [\mathbf{y}^T, \mathbf{u}^T]^T$, then $\mathcal{L}(f_{\mathrm{true}}) = \sigma_{\mathbf{e}^s}^2$

Proof of (82) Note that $\mathbf{y}(t) = \mathbf{e}_s(t) + \hat{\mathbf{y}}_{f_\star}(t)$ and $\hat{\mathbf{y}}_{f_\star}(t)$ and $\hat{\mathbf{y}}_f(t)$ are outputs of LTI systems with transfer functions $H_{f_{\mathrm{true}}}$ and H_f applied to the input $\mathbf{x}$. Moreover, notice that $\mathbf{e}_s(t)$ is uncorrelated with past values of $\mathbf{x}$ and hence it is uncorrelated with $\hat{\mathbf{y}}_{f_\star} - \hat{\mathbf{y}}_f(t)$, as the latter belongs to the Hilbert-space of square-integrable random variables generated by the past values of $\mathbf{x}$ (Lindquist and Picci, 2015). Hence,

$$\begin{aligned}\mathcal{L}(f) &= \mathbf{E}[\|y(t) - \hat{\mathbf{y}}_{f_\star}(t)\|_2^2] \\ &= \mathrm{trace}(\mathbf{E}[(\mathbf{e}_s(t) + \hat{\mathbf{y}}_{f_\star}(t) - \hat{\mathbf{y}}_f(t)) \\ &\qquad\qquad (\mathbf{e}_s(t) + \hat{\mathbf{y}}_{f_\star}(t) - \hat{\mathbf{y}}_f(t))^T]) \\ &= \mathrm{trace}(\mathbf{E}[\mathbf{e}_s(t)\mathbf{e}_s^T(t)]) \\ &\quad + \mathrm{trace}(\mathbf{E}[(\hat{\mathbf{y}}_{f_\star}(t) - \hat{\mathbf{y}}_f(t))(\hat{\mathbf{y}}_{f_\star}(t) - \hat{\mathbf{y}}_f(t))^T]).\end{aligned}$$

Note that $\hat{\mathbf{y}}_{f_\star}(t) - \hat{\mathbf{y}}_f(t)$ is the output of the LTI system with the transfer function $H_\Delta \triangleq H_{f_{\mathrm{true}}} - H_f$ which is driven by the input $\mathbf{x}$. Hence, from the standard properties of LTI systems it follows that

$$\mathbf{E}[(\hat{\mathbf{y}}_{f_\star}(t) - \hat{\mathbf{y}}_f(t))(\hat{\mathbf{y}}_{f_\star}(t) - \hat{\mathbf{y}}_f(t))^T] = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_\Delta(iz)\Phi_{\mathbf{w}}(iz)H_\Delta(iz)^* dz.$$

Hence,

$$\mathcal{L}(f) = \mathrm{trace}(\mathbf{E}[\mathbf{e}_s(t)\mathbf{e}_s^T(t)]) + \mathrm{trace}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} H_\Delta(iz)\Phi_{\mathbf{w}}(iz)H_\Delta(iz)^* dz\right).$$

Note that $\mathrm{trace}(\mathbf{E}[\mathbf{e}_s(t)\mathbf{e}_s^T(t)]) = \sigma_{\mathbf{e}_s}^2$ and

$$H_\Delta(iz)\Phi_{\mathbf{w}}(iz)H_\Delta(iz)^* > m_{\mathbf{w}} H_\Delta(iz)H_\Delta(iz)^*, \quad \forall z \in [-\pi, \pi]$$

Hence,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} H_\Delta(iz)\Phi_{\mathbf{w}}(iz)H_\Delta(iz)^* dz \geq m_{\mathbf{w}}\frac{1}{2\pi}\int_{-\pi}^{\pi} H_\Delta(iz)H_\Delta(iz)^* dz$$

and therefore

$$\mathrm{trace}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} H_\Delta(iz)\Phi_{\mathbf{w}}(iz)H_\Delta(iz)^* dz\right) \geq m_{\mathbf{w}}\mathrm{trace}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} H_\Delta(iz)H_\Delta(iz)^* dz\right) = m_{\mathbf{w}}\|H_\Delta\|_{H_2}^2. \tag{83}$$

Hence,

$$\mathcal{L}(f) \leq \sigma_{\mathbf{e}_s}^2 + m_{\mathbf{w}}\|H_{f_{\mathrm{true}}} - H_f\|_{H_2}^2$$

and therefore

$$\|H_{f_{\mathrm{true}}} - H_f\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}(\mathcal{L}(f) - \sigma_{\mathbf{e}^s}^2)$$

i.e., (82) holds.

Inequalities on $\|H_{f_{true}} - H_{f_{\theta_\star}}\|_{H_2}$.

If $\theta_\star$ is chosen so that (10), then from (11) it follows with probability $1-2\delta$ over data

$$\|H_{f_{\theta_{true}}} - H_{f_{\theta_\star}}\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}\left(r_N(\rho^{\mathcal{D}},\pi) + \left(\mathbb{E}_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f) - \sigma_{\mathbf{e}^s}^2\right)\right) \tag{84}$$

and in the latter case, from (12) it follows that with probability $(1-2\delta)(1-\delta_1)$ over data and $\hat{\rho}^{\mathcal{D}}$,

$$\|H_{f_{\theta_{true}}} - H_{f_{\theta_\star}}\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}\left(\frac{1}{\delta_1}r_N(\rho^{\mathcal{D}},\pi) + \left(\hat{\mathcal{L}}_N(f_{\theta_\star}) - \sigma_{\mathbf{e}^s}^2\right)\right) \tag{85}$$

Note that $\hat{\mathcal{L}}_N(f_{\theta_\star}) \geq \sigma_{\mathbf{e}^s}^2$, in fact, $\sigma_{\mathbf{e}^s}^2$ represents a lower bound on the empirical loss and the minimal generalisation loss possible. The bounds (84)-(85) say that the closer the empirical loss is to the minimal possible value $\sigma_{\mathbf{e}^s}^2$, the closer the estimated system is to the true one in the H_2 norm.

Proof of (84) Assume that $\theta_\star$ is a random variable such that (80) holds. We will show that (84) holds, i.e.,

$$\mathbf{P}\left(\left\{\omega \mid \|H_{f_{\theta_{true}}} - H_{f_{\theta_\star(\omega)}}\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}\left(E_{f\sim\rho^{\mathcal{D}}(\omega)}(\hat{\mathcal{L}}_N(f))(\omega) - \sigma_{\mathbf{e}^s}^2 + r_N(\rho^{\mathcal{D}}(\omega),\pi)\right)\right\}\right) > 1-2\delta. \tag{86}$$

To this end, if $\omega \in B = \{\omega \mid \mathcal{L}(f_{\theta_\star(\omega)}) \leq E_{f\sim\rho^{\mathcal{D}}(\omega)}(\hat{\mathcal{L}}_N(f)(\omega) + r_N(\rho^{\mathcal{D}}(\omega)\}$, then by (82),

$$\|H_{f_{\theta_{true}}} - H_{f_{\theta_\star(\omega)}}\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}\left(E_{f\sim\rho^{\mathcal{D}}(\omega)}(\hat{\mathcal{L}}_N(f))(\omega) - \sigma_{\mathbf{e}^s}^2 + r_N(\rho^{\mathcal{D}}(\omega)\right)$$

from which the claim follows using (81).

Proof of (85) We will show that

$$\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\Big(\left\{(\omega,\theta) \mid \|H_{f_\theta} - H_{f_{\theta_{true}}}\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}\left(\hat{\mathcal{L}}_N(f_\theta)(\omega) - \sigma_{\mathbf{e}_s}^2 + \frac{1}{\delta_1}r_N(\rho^{\mathcal{D}}(\omega),\pi)\right)\right\}\Big) > (1-2\delta)(1-\delta_1) \tag{87}$$

To this end, from (82) it follows that for any (ω,θ) such that $|\mathcal{L}(f_\theta) - \hat{\mathcal{L}}_N(f_\theta)|(\omega) \leq \frac{1}{\delta_1}r_N(\rho^{\mathcal{D}}(\omega),\pi)$, it holds that

$$\|H_{f_\theta} - H_{f_{\theta_{true}}}\|_{H_2}^2 \leq \frac{1}{m_{\mathbf{w}}}\left(\hat{\mathcal{L}}_N(f_\theta)(\omega) - \sigma_{\mathbf{e}_s}^2 + r_N(\rho^{\mathcal{D}}(\omega),\pi)\right)$$

and hence by Lemma A.6 the inequality follows.

Inequalities for state-space matrices From (84)–(85) we derive the inequalities for the difference between the matrices of the true and estimated LTI systems.

To this end, let $T = 2\hat{n}+1$, where $\hat{n}$ is the common state-space dimension of the LTI systems $\Sigma(\theta)$, $\theta\in\Theta$ representing our class of predictors.

Let $\mathcal{H}(\theta)$ be the $T\times T$ Hankel-matrix of $\Sigma(\theta)$ and let $\mathcal{H}^-(\theta)$ formed by the first $2\hat{n}$ block columns of size $n_y \times n_w$. From Lemma 5.2 of(Oymak and Ozay, 2022) it follows that

$$\|\mathcal{H}^-(\theta) - \mathcal{H}^-(\theta_{true})\| \leq \sqrt{2\hat{n}+1}\|H_{f_\theta} - H_{f_{\theta_{true}}}\|_{H_2} \tag{88}$$

Let $\sigma_{min}(\mathcal{H}^-(\theta_{true}))$ be the minimal singular value of $\mathcal{H}^-(\theta_{true})$. Let us assume that for any θ, $\Sigma(\theta)$ is a minimal LTI system. From Theorem 5.2 of (Oymak and Ozay, 2022) it follows that if

$$\|\mathcal{H}^-(\theta) - \mathcal{H}^-(\theta_{true})\|_2 \leq \sigma_{min}(\mathcal{H}^-(\theta_{true}))/4 \tag{89}$$

then there exist a constant $C>0$ and a unitary matrix $T(\theta)$ such that for any

$$\begin{aligned}
&\max\{\|\hat{C}(\theta)T(\theta)^{-1} - \hat{C}(\theta_{true}\|_F, \|T(\theta)\hat{B}(\theta) - \hat{B}(\theta_{true})\|_F\} \leq \sqrt{C\hat{n}}\sqrt{2\hat{n}+1}\frac{\|H_{f_\theta} - H_{f_{\theta_{true}}}\|_{H_2}}{(\sigma_{min}(\mathcal{H}^-(\theta_{true})))^{1/2}} \\
&\|T(\theta)\hat{A}(\theta)T(\theta)^{-1} - \hat{A}(\theta_{true})\|_F \leq \frac{\sqrt{C\hat{n}}\sqrt{2\hat{n}+1}\|H_{f_\theta} - H_{f_{\theta_{true}}}\|_{H_2}\left(2\|\mathcal{H}(\theta_{true})\|_2 + \sigma_{min}(\mathcal{H}^-(\theta_{true}))\right)}{(\sigma_{min}(\mathcal{H}^-(\theta_{true})))^2} \\
&\|\hat{D}(\theta) - \hat{D}(\theta_{true})\|_F \leq \|H_{f_\theta} - H_{f_{\theta_{true}}}\|_{H_2}
\end{aligned} \tag{90}$$

From (88)–(90) it follows that if $\rho^{\mathcal{D}}$ is such that $E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f) + r_N(\rho^{\mathcal{D}},\pi)$ is sufficiently small with high probability, then we can establish high probability bounds for $\|\hat{C}(\theta_\star)T(\theta_\star)^{-1} - \hat{C}(\theta_{true})\|_F$, $\|T(\theta_\star)\hat{B}(\theta_\star) - \hat{B}(\theta_{true})\|_F$ and $\|T(\theta_\star)\hat{A}(\theta_\star)T(\theta_\star)^{-1} - \hat{A}(\theta_{true})\|_F$ for a suitable unitary matrix $T(\theta_\star)$, where $\theta_\star$ is the parameter learned based on $\rho^{\mathcal{D}}$.

To this end, assume that $\rho^{\mathcal{D}}$ is such that for some $\delta_2 \in (0,1)$ and $\delta_1 \in (0,1)$

$$\begin{aligned}
&\mathbf{P}\Big(\Big\{\omega \mid E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)(\omega) + r_N(\rho^{\mathcal{D}}(\omega),\pi) < \\
&\left(\frac{1}{4\sqrt{2\hat{n}+1}}\sigma_{min}(\mathcal{H}^-(\theta_{true})) + \sigma^2_{\mathbf{e}^s}\right)^2 m_{\mathbf{w}}\delta_1\Big\}\Big) > 1-\delta_2
\end{aligned} \tag{91}$$

First, let us consider the case when $\theta_\star$ is chosen such that (11) holds. From (86), (90) and (91) and the union bound it follows

$$\begin{aligned}
&\mathbf{P}\left(\{\omega \mid \exists T \text{ unitary matrix such that,}\right. \\
&\max\{\|\hat{C}(\theta_\star(\omega))T^{-1} - \hat{C}(\theta_{true})\|_F, \|T\hat{B}(\theta_\star(\omega)) - \hat{B}(\theta_{true})\|_F\} \\
&\leq \sqrt{C\hat{n}}\sqrt{2\hat{n}+1}\frac{\left(\frac{1}{m_{\mathbf{w}}}\left[E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)(\omega) - \sigma^2_{\mathbf{e}^s} + r_N(\rho^{\mathcal{D}}(\omega),\pi)\right]\right)^{\frac{1}{2}}}{(\sigma_{min}(\mathcal{H}^-(\theta_{true})))^{1/2}} \\
&\|T\hat{A}(\theta_\star(\omega))T^{-1} - \hat{A}(\theta_{true})\|_F \\
&\leq \frac{\sqrt{Cn}\sqrt{2\hat{n}+1}\left(\frac{1}{m_{\mathbf{w}}}\left[E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)(\omega) - \sigma^2_{\mathbf{e}^s} + r_N(\rho^{\mathcal{D}}(\omega),\pi)\right]\right)^{\frac{1}{2}}(2\|\mathcal{H}(\theta_{true})\|_2 + \sigma_{min}(\mathcal{H}^-(\theta_{true})))}{(\sigma_{min}(\mathcal{H}^-(\theta_{true})))^2}, \\
&\|\hat{D}(\theta_\star(\omega)) - \hat{D}(\theta_{true})\|_F \leq \left(\frac{1}{m_{\mathbf{w}}}\left[E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)(\omega) - \sigma^2_{\mathbf{e}^s} + r_N(\rho^{\mathcal{D}}(\omega),\pi)\right]\right)^{\frac{1}{2}} \\
&\}) > 1-2\delta-\delta_2
\end{aligned} \tag{92}$$

If $\theta_\star$ is randomly drawn from $\rho^{\mathcal{D}}$ we can derive an upper bounds on the estimation error as follows. From Markov inequality and (91) it follows that

$$\begin{aligned}
&\left(\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\right)\Big(\Big\{(\omega,\theta) \mid \hat{\mathcal{L}}_N(f_\theta)(\omega) + \frac{1}{\delta_1}r_N(\rho^{\mathcal{D}}(\omega),\pi) \\
&< \left(\frac{1}{4\sqrt{2\hat{n}+1}}\sigma_{min}(\mathcal{H}^-(\theta_{true})) + \sigma^2_{\mathbf{e}^s}\right)^2 m_{\mathbf{w}}\Big\}\Big) > (1-\delta_2)(1-\delta_1)
\end{aligned} \tag{93}$$

Then from (93), (87) and the union bound it follows that

$$\begin{aligned}
&\left(\mathbf{P}\times P_{\theta\sim\rho^{\mathcal{D}}}\right)(\{\omega \mid \exists T \text{ unitary matrix such that,} \\
&\max\{\|\hat{C}(\theta_\star(\omega))T^{-1} - \hat{C}(\theta_{true})\|_F, \|T\hat{B}(\theta_\star(\omega)) - \hat{B}(\theta_{true})\|_F\} \\
&\leq \sqrt{C\hat{n}}\sqrt{2\hat{n}+1}\frac{\left(\frac{1}{m_{\mathbf{w}}}\left[\hat{\mathcal{L}}_N(f_\theta)(\omega) - \sigma^2_{\mathbf{e}^s} + \frac{1}{\delta_1}r_N(\rho^{\mathcal{D}}(\omega),\pi)\right]\right)^{\frac{1}{2}}}{(\sigma_{min}(\mathcal{H}^-(\theta_{true})))^{1/2}} \\
&\|T\hat{A}(\theta_\star(\omega))T^{-1} - \hat{A}(\theta_{true})\|_F \\
&\leq \sqrt{C\hat{n}}\sqrt{2\hat{n}+1}\frac{\left(\frac{1}{m_{\mathbf{w}}}\left[\hat{\mathcal{L}}_N(f_\theta)(\omega) - \sigma^2_{\mathbf{e}^s} + \frac{1}{\delta_1}r_N(\rho^{\mathcal{D}}(\omega),\pi)\right]\right)^{\frac{1}{2}}(2\|\mathcal{H}(\theta_{true})\|_2 + (\sigma_{min}(\mathcal{H}^-(\theta_{true})))}{(\sigma_{min}(\mathcal{H}^-(\theta_{true}))^2}, \\
&\|\hat{D}(\theta_\star(\omega)) - \hat{D}(\theta_{true})\|_F \leq \left(\frac{1}{m_{\mathbf{w}}}\left[\hat{\mathcal{L}}_N(f_\theta)(\omega) - \sigma^2_{\mathbf{e}^s} + \frac{1}{\delta_1}r_N(\rho^{\mathcal{D}}(\omega),\pi)\right]\right)^{\frac{1}{2}} \\
&\}) > (1-2\delta-\delta_2)(1-\delta_1)
\end{aligned} \tag{94}$$

That is, with probability $1-2\delta-\delta_2$ over data (when $\theta_\star$ satisfies (11)) or with probability $(1-2\delta-\delta_2)(1-\delta_1)$ over data and parameters (when $\theta_\star$ is randomly sampled from $\rho^{\mathcal{D}}$), the difference between the matrices of the true system and those of an isomorphic copy of the estimated system can be bounded by above by an expression which is

$$O\left(\left(E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)-\sigma^2_{\mathbf{e}^s}+r_N(\rho^{\mathcal{D}},\pi)\right)^{1/2}\right) \text{ or } O\left(\left(\hat{\mathcal{L}}_N(f_{\theta_\star})-\sigma^2_{\mathbf{e}^s}+\frac{1}{\delta_1}r_N(\rho^{\mathcal{D}},\pi)\right)^{1/2}\right)$$

i.e., which decreases as $E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)+r_N(\rho^{\mathcal{D}}(\omega),\pi)$ or $\hat{\mathcal{L}}_N(f_{\theta_\star})+\frac{1}{\delta_1}r_N(\rho^{\mathcal{D}},\pi)$ decreases.

If $E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)$ (resp. $\hat{\mathcal{L}}_N(f_{\theta_\star})$ with high probability w.r.t. $\rho^{\mathcal{D}}$) tends to $\sigma^2_{\mathbf{e}^s}$ as $N\rightarrow\infty$, then the upper bounds tends to zero as $N\rightarrow+\infty$. Note that when $\mathbf{w}=[\mathbf{y}^T,\mathbf{u}^T]^T$, then $\sigma^2_{\mathbf{e}^s}=\mathcal{L}(\theta_{true})$, and hence $E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)$ (resp. $\hat{\mathcal{L}}_N(f_{\theta_\star})$) converging to $\sigma^2_{\mathbf{e}^s}$ as $N\rightarrow\infty$ means that the average (w.r.t. $\rho^{\mathcal{D}}$) empirical error (resp. the empirical error with high probability w.r.t. $\rho^{\mathcal{D}}$) tends to the smallest possible value. If $E_{f\sim\rho^{\mathcal{D}}}\hat{\mathcal{L}}_N(f)$ (resp. $\hat{\mathcal{L}}_N(f_{\theta_\star})$) tends to $\sigma^2_{\mathbf{e}^s}$ at a rate $O(\frac{1}{\sqrt{N}})$ and $\bar{D}_2(\rho^{\mathcal{D}}\|\pi)$ remains bounded, then the upper bound converges to zero as $O(\frac{1}{N^{1/4}})$, i.e., (92)-(94) is comparable with the results of (Oymak and Ozay, 2022; Simchowitz, 2021), although it provides an upper bound with a slower convergence rate. However, in contrast to (Oymak and Ozay, 2022; Simchowitz, 2021), (92)-(94) relates the parameter estimation error with the empirical loss and the upper bound $r_N(\rho^{\mathcal{D}},\pi)$ for any $\rho^{\mathcal{D}}$. On the one hand, it is an advantage, as in contrast to (Oymak and Ozay, 2022; Simchowitz, 2021) (92)–(94) can be applied to any identification algorithm which can be represented as θ_{true} which is either the mean model w.r.t. the posterior $\rho^{\mathcal{D}}$, or it is randomly sampled from $\rho^{\mathcal{D}}$. On the other hand, in order to get $O(\frac{1}{N^{1/4}})$ error bounds, the average empirical error w.r.t. $\rho^{\mathcal{D}}$ has to be small. We conjecture that the error bound above can be improved to make it $O(\frac{1}{\sqrt{N}})$, this remains a topic of future research.

A.2. Choice of $\mu_{max}(Q_e)$

Lemma A.9. *Assume that $\mathbf{e}_g(t)$ is sub-Gaussian such that for some $\sigma>0$ it holds that for all $w\in\mathbb{R}^{n_u+n_y}$, $\mathbf{E}[e^{w^T\mathbf{e}_g(t)}]\leq e^{\frac{1}{2}w^Tw\sigma^2}$. Then with $\mu_{max}(Q_e)=(4\sigma\sqrt{n_u+n_y})^2$, there exists $\mathbf{z}(t)\sim\mathcal{N}(0,I_{n_y+n_u})$ such that*

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]\leq\mu_{\max}(Q_e)^{\frac{r}{2}}\mathbf{E}[\|\mathbf{z}(t)\|_2^r], \tag{95}$$

Proof Lemma A.9. From Lemma 1 of (Jin et al., 2019) and Lemma 5.5 of (Vershynin, 2011) it follows that

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]\leq(\sqrt{n_y+n_u})^r\sigma^r2^{r/2}r\Gamma(\frac{r}{2})$$

where Γ is the gamma function. Let $\mathbf{z}(t)$ be any random variable such that $\mathbf{z}(t)\sim\mathcal{N}(0,I_{n_y+n_u})$. Then with $m=n_y+n_u$,

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^r]=2^{r/2}\frac{\Gamma(\frac{m+r}{2})}{\Gamma(\frac{m}{2})}.$$

Hence,

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^r]\leq4^r(\sqrt{n_y+n_u})^r\sigma^r\mathbf{E}[\|\mathbf{z}(t)\|_2^r]\left(\frac{\Gamma(\frac{r}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{m+r}{2})}\frac{r}{4^r}\right)$$

We argue that

$$\frac{\Gamma(\frac{r}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{m+r}{2})}\leq\pi \tag{96}$$

and hence

$$\left(\frac{\Gamma(\frac{r}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{m+r}{2})}\frac{r}{4^r}\right)\leq\frac{r\pi}{4^r}\leq1$$

In order to show (96), we can use induction on m. For $m=1$ and $m=2$ (96) follows from the definition of the gamma function. Indeed, for $m=2$, $\frac{\Gamma(\frac{r}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{m+r}{2})}=\frac{\Gamma(\frac{r}{2})\Gamma(1)}{\Gamma(\frac{r}{2}+1)}=\frac{\Gamma(\frac{r}{2})}{\frac{r}{2}\Gamma(\frac{r}{2})}=\frac{2}{r}\leq2\leq\pi$. If $m=1$, then for $r=1$, $\frac{\Gamma(\frac{r}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{m+r}{2})}=\frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)}=\pi$ and for $r=2$, $\frac{\Gamma(\frac{r}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{m+r}{2})}=\frac{\Gamma(1)\Gamma(1/2)}{\Gamma(3/2)}=\frac{\sqrt{\pi}}{\sqrt{\pi}/2}=2\leq\pi$. Since for any $r>2$ and for $m=1$, $\frac{\Gamma(\frac{r}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{m+r}{2})}=\frac{\Gamma(\frac{r-2}{2}+1)\Gamma(\frac{1}{2})}{\Gamma(\frac{1+r-2}{2}+1)}=\frac{r-2}{r-2+1}\frac{\Gamma(\frac{r-2}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{r-2+1}{2})}\leq\frac{\Gamma(\frac{r-2}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{r-2+1}{2})}$, then by induction on r, (96) follows for $m=1$.

If $m>2$, then $\Gamma(\frac{m}{2})=\frac{m-2}{2}\Gamma(\frac{m-2}{2})$ and $\Gamma(\frac{m+r}{2})=\frac{m+r-2}{2}\Gamma(\frac{m-2+r}{2})$ and hence the statement follows from the induction hypothesis. □

A.3. Supporting Lemmas

Lemma A.10. *(Eringis et al. (2023a, Lemma A.2)]) If $\mathbf{z}(t) \sim \mathcal{N}(0, I_m)$, then for all $r \in \mathbb{N}$, $r > 0$ the following holds*

$$\mathbf{E}[\|\mathbf{z}(t)\|_2^r] \leq 2\sqrt{(m+r-1)!}$$

Lemma A.11. *(Eringis et al. (2023b, Lemma A.3)) For random variable $\mathbf{z} \sim \mathcal{N}(0, I_m)$, the even moments of $\|\mathbf{z}\|_2$ are bounded as follows: for any $r \in \mathbb{N}$, $r > 1$,*

$$\mathbf{E}[\|\mathbf{z}\|_2^{2r}] \leq 2^r (m+r-1)!$$

Combining (95) and Lemma A.10, we obtain the following lemma.

Lemma A.12. *For any $r \in \mathbb{N}$, the following holds:*

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^{2r}] \leq \mu_{\max}(Q_e)^r 2^r (m+r-1)!$$

Combining (95) and Lemma A.11, we obtain the following lemma.

Lemma A.13. *For any $r \in \mathbb{N}$, $r \geq 1$, r odd, the following holds:*

$$\mathbf{E}[\|\mathbf{e}_g(t)\|_2^{r}] \leq 2\mu_{\max}(Q_e)^{\frac{r}{2}} \sqrt{(m+r-1)!}$$

Lemma A.14. *(Eringis et al. (2023b, Lemma A.6)) Let $\mathbf{z}(t)$ be any stationary process, and $r \in \mathbb{N}$, then for a stochastic process $\mathbf{s}(t) = \sum_{k=0}^{\infty} \alpha_k \mathbf{z}(t-k)$, with $\sum_{k=0}^{\infty} \|\alpha_k\|_2 \leq +\infty$, the following holds*

$$\mathbf{E}[\|\mathbf{s}(t)\|^r] \leq \left(\sum_{k=0}^{\infty} \|\alpha_k\|_2\right)^r \mathbf{E}[\|\mathbf{z}(t)\|_2^r] \tag{97}$$

The following results from (Eringis et al., 2023b) are consequences of Lemma A.14.

Lemma A.15. *(Eringis et al. (2023b, Lemma A.7)) Let $r \in \mathbb{N}$, then with notation of the proof of Lemma A.2, for all $f \in \mathcal{F}$, the following holds*

$$\mathbf{E}[\|\mathbf{z}_\infty(t) - \mathbf{z}_{fin}(t)\|_2^r] \leq \hat{\gamma}^{rt} g_p^r(f) \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^r\right] \tag{98}$$

Lemma A.16. *(Eringis et al. (2023b, Lemma A.8)) Let $r \in \mathbb{N}$, then with notation of the proof of Lemma A.2, for all $f \in \mathcal{F}$ the following holds*

$$\mathbf{E}\left[\|\mathbf{z}_\infty(t)\|_2^r\right] \leq \left(1 + \|\hat{D}\|_2 + g_p(f)\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^r\right] \tag{99}$$

Lemma A.17. *(Eringis et al. (2023b, Lemma A.9)) Let $r \in \mathbb{N}$, then with notation of the proof of Lemma A.2, for all $f \in \mathcal{F}$ the following holds*

$$\mathbf{E}\left[\|\mathbf{z}_{fin}(t)\|_2^r\right] \leq \left(\|I\| + \|\hat{D}\|_2 + g_p(f)\right)^r \mathbf{E}\left[\left\|\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{u}(t)\end{bmatrix}\right\|_2^r\right] \tag{100}$$

B. Rényi divergence vs. KL-divergence: differences, technical difficulties

Bounds with Rényi divergence tend to be more conservative than those with KL-divergence, but they are easier to derive. Roughly speaking, bounds based on Rényi divergence require only an estimate on the second order central moment of the empirical loss. In contrast, bounds based on KL-divergence require bounding all the high-order moments of the empirical loss. The drawback is that Rényi divergence may lead to more conservative error bounds (Bégin et al., 2016), and that it is difficult to compute $\hat{\rho}$ which minimizes the right-hand of (8), see (Alquier and Guedj, 2018).

Unfortunately, the techniques used in the literature for KL-based bounds do not seem to be applicable to LTI systems with unbounded signals. More precisely, the existing literature tends to exploit a subset of the following properties when deriving bounds based on KL-divergence: the empirical loss is sub-Gaussian or sub-exponential (even bounded) (Alquier, 2021), the data is either assumed to be i.i.d. (Alquier et al., 2013) or a martingale difference sequence (Seldin et al., 2012; Haddouche and Guedj, 2022b;a), or the data is assumed to be weakly dependent and the loss is Lipschitz (Alquier et al., 2013; Alquier and Wintenberger, 2012). However, for LTI systems with unbounded signals and square loss function none of the above techniques seem directly applicable, since:

1. it is not obvious that LTI systems with sub-Gaussian noises lead to sub-Gaussian signals, since the signals are generated by LTI systems started in infinite past, that is, the signals are infinite sums of sub-Gaussian noises,
2. sub-exponential losses, without additional assumptions, do not automatically lead to useful KL bounds which converge to zero as the number of data points grows, (Germain et al., 2016; Shalaeva et al., 2020)

More precisely, we only assume that the data generator, which is an LTI system started at infinitely distant past, is driven by a sub-Gaussian noise. Then the data is an infinite sum of sub-Gaussian random variables. While a finite sum of sub-Gaussian variables is known to be sub-Gaussian, the same is not obvious for infinite sums. It is a non-trivial result from standard LTI systems theory that the outputs of LTI systems driven by Gaussian noise (input) is also Gaussian. This being said, Lemma A1 of the appendix states a bound on the moments of the data, suggesting that sub-Gaussianity could perhaps be shown. Second, even if the data is sub-Gaussian, it is not immediately obvious that the empirical loss function $V_N(f)$, which is is used to derive the bound, is sub-exponential. The loss $V_N(f)$ depends on the prediction $\hat{\mathbf{y}}_f(t)$ generated by the LTI model f using infinite amount of past data. Hence, $\hat{\mathbf{y}}_f(t)$ is an infinite sum of past data (inputs and outputs). Even if the data is sub-Gaussian, it is not obvious that $\hat{\mathbf{y}}_f(t)$ is also sub-Gaussian. Again, if the data is Gaussian, it follows from standard LTI theory than $\hat{\mathbf{y}}_f(t)$ is Gaussian. Lemma A4 provides bounds on the moments of $V_N(f)$, which suggests that it might be possible to show that $V_N(f)$ is sub-exponential.

However, even if the loss functions are sub-exponential, without further assumptions, the resulting PAC bounds will not converge to zero (for fixed posteriors) as $N \to \infty$ (Germain et al., 2016; Shalaeva et al., 2020; Eringis et al., 2021), rendering them of limited use.

To sum up, the usual techinques used in the literature to control moment generating functions of the loss do not seem to be directly applicable for LTI systems considered in this paper. At the same time, the assumptions this paper, i.e. unbounded signals and square loss, is standard in econometrics and system identification. In fact, the data is often assumed to be Gaussian (hence unbounded), and the square loss is the traditional choice of the loss function.

C. Numerical examples

C.1. Further details on the example from Section 5 and investigation of the behavior of Rényi divergence

In the following section we showcase the behaviour of the proposed bound with respect to the exponential of Rényi divergence. That is, we shall repeat the illustrative example depicted in Section 5, with the following modification. We will explore different posterior distribution schemes. We shall keep using the Gibbs posterior, and try to see how the growth rate of λ_N in $\rho_N(\theta) \propto e^{-\lambda_N \hat{\mathcal{L}}_N(f_\theta)}\pi(\theta)$. In figure 2, we see the example repeated for:

- $\lambda_N = 1$, i.e. posterior only depends on number of datapoints N through the empirical loss. This case acts as a baseline for other scenarios.
- $\lambda_N = \sqrt{N}$, commonly used rate for λ.
- $\lambda_N = \ln N + 1$, to act as an intermediary between the other two cases.

Firstly, in Figure 2 we see that for all considered rates, the numerical example seem to indicate that the bound will eventually converge to 0.

Secondly, for $\lambda_N = \sqrt{N}$, the plot indicates that the divergence $\bar{D}_2(\rho\|\pi)$ converges. Currently we are unsure if it is because we restrict predictors to only stable systems, thus imposing some constraints on the parameters. Or, if it is due to numerical issues, since we employ Markov-Chain Monte-Carlo methods to compute these quantities, and due to the behaviour of the sampler, it results with the bounded divergence. The code for the numerical example can be found in `https://gitlab.com/mpetrec/lti-pac-renyi` (files main.m and mainPlots.m).

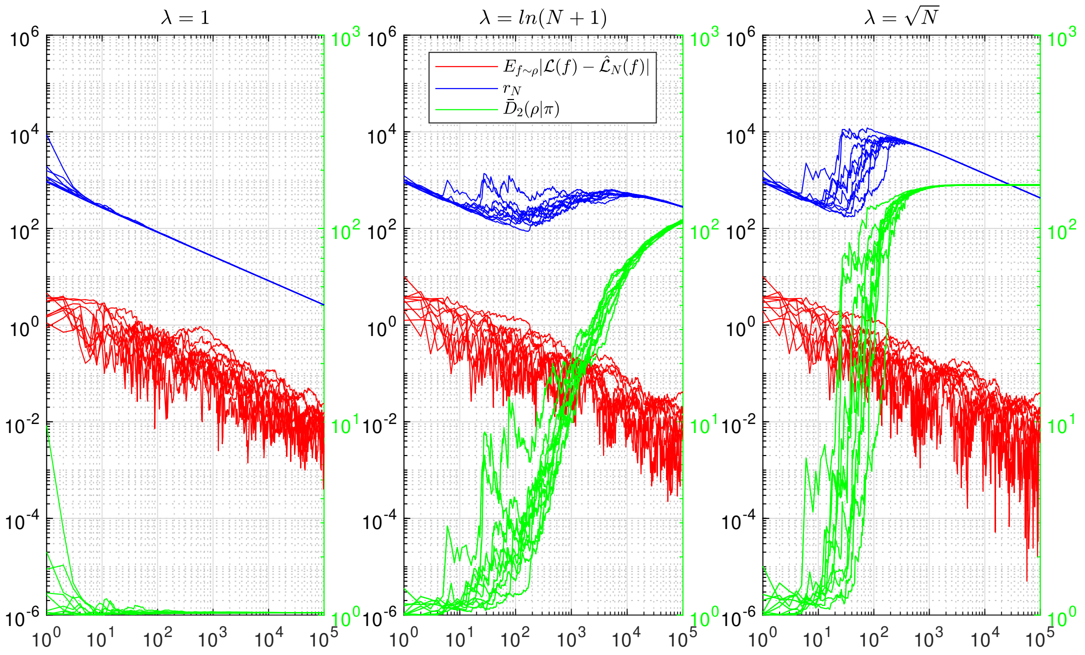

Figure 2. Repetition of the example described in Section 5, with different rates for λ_N. Repeated for 10 realisations of data. In green we show the divergence $\bar{D}_2(\rho\|\pi)$. $E_{f\sim\rho}|\mathcal{L}(f) - \hat{\mathcal{L}}_N(f)|$ and r_N are plotted on the left y-axis, and $\bar{D}_2(\rho\|\pi)$ is plotted on the right y-axis.

C.2. Example with real-life data

We tested the derived bound on the example of hair dryer benchmark from Ljung (1999, Section 17.3). This benchmark contains one input $\mathbf{u}$ and output $\mathbf{y}$. After detrending the data we applied subspace identification (n4sid command of Matlab) to estimate a data generator of the form (14) order 7. We then used a parametrization $\Sigma(\theta) = (A(\theta), B(\theta), C(\theta), D(\theta))$, where $A(\theta)(\theta), B(\theta), C(\theta), D(\theta))$ range through set of all possible matrices such that $A(\theta)$ is Schur, i.e., θ belongs to $\mathbb{R}^{16+4+4+1}$. Moreover, we took $\mathbf{w} = \begin{bmatrix}\mathbf{y}^T, & \mathbf{u}^T\end{bmatrix}^T$, i.e. we used past and current inputs and pas outputs for prediction. We chose a Gaussian prior ϕ centered around θ_0 with variance 0.2. The matrices $A_0 = A(\theta_0), B_0 = B(\theta_0)\ C_0 = C(\theta_0), D_0 = C(\theta_0)$ were chosen as follows:

$$A_0 = \begin{bmatrix} -0.0373 & 0.0672 & 0.0181 & -0.0012 \\ -0.0737 & 0.0876 & -0.0151 & -0.0018 \\ 0.0317 & 0.0141 & 0.0010 & 0.0287 \\ 0.0158 & 0.0044 & 0.0138 & -0.0109 \end{bmatrix}$$

$$B_0 = 0.001 \cdot \begin{bmatrix} 0.4468 & 0.0340 \\ 0.2581 & 0.1310 \\ -0.1509 & -0.1739 \\ -0.0792 & -0.1922 \end{bmatrix}$$

$$C_0 = \begin{bmatrix} 0.2608 & -0.1033 & -0.0301 & 0.0032 \end{bmatrix}$$

$$D_0 = \begin{bmatrix} 0 & 5 \end{bmatrix}$$

We set the posterior to the Gibbs posterior with $\lambda = 1$. That is, the posterior ρ_N is data dependent and it changes with the number of data points. We chose the confidence parameter δ as 0.1, leading to bounds which hold at least with probability 0.8. First we used only the real-life data of the benchmark to compute the posterior and to evaluate the average (w.r.t. posterior) generalization gap and the constants G_1 and G_2 and to evaluate the bound $r_N = r_N(\rho_N, \pi)$. The benchmark data set contained $N = 10^3$ data points. The constant G_2 is of $O(10^4)$ in this example. This, together with the small values of N lead to a fairly conservative overall bound, depicted in Figure 3. We conjecture that a more careful choice of the prior and of the parametrization could improve the obtained bound, but this remains a topic of future research. The code is in main_realDryerReal.m (for the computations) and makePlotsDryerReal.m (for plotting) in `https://gitlab.com/mpetrec/lti-pac-renyi`.

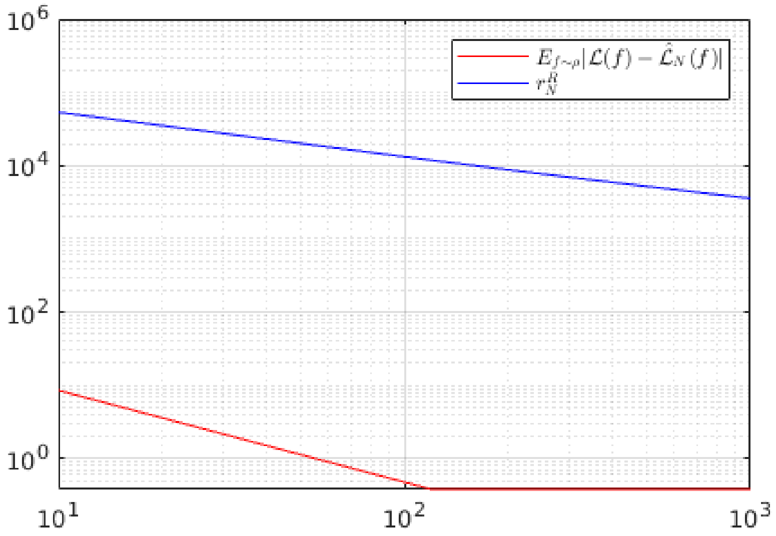

Figure 3. Example of the hair dryer, real data $N = 10^3$

We then used the data generator estimated from real-life data to generate $N = 10^6$ synthetic data points. Note that the estimated data generator produced a fit ratio of over 90% on the benchmark data. We then repeated the steps above for synthetic data, the generalisation gap and the bound r_N are depicted in Figure 4. The code is in main_realDryerS.m (for the computations) and makePlotsDryerSim.m (for plotting) in `https://gitlab.com/mpetrec/lti-pac-renyi`. We can see that the bound, while still several order of magnitude above the actual generalisation gap, converges to the actual generalisation gap.

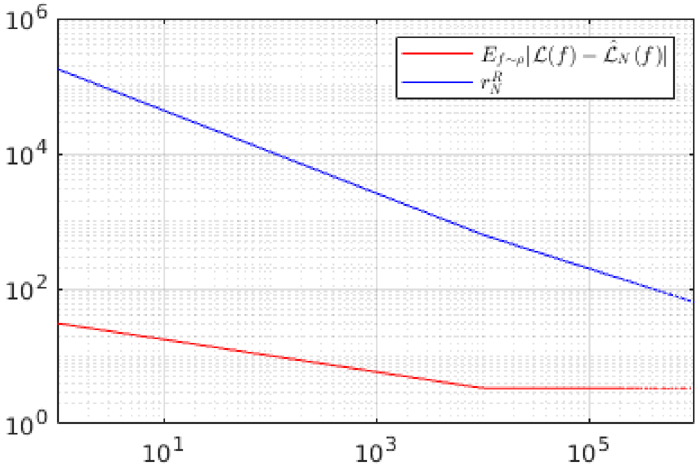

Figure 4. Example of the hair dryer, simulated data $N = 10^6$

The behavior of the Rényi divergence is not substantially different from the one described for the example of Subsection C.1.