2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Rate-Optimal Non-Asymptotics for the Quadratic Prediction Error Method

Charis Stamouli, Ingvar Ziemann, George J. Pappas

***Abstract*— We study the quadratic prediction error method—i.e., nonlinear least squares—for a class of time-varying parametric predictor models satisfying a certain identifiability condition. While this method is known to asymptotically achieve the optimal rate for a wide range of problems, there have been no *non-asymptotic* results matching these optimal rates outside of a select few, typically linear, model classes. By leveraging modern tools from learning with dependent data, we provide the first rate-optimal non-asymptotic analysis of this method for our more general setting of nonlinearly parametrized model classes. Moreover, we show that our results can be applied to a particular class of *identifiable* AutoRegressive Moving Average (ARMA) models, resulting in the first optimal non-asymptotic rates for identification of ARMA models.**

I. Introduction

Identifying predictive models from data is of critical importance in a wide range of fields, from classical control theory and signal processing to modern machine learning. To this end, a significant line of work in system identification has been devoted to identifying predictor models of the form:

$$Y_t = f_t(X_t, \theta_\star) + W_t, \tag{1}$$

from sequential data $(X_0, Y_0), \dots, (X_{T-1}, Y_{T-1})$. We typically refer to the variables X_t as the inputs and the variables Y_t as the outputs, with the inputs allowed to have a causal dependence on past outputs. However, we do not restrict attention to input-output models in the sense that (1) may well be autonomous, cf. (2) below.

Assuming that the regression functions $f_t(\cdot, \cdot)$ are known, a standard approach for estimating the unknown parameter $\theta_\star$ is to minimize the quadratic criterion:

$$L_T(\theta) := \frac{1}{T} \sum_{t=0}^{T-1} (f_t(X_t, \theta) - Y_t)^2$$

over a parameter class M which is assumed to contain $\theta_\star$. This approach yields the quadratic prediction error method, also referred to as nonlinear least squares.

As a motivating example, consider the classical prediction error method for AutoRegressive Moving Average (ARMA) models of the form:

$$Y_t = \sum_{i=1}^{p} a_i^\star Y_{t-i} + \sum_{j=0}^{q} b_j^\star W_{t-j} \tag{2}$$

from system identification [1], [2]. Such models can be cast in the form (1) with parameter $\theta_\star := \left[a_1^\star, \dots, a_p^\star, b_0^\star, \dots, b_q^\star\right]^\mathsf{T}$ and inputs $X_t := \left[Y_0, \dots, Y_{t-1}\right]^\mathsf{T}$. To convert (2) to the form (1), one selects $f_t(\cdot, \cdot)$ to be the conditional expectation of Y_t given all the past data $Y_0, \dots, Y_{t-1}$. We return to this example in more detail in Section V.

While the asymptotic rates of prediction error methods are by now well understood—including optimal rates of convergence [1] as characterized by the Cramér-Rao Inequality—less is known about their non-asymptotic counterparts. Some early progress on extending these ideas to the finite-sample regime was made in [3]. However, the bounds therein are both qualitatively and quantitatively loose as compared to older asymptotic results.

A few years ago, drawing upon recent advances in high-dimensional statistics and probability [4], [5], non-asymptotic rates nearly as sharp as the older known asymptotics were derived for the particular case of fully observed ARMA models, given by $Y_t = a_1^\star Y_{t-1} + W_t$ [6], [7]. Soon thereafter, classical subspace methods from system identification, based on higher-order linear autoregressions [8]–[10], were also given a refined non-asymptotic analysis [11]. Note that in contrast to the general prediction error method, the algorithms in [6], [7], [11] are based on linear least squares. For a broader overview of recent results on non-asymptotic learning and identification of linear models, refer to [12], [13].

As for learning and identification of nonlinear models of the form (1), progress on non-asymptotic analysis has proven somewhat slower. The special case of a general-

The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA. Emails: {stamouli,ingvarz,pappasg}@seas.upenn.edu.

Charis Stamouli and George J. Pappas acknowledge support from NSF award SLES-2331880. Ingvar Ziemann is supported by a Swedish Research Council international postdoc grant.

ized linear model (i.e., a first-order linear autoregression composed with a static known nonlinearity) is analyzed in [14], [15]. At a technical level, the goal has primarily been to sidestep—as much as possible—the blocking technique [16], which has otherwise been a dominant approach to deriving non-asymptotic guarantees for learning with dependent data [17]–[21]. In brief, the blocking technique splits a dependent sample $\{Z_t\}_{t=0}^{T-1}$ into independent blocks, say $\{Z_t\}_{t=1}^{k}, \{Z_t\}_{k+1}^{2k}, \ldots$, and then proceeds to treat each block as an independent datapoint. The caveat of this technique is that it reduces the effective sample size (e.g., here by a factor of k) and thus typically does not yield optimal rates of convergence. To provide some intuition, k above can be thought of as an analogue to the inverse stability margin of a linear system, and in fact, the blocking technique cannot be applied to marginally stable linear autoregressions. By contrast, an optimal asymptotic characterization of the rate of convergence for such autoregressions has been known since 1943 [22]. Moreover, note that sidestepping this approach is precisely what allowed [6] to first derive optimal rates for linear system identification.

More recently, [23] showed how to, at least partially, avoid the blocking approach for the time-invariant version of (1)—with $f_t(\cdot,\cdot) = f(\cdot,\cdot)$ for a fixed function $f(\cdot,\cdot)$ independent of time. The result of [23] is almost sufficient to provide a rate-optimal non-asymptotic analysis of the ARMA prediction error method. However, it has two shortcomings for this purpose, one of which we have already hinted at. First, the result does not allow for time-varying regression functions $f_t(\cdot,\cdot)$, which is crucial, as the conditional expectation function of Y_t given the past data $Y_0, \ldots, Y_{t-1}$ generally varies in time. Second, the final bounds in [23] are loose by logarithmic factors in problem quantities (including dimensional factors and the time horizon T) and hence cannot match known asymptotics [1], [24] even up to constant factors. For the case of time-invariant regression functions, the authors in [25] removed these logarithmic factors via a mixed-tail generic chaining argument.

In this paper, we pursue a simpler approach and provide the first rate-optimal non-asymptotic prediction error bound for a relatively general class of *time-varying* parametric predictor models. Our model class is rich enough to allow for ARMA models of the form (2) that satisfy a certain identifiability condition. Similar to [23], our approach is based on the martingale offset complexity introduced to the statistical literature by [26]. We arrive at our result by providing a refined analysis of this complexity notion for models of the form (1). An informal version of our main result is presented next.

Informal Version of Theorem 1. *Given data from a sufficiently stable system, for a wide range of identifiable models $f_t(\cdot,\theta_\star)$, the mean-squared prediction error corresponding to any least-squares estimate $\widehat{\theta} \in \arg\min_{\theta\in\mathsf{M}} L_T(\theta)$ satisfies:*

$$\text{Mean-Squared Prediction Error}(\widehat{\theta}) \leq \frac{\text{parameter dimension} \times \text{noise}}{\text{number of samples}} + \text{higher-order terms}.$$

The above statistical rate matches known asymptotics [1], [24] up to constant factors and higher-order terms that become negligible for a large enough sample size T. The requirement that T is larger than a so-called burn-in time is necessary to establish several components of our result, such as persistence of excitation. We note that the stability properties of the model, which are measured via the stochastic dependency of the input process $\{X_t\}_{t=0}^{T-1}$, affects only the burn-in time of our result.

In the next section, we formally present our mathematical assumptions and the problem formulation. In Section III, we introduce our main result, a proof sketch of which is given in Section IV. In Section V, we apply our main result to scalar ARMA models. Full proofs of all components of the main theorem's proof can be found in [27, Appendix].

Notation. The norm $\|\cdot\|$ is the Euclidean norm whenever it is applied to vectors and the spectral norm whenever it is applied to matrices. Moreover, $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$, and $\mathbb{B}_r^d$ the Euclidean ball of radius r in $\mathbb{R}^d$. We use $\mathbb{I}_d$ to denote the identity matrix of size d and $\mathbf{tr}(A)$ to denote the trace of any square matrix $A \in \mathbb{R}^{d\times d}$. Expectation and probability with respect to all the randomness of the underlying probability space are denoted by $\mathbf{E}$ and $\mathbf{P}$, respectively. Expectation with respect to a random variable X is denoted by $\mathbf{E}_X$. Conditional expectation of a random variable X with respect to an event $\mathcal{E}$ and a σ-field $\mathcal{F}$ is denoted by $\mathbf{E}[X|\mathcal{E}]$ and $\mathbf{E}[X|\mathcal{F}]$, respectively. For any event $\mathcal{E}$, we define $\mathbb{1}_\mathcal{E}$ as the indicator function of $\mathcal{E}$, which takes value 1 when the event occurs and 0 otherwise. If $g(\cdot)$, $h(\cdot)$ are functions defined on some unbounded subset of the positive real numbers and $h(x)$ is strictly positive for all large enough values of x, we write $g = \mathcal{O}(h)$ if there exists $x_0 \in \mathbb{R}$ such that $\limsup_{x\to x_0} |g(x)/h(x)| < \infty$.

II. Problem Formulation

Consider the predictor model (1), where the input variables X_t take values in $\mathsf{X} \subset \mathbb{R}^{d_\mathsf{x}}$, whereas the output and noise variables, denoted by Y_t and W_t, respectively, take values in $\mathbb{R}$. For each t, the regression function $f_t : \mathbb{R}^{d_\mathsf{x}} \times \mathbb{R}^{d_\theta} \to \mathbb{R}$ is known and depends on the

input X_t and the unknown parameter $\theta_\star$. The parameter $\theta_\star$ is assumed to belong to some known and compact parameter class $\mathsf{M} \subseteq \mathbb{B}_{B_\theta}^{d_\theta}$, where $B_\theta > 0$.

Before formalizing our problem, we introduce a few further assumptions about model (1) and the parameter class M. We start by characterizing the stochastic dependency of $\{X_t\}_{t=0}^{T-1}$, which can be thought of as a measure of the stability of model (1). Let us first state the main definition we will need for this characterization.

Definition 1 (Dependency Matrix)**.** *[28, Section 2] Let $\{Z_t\}_{t=0}^{T-1}$ be a stochastic process with joint distribution P_Z. For each (i,j), let $\mathsf{P}_{Z_{i:j}}$ denote the joint distribution of $\{Z_t\}_{t=i}^{j}$ and $\mathcal{Z}_{ij} := \sigma(Z_i, \dots, Z_j)$ the σ-algebra generated by $\{Z_t\}_{t=i}^{j}$. The dependency matrix of $\{Z_t\}_{t=0}^{T-1}$ is the matrix $\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{Z}) := \{\Gamma_{ij}\}_{i,j=0}^{T-1} \in \mathbb{R}^{T\times T}$, where:*

$$\Gamma_{ij} = \sqrt{2 \sup_{\substack{A\in\mathcal{Z}_{0:i} \\ B\in\mathcal{Z}_{j:T-1}}} \left|\mathsf{P}_{Z_{j:T-1}}(B|A) - \mathsf{P}_{Z_{j:T-1}}(B)\right|},$$

for $i < j$, $\Gamma_{ii} = 1$, and $\Gamma_{ij} = 0$, for $i > j$.

Let P_X denote the joint distribution of the input process $\{X_t\}_{t=0}^{T-1}$. We can measure the dependency of $\{X_t\}_{t=0}^{T-1}$ via the norm $\|\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{X})\|$ of its dependency matrix. Notice that $\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{X})$ always satisfies $1 \le \|\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{X})\| \le cT$, for some $c > 0$. The lower bound of $\|\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{X})\|$ corresponds to independent input processes, whereas the upper bound corresponds to fully dependent input processes (i.e., processes with $X_t = X_{t+1}$, for all $t = 0, \dots, T-2$). Our results apply to processes for which $\|\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{X})\|^2$ grows sublinearly in T, as formalized in the following assumption.

Assumption 1. *There exist $b_1 > 0$ and $b_2 \in [0,1)$ such that $\|\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{X})\|^2 \le b_1 T^{b_2}$.*

Assumption 1 holds for a large family of input processes $\{X_t\}_{t=0}^{T-1}$ including, e.g., geometrically ϕ-mixing processes [28], processes that satisfy Doeblin's condition [28], [29], and stationary time-homogeneous Markov chains (see [23] for details). In the context of stable linear dynamical systems with bounded noise, it has been shown that the spectral norm of the dependency matrix $\Gamma_{\text{dep}}(\mathsf{P}_\mathsf{X})$ is uniformly bounded (i.e., $b_2 = 0$) [23], which implies an intuitive connection between stability and dependency in the process $\{X_t\}_{t=0}^{T-1}$.

Assumption 2. *For each t, let $\mathcal{F}_t := \sigma(X_0, \dots, X_{t+1}, W_0, \dots, W_t)$ be the σ-field generated by the inputs $X_0, \dots, X_{t+1}$ and the noise variables $W_0, \dots, W_t$. For every t, the noise variable W_t is σ_w^2-conditionally sub-Gaussian with respect to $\mathcal{F}_{t-1}$, that is:*

$$\mathbf{E}[e^{\lambda W_t}|\mathcal{F}_{t-1}] \le e^{\frac{\lambda^2\sigma_w^2}{2}},$$

for all $\lambda \in \mathbb{R}$, for some $\sigma_w > 0$.

Assumption 2 is satisfied if the noise variables W_t are i.i.d. zero-mean Gaussian with variance σ_w^2 and independent of the inputs $X_0, \dots, X_t$. In addition, it is satisfied by a large number of non-Gaussian random variables W_t [5]; it is also standard in prior work [23], [30], [31].

Assumption 3. *For each t, the regression function $f_t(\cdot,\cdot)$ is twice differentiable with respect to its second argument. Moreover, there exist $L_1, L_2 > 0$ such that the partial gradients $\nabla_\theta f_t(\cdot,\cdot)$ and the partial Hessians $\nabla_\theta{}^2 f_t(\cdot,\cdot)$ satisfy $\|\nabla_\theta f_t(x,\theta)\| \le L_1$ and $\|\nabla_\theta{}^2 f_t(x,\theta)\| \le L_2$, respectively, for all $(x,\theta) \in \mathsf{X}\times\mathsf{M}$. In addition, the partial Hessians $\nabla_\theta{}^2 f_t(\cdot,\cdot)$ are L_3-Lipschitz continuous in their second argument with respect to the norm $\|\cdot\|$.*

Note that for all functions $f_t(\cdot,\cdot)$ that are three times differentiable with respect to their second argument, Assumption 3 trivially holds if X is bounded given that $\mathsf{M} \subseteq \mathbb{B}_{B_\theta}^{d_\theta}$ is bounded. One expects that our results extend to unbounded inputs via a truncation argument, see for instance [23, Section 5.1]. We leave a thorough analysis of this case for future work.

Assumption 4 (Positive Definite Information Matrix)**.** *There exists $\lambda_0 > 0$ such that:*

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \nabla_\theta f_t(X_t,\theta_\star)\, \nabla_\theta{}^\intercal f_t(X_t,\theta_\star)\right] \succeq \lambda_0 \mathbb{I}_{d_\theta}.$$

Assumption 4 imposes a minimal noise excitation condition, quantifying the notion of persistence of excitation [1]. Put differently, it asks that the parameter $\theta_\star$ is identifiable (in the second-order sense). We note in passing that analogous conditions are employed in recent related work (see, e.g., [15], [23], [32]).

Assumption 5 (Quadratic Identifiability)**.** *There exists $a > 0$ such that for every $\theta \in \mathsf{M}$:*

$$\|\theta - \theta_\star\|^2 \le a\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(f_t(X_t,\theta) - f_t(X_t,\theta_\star))^2\right]. \tag{3}$$

Assumption 5 imposes a regularity condition on the regression functions $f_t(\cdot,\cdot)$ with respect to the parameter space. More specifically, it quantifies the growth of the prediction error as quadratic in the parameter error. We point out that condition (3) is weaker than the global positive-definiteness condition:

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \nabla_\theta f_t(X_t,\theta)\, \nabla_\theta{}^\intercal f_t(X_t,\theta)\right] \succeq \delta \mathbb{I}_{d_\theta},$$

which is often assumed in the asymptotic literature [24], for all $\theta \in \mathsf{M}$, for some $\delta > 0$. Moreover, note that Assumption 5 always holds for linear dynamical systems as well as generalized linear models satisfying a certain expansivity condition (see, e.g., [14], [15], [33]).

The goal of system identification can often be cast as to identify the parameter $\theta_\star$, specifying the data-generating distribution in (1), from sequential input-output data $(X_0, Y_0), \ldots, (X_{T-1}, Y_{T-1})$ [34]. In this paper, we analyze the finite-sample performance of the regression functions $f_t(\cdot, \widehat{\theta})$, where $\widehat{\theta}$ satisfies:

$$\widehat{\theta} \in \arg\min_{\theta \in \mathsf{M}} \frac{1}{T} \sum_{t=0}^{T-1} (f_t(X_t, \theta) - Y_t)^2. \tag{4}$$

In particular, we are interested in providing an upper bound for the (mean-squared) prediction error of the models $f_t(\cdot, \widehat{\theta})$, given by:

$$\mathbf{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} (f_t(\bar{X}_t, \widehat{\theta}) - f_t(\bar{X}_t, \theta_\star))^2\right]. \tag{5}$$

Herein, we use $\{\bar{X}_t\}_{t=0}^{T-1}$ to denote a fresh sample drawn from P_X independently of $\{X_t\}_{t=0}^{T-1}$. We formalize the problem in the following statement.

Problem 1 (Rate-Optimal Non-asymptotic Analysis of the Quadratic Prediction Error Method)**.** *Assume that $\theta_\star$ in predictor model* (1) *is unknown. Consider a finite number $T \in \mathbb{N}_+$ of sequential input-output data $(X_0, Y_0), \ldots, (X_{T-1}, Y_{T-1})$ generated by model* (1) *and let $\widehat{\theta}$ satisfy* (4)*. Provide bounds T_0 and $\varepsilon(T)$ such that if $T \geq T_0$, then:*

$$\mathbf{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} (f_t(\bar{X}_t, \widehat{\theta}) - f_t(\bar{X}_t, \theta_\star))^2\right] \leq \varepsilon(T).$$

The bounds T_0 and $\varepsilon(T)$ may also depend on the parameters d_θ, σ_w, B_θ, L_1, L_2, L_3, λ_0, a, b_1, and b_2. Moreover, the prediction error bound $\varepsilon(T)$ should match known asymptotics up to constant factors in its leading term (see Remark 1 for details).

Remark 1. *We refer to non-asymptotic rates for the prediction error* (5) *as optimal if they match known asymptotics up to constant factors and higher-order terms. In particular, existing results for the quadratic prediction error method from the asymptotic literature [1], [24] guarantee that $\sqrt{T}(\widehat{\theta} - \theta_\star)$ converges in distribution to $\mathcal{N}(0, \mathcal{I}^{-1}(\theta_\star))$, where:*

$$\mathcal{I}(\theta_\star) := \frac{1}{\sigma_w^2} \mathbf{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} \nabla_\theta f_t(X_t, \theta_\star) \nabla_\theta^\intercal f_t(X_t, \theta_\star)\right]$$

is the Fisher information matrix. An informal calculation—ignoring the higher-order terms in Taylor's theorem—suggests that the prediction error can be written as follows:

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} (f_t(\bar{X}_t, \widehat{\theta}) - f_t(\bar{X}_t, \theta_\star))^2\right] \\
&\approx \mathbf{E}\left[(\widehat{\theta} - \theta_\star)^\intercal (\sigma_w^2 \mathcal{I}(\theta_\star))(\widehat{\theta} - \theta_\star)\right] \\
&= \mathbf{E}\,\mathbf{tr}\left(\sigma_w^2 \mathcal{I}(\theta_\star)(\widehat{\theta} - \theta_\star)(\widehat{\theta} - \theta_\star)^\intercal\right).
\end{aligned} \tag{6}$$

Under suitable regularity conditions, we can deduce that the expectation of the trace on the right-hand side of (6) *asymptotically converges to $\frac{\sigma_w^2 d_\theta}{T}$, that is:*

$$\begin{aligned}
&T^{-1} \mathbf{E}\,\mathbf{tr}\left(\sigma_w^2 \mathcal{I}(\theta_\star)(\widehat{\theta} - \theta_\star)(\widehat{\theta} - \theta_\star)^\intercal\right) \\
&\to \mathbf{tr}\left(\sigma_w^2 \mathcal{I}(\theta_\star) \mathcal{I}^{-1}(\theta_\star)\right) = \sigma_w^2 d_\theta.
\end{aligned}$$

In light of the above result, our goal is to obtain a rate of convergence that decays as fast as $\frac{c\sigma_w^2 d_\theta}{T}$, for some universal constant $c > 0$.

III. Optimal Non-asymptotic Rates for the Quadratic Prediction Error Method

In this section, we present our main result, which is a rate-optimal bound for the prediction error (5) of the models $f_t(\cdot, \widehat{\theta})$, where $\widehat{\theta}$ is an estimate of the true parameter $\theta_\star$, satisfying (4). Before we state our main theorem, let us note that herein, poly_ψ denotes a polynomial of degree of order ψ in its arguments.

Theorem 1 (Optimal Non-asymptotic Rates for the Quadratic Prediction Error Method)**.** *Consider the predictor model* (1) *and the parameter class* M *under Assumptions 1 to 5. Fix any $\gamma \in (0, 1/2)$ and let $\widehat{\theta}$ satisfy* (4)*. Then, there exist:*

$$T_1 := \mathsf{poly}_{\frac{1}{1-b_2}}(d_\theta, L_1, a, b_1, 1/(1-b_2)), \tag{7a}$$

$$T_2 := \mathsf{poly}_{\frac{1}{1-b_2}}(d_\theta, \sigma_w, \mathrm{B}_\theta, L_1, 1/\lambda_0, b_1, 1/(1-b_2)), \tag{7b}$$

$$T_3 := \mathsf{poly}_{\frac{1}{1-2\gamma}}(d_\theta, \sigma_w, \mathrm{B}_\theta, L_1, L_2, L_3, a, 1/(1-2\gamma)), \tag{7c}$$

and a universal constant $c > 0$ such that if $T \geq \max\{T_1, T_2, T_3\}$, we have:

$$\mathbf{E}\Big[\frac{1}{T} \sum_{t=0}^{T-1} (f_t(\bar{X}_t, \widehat{\theta}) - f_t(\bar{X}_t, \theta_\star))^2\Big] \leq \frac{c d_\theta \sigma_w^2}{T} + \frac{B}{T^{1+\gamma}}, \tag{8}$$

where $B = 2L_1^2\,\mathrm{B}_\theta{}^2 + 16$.

The exact expressions of the polynomials T_1, T_2, and T_3 of Theorem 1 are given in [27, Appendix].

Remark 2 (Result interpretation)**.** *Observe in* (8) *that for sufficiently large sample size T, the least-squares prediction error decays at a rate of $\mathcal{O}(T^{-1})$. In particular, the leading term in* (8) *is determined by the signal-to-noise ratio (SNR) of model* (1)*, which is defined as $\mathsf{SNR} = \sigma_w^2/T$. Notice that the longer the predictor model is excited by noise and the smaller the sub-Gaussian parameter σ_w is, the smaller the prediction error bound becomes. We note that this rate is optimal in the sense that it matches known asymptotics up to constant factors in its leading term (see Remark 1), after a finite burn-in time $T_0 := \max\{T_1, T_2, T_3\}$. The burn-in time grows polynomially in: i) the parameter dimension d_θ, ii) the sub-Gaussian parameter σ_w, iii) the noise bound B_θ, iv) the dependency parameter b_1, v) the smoothness parameters L_1, L_2, L_3, a, and vi) the inverse of the noise excitation constant λ_0. Notice also the exponential growth of T_0 in the dependency parameter b_2. The parameter b_2 is typically zero for exponentially stable dynamical systems (consider, e.g., exponentially stable ARMA models, cf. [23] for the case of autoregressive models). Nonetheless, improving this growth rate is an interesting future research direction.*

In the following section, we sketch the proof steps of Theorem 1.

IV. Proof Sketch of Theorem 1

In this section, we present the main proof steps of Theorem 1, which provides us with optimal non-asymptotic rates for the quadratic prediction error method.

A key quantity appearing in our analysis is the martingale offset corresponding to a parameter $\theta \in \mathsf{M}$, which can be thought of as a measure of the complexity of the corresponding regression functions $f_t(\cdot, \theta)$. To formally define the martingale offset, let us first introduce relevant notation. Let $\{W_t\}_{t=0}^{T-1}$ denote the noise sequence corresponding to the input-output data $(X_0, Y_0), \ldots, (X_{T-1}, Y_{T-1})$, i.e., let $W_t = Y_t - f_t(X_t, \theta_\star)$, for all $t = 0, \ldots, T-1$. Moreover, consider the shifted process $\{g_t(X_t, \theta)\}_{t=0}^{T-1}$, where $g_t(X_t, \theta) = f_t(X_t, \theta) - f_t(X_t, \theta_\star)$, for all $t = 0, \ldots, T-1$. For any $\theta \in \mathsf{M}$, the *martingale offset* corresponding to the parameter θ is defined as:

$$M_T(\theta) = \frac{1}{T} \sum_{t=0}^{T-1} \left(4 W_t g_t(X_t, \theta) - g_t^2(X_t, \theta)\right).$$

The above definition is motivated by the *martingale offset complexity* $\sup_{\theta \in \mathsf{M}} M_T(\theta)$, which is employed in previous works [23], [30], [35]. As we will see in the analysis that follows, deriving a bound on the expected martingale offset $\mathbf{E} M_T(\widehat{\theta})$ of any least-squares estimate $\widehat{\theta}$, instead of the expected martingale offset complexity $\mathbf{E}[\sup_{\theta \in \mathsf{M}} M_T(\theta)]$, is essential for obtaining optimal finite-sample rates for the quadratic prediction error method.

In the theorem below, we present a bound for the prediction error of the models $f_0(\cdot, \widehat{\theta}), \ldots, f_{T-1}(\cdot, \widehat{\theta})$, conditioned on the given sample $\{(X_t, Y_t)\}_{t=0}^{T-1}$. Note that the following theorem is a modified version of [23, Corollary 4.2] for time-varying predictor models. We achieve the extension to the time-varying case by deriving concentration inequalities for the sum of time-varying functions of the input data, leveraging a result from [28] (see [27, Appendix B] for details).

Theorem 2. *Consider the predictor model* (1) *and the parameter class* M *under Assumptions 1, 3 and 5. Fix any $\gamma \in [0, 1)$ and let $\widehat{\theta}$ satisfy* (4)*. Then, there exists T_1, defined as in* (7a)*, such that if $T \geq T_1$, we have:*

$$\begin{aligned} &\mathbf{E}_{\bar{X}_{0:T-1}} \left[\frac{1}{T} \sum_{t=0}^{T-1} (f_t(\bar{X}_t, \widehat{\theta}) - f_t(\bar{X}_t, \theta_\star))^2 \right] \\ &\leq 8 M_T(\widehat{\theta}) + \frac{2 L_1^2 B_\theta^2}{T^{1+\gamma}}, \end{aligned} \tag{9}$$

where $\bar{X}_{0:T-1} = (\bar{X}_0, \ldots, \bar{X}_{T-1})$.

Given (4), by taking the expectation over the sample $\{(X_t, Y_t)\}_{t=0}^{T-1}$, (9) yields:

$$\begin{aligned} &\mathbf{E} \left[\frac{1}{T} \sum_{t=0}^{T-1} (f_t(\bar{X}_t, \widehat{\theta}) - f_t(\bar{X}_t, \theta_\star))^2 \right] \\ &\leq 8 \mathbf{E} M_T(\widehat{\theta}) + \frac{2 L_1^2 B_\theta^2}{T^{1+\gamma}}, \end{aligned} \tag{10}$$

for any $\gamma \in [0, 1)$. Moreover, Assumption 5 can be invoked to obtain the parameter error bound:

$$\|\widehat{\theta} - \theta_\star\|^2 \leq 8 a M_T(\widehat{\theta}) + \frac{2 a L_1^2 B_\theta^2}{T}, \tag{11}$$

since $T \leq T^{1+\gamma}$, for all $\gamma \in [0, 1)$.

Employing the result from [30, Lemma 10] to bound the expected martingale offset complexity $\mathbf{E}[\sup_{\theta \in \mathsf{M}} M_T(\theta)]$, inequality (10) directly provides us with a prediction error bound of order $\mathcal{O}(\log T/T)$, after a finite burn-in time T_1. We note that this is the first non-asymptotic result for predictor models of the form (1), where the regression functions $f_t(\cdot, \cdot)$ are time-varying.

In the rest of this section, our goal is to improve upon that rate and ensure an optimal convergence rate of $\mathcal{O}(T^{-1})$, after a longer but finite burn-in time. Recall that herein optimality of rate implies matching existing results from the asymptotic literature, modulo constant factors and higher-order terms (see Remark 1). Our

refinement of the prediction error bound resulting from Theorem 2 consists of three distinct steps:

i) First, we derive an upper bound for $\mathbf{E}M_T(\widehat{\theta})$ by using the Taylor expansion of the models $f_t(X_t, \widehat{\theta})$ around $\theta_\star$. This bound depends on a "linearized" term and higher-order terms related to the parameter error $\|\widehat{\theta} - \theta_\star\|$.

ii) Second, we provide a rate-optimal bound which scales like $d_\theta \sigma_w^2 T^{-1}$ for the "linearized" term, by leveraging ideas from linear system identification.

iii) Third, we combine our bound for the "linearized" term with faster decaying bounds for the higher-order terms to obtain a refined bound for $\mathbf{E}M_T(\widehat{\theta})$. The main idea is to bound the higher-order terms employing the nearly optimal bounds resulting from Theorem 2. Owing to the higher order of these components, a careful analysis does not degrade the leading $d_\theta \sigma_w^2 T^{-1}$-order term of the linearized component.

Putting everything together, we provide the first optimal non-asymptotic rates for the quadratic prediction error method. For clarity of presentation, we separately analyze the aforementioned proof steps.

Step I: Bounding $\mathbf{E}M_T(\widehat{\theta})$ via Taylor expansion. By Taylor's theorem with remainder, for each t, we have:

$$f_t(X_t, \widehat{\theta}) = f_t(X_t, \theta_\star) + Z_t^\intercal(\widehat{\theta} - \theta_\star) + \frac{1}{2}(\widehat{\theta} - \theta_\star)^\intercal V_t(\widehat{\theta} - \theta_\star), \tag{12}$$

where $Z_t = \nabla_\theta f_t(X_t, \theta_\star)$ and $V_t = \nabla_\theta^2 f_t(X_t, \tilde{\theta}_t)$, with $\tilde{\theta}_t = \alpha_t \widehat{\theta} + (1 - \alpha_t)\theta_\star$, for some $\alpha_t \in [0, 1]$. Exploiting the Taylor expansion of each $f_t(X_t, \widehat{\theta})$ from (12), we can prove the following lemma.

Lemma 1. *Consider the predictor model* (1) *and the parameter class* M *under Assumption 3. Moreover, let* $\widehat{\theta}$ *satisfy* (4)*, and for each* t*, consider the Taylor expansion of* $f_t(X_t, \widehat{\theta})$ *around* $\theta_\star$ *given in* (12)*. Then, the martingale offset of* $\widehat{\theta}$ *satisfies:*

$$M_T(\widehat{\theta}) \leq \bar{M}_T(\widehat{\theta}) + \left\| \frac{2}{T} \sum_{t=0}^{T-1} W_t V_t \right\| \|\widehat{\theta} - \theta_\star\|^2 + \frac{L_2^2}{4} \|\widehat{\theta} - \theta_\star\|^4, \tag{13}$$

where:

$$\bar{M}_T(\widehat{\theta}) = \frac{1}{T} \sum_{t=0}^{T-1} \left[4 W_t Z_t^\intercal (\widehat{\theta} - \theta_\star) - \frac{1}{2} (Z_t^\intercal (\widehat{\theta} - \theta_\star))^2 \right]. \tag{14}$$

By taking the expectation over the sample $\{X_t, Y_t\}_{t=0}^{T-1}$ in (13), we obtain the following bound for the expected martingale offset of $\widehat{\theta}$:

$$\mathbf{E}M_T(\widehat{\theta}) \leq \mathbf{E}\bar{M}_T(\widehat{\theta}) + \mathbf{E}\left[\left\| \frac{2}{T} \sum_{t=0}^{T-1} W_t V_t \right\| \|\widehat{\theta} - \theta_\star\|^2 \right] + \frac{L_2^2}{4} \mathbf{E}\|\widehat{\theta} - \theta_\star\|^4. \tag{15}$$

Notice that the bound for $\mathbf{E}M_T(\widehat{\theta})$ in (15) consists of the "linearized" term $\mathbf{E}\bar{M}_T(\theta)$ (note that the quadratic term on the right-hand side of (14) is negative) and two higher-order terms depending on the parameter error $\|\widehat{\theta} - \theta_\star\|$. In the next step of our proof, we focus on bounding the linearized component.

Step II: Bounding the "linearized" term $\mathbf{E}\bar{M}_T(\widehat{\theta})$. In the following theorem, we provide a bound for the "linearized" term $\mathbf{E}\bar{M}_T(\theta)$ appearing on the right-hand side of (15). Our analysis employs tools for self-normalized martingales, similar to previous works in linear system identification (see, e.g., [12]).

Theorem 3. *Consider the predictor model* (1) *and the parameter class* M *under Assumptions 1 to 4. Fix any* $\gamma \in (0, 1)$ *and let* $\widehat{\theta}$ *satisfy* (4)*. Then, there exists* T_2*, defined as in* (7b)*, and a universal constant* $c > 0$ *such that if* $T \geq T_2$*, we have:*

$$\mathbf{E}\bar{M}_T(\widehat{\theta}) \leq \frac{c d_\theta \sigma_w^2}{T} + \frac{1}{T^{1+\gamma}}. \tag{16}$$

Notice that the bound in (16) decays at the optimal rate of $\frac{\sigma_w^2 d_\theta}{T}$ (cf. Remark 1), up to a constant factor $c > 0$ and a higher-order term $1/T^{\gamma+1}$, which becomes negligible in finite time (set for instance $\gamma = 1/4$). Next, we present the final step of our proof, which combines the bound (16) with faster decaying bounds for the higher-order terms on the right-hand side of (15).

Step III: Bounding $\mathbf{E}M_T(\widehat{\theta})$ using the bound (16) **for $\mathbf{E}\bar{M}_T(\widehat{\theta})$ and the bound** (11) **for the higher-order terms in** (15)**.** In Step II we provided a bound of order $\mathcal{O}(T^{-1})$ for the "linearized" term $\mathbf{E}\bar{M}_T(\theta)$ appearing on the right-hand side of (15). To bound $\mathbf{E}M_T(\widehat{\theta})$ at a rate of $\mathcal{O}(T^{-1})$, it suffices to derive faster decaying bounds of order $\mathcal{O}(T^{1/(1+\gamma)})$ for the higher-order terms, where $\gamma \in (0, 1/2)$. Combining (16) from Theorem 3 and the parameter error bound given in (11), we obtain the following corollary.

Corollary 1. *Consider the predictor model* (1) *and the parameter class* M *under Assumptions 1 to 5. Fix any* $\gamma \in (0, 1/2)$ *and let* $\widehat{\theta}$ *satisfy* (4)*. Then, there exist* T_1*,* T_2*,* T_3*, defined as in Theorem 1, and a universal constant* $c > 0$ *such that if* $T \geq \max\{T_1, T_2, T_3\}$*, we*

have:

$$\mathbf{E}M_T(\widehat{\theta}) \leq \frac{cd_\theta\sigma_w^2}{T} + \frac{2}{T^{1+\gamma}}. \tag{17}$$

Notice that the dominant term on the right-hand side of (17) decays at a rate of $\frac{cd_\theta\sigma_W^2}{T}$, which is optimal up to a constant factor $c > 0$ (see Remark 1). We note that the above bound improves upon the rate $\mathcal{O}(\log T/T)$ that can been shown for the martingale offset complexity $\mathbf{E}[\sup_{\theta\in\mathsf{M}} M_T(\theta)]$ via maximal inequalities [30]. The key point here is that the offset process locally, once $\widehat{\theta}$ is sufficiently near $\theta_\star$, behaves like a linear offset process.

Combining (10) with (17) from Corollary 1, we complete the proof of (8) in Theorem 1. Next, we instantiate Theorem 1 to provide finite-sample guarantees for the quadratic prediction error method for AutoRegressive Moving Average (ARMA) models.

V. Case Study: The ARMA Model

In this section, we demonstrate the applicability of our rate-optimal non-asymptotic analysis of the quadratic prediction error method to scalar ARMA models. Our result relies on a standard analysis from [2], [36] that allows converting any ARMA model into a predictor model of the form (1). For completeness of presentation, we briefly review the conversion methodology and then present a rate-optimal non-asymptotic bound for a particular class of ARMA models.

Consider the scalar ARMA(p, q) model given by:

$$Y_t = \sum_{i=1}^{p} a_i^\star Y_{t-i} + \sum_{j=0}^{q} b_j^\star W_{t-j}, \tag{18}$$

where the noise variables $W_t \in \mathbb{R}$ are assumed to be independent and zero-mean, and the initial conditions are assumed to be zero, i.e., $Y_t = 0$, $W_t = 0$, for all $t < 0$. Suppose that $b_0^\star = 1$ and the parameter $\theta_\star := \left[a_1^\star, \ldots, a_p^\star, b_0^\star, \ldots, b_q^\star\right]^\mathsf{T} \in \mathbb{R}^{p+q+1}$ belongs to some known set $\mathsf{M} \subseteq \mathbb{B}_{\mathrm{B}_\theta}^{d_\theta}$, where B_θ is a positive constant. The assumption that $b_0^\star = 1$ can always be ensured by providing additional artificial noise components of zero mean and variance, and applying linear transformations to the noise variables W_t [36]. Let z^{-1} denote the backward-shift operator, defined by $z^{-1}e_t := e_{t-1}$, for any stochastic process $\{e_t\}_{-\infty}^{\infty}$. Powers of z^{-1} are defined recursively by $z^{-(i+1)}e_t := z^{-1}(z^{-i}e_t)$ so that $z^{-i}e_t = e_{t-i}$. It is straightforward to show that (18) is equivalent to $A_{\theta_\star}(z^{-1})Y_t = B_{\theta_\star}(z^{-1})W_t$, where $A_{\theta_\star}(\cdot)$ and $B_{\theta_\star}(\cdot)$ are polynomials given by:

$$A_{\theta_\star}(\lambda) = 1 - \sum_{i=1}^{p} a_i^\star \lambda^i,\ B_{\theta_\star}(\lambda) = \sum_{j=0}^{q} b_j^\star \lambda^j,$$

respectively, for all $\lambda \in \mathbb{R}$. For each t, let $\bar{\mathcal{F}}_t := \sigma(Y_0, \ldots, Y_t)$ be the σ-field generated by the outputs $Y_0, \ldots, Y_t$. It is known [2, Section 2.6] that the conditional expectation $\widehat{Y}_t := \mathbf{E}[Y_t|\bar{\mathcal{F}}_{t-1}]$ satisfies:

$$B_{\theta_\star}(z^{-1})\widehat{Y}_t = [B_{\theta_\star}(z^{-1}) - A_{\theta_\star}(z^{-1})]Y_t, \tag{19}$$

for all $t = 0, \ldots, T-1$. Hence, we can rewrite model (18) in the predictor model form (1) with regression functions:

$$f_t(X_t, \theta_\star) := \widehat{Y}_t, \tag{20}$$

where $X_t = \left[Y_0, \ldots, Y_{t-1}\right]^\mathsf{T}$. The conditional expectations $\widehat{Y}_0, \ldots, \widehat{Y}_{T-1}$ can be computed recursively from (19) with zero initial condition, i.e., $\widehat{Y}_t = 0$, for all $t < 0$. We can similarly define the regression functions $f_t(\cdot, \theta)$ corresponding to any parameter θ in the class M. In the corollary that follows, we combine Theorem 1 with the predictor model form of the ARMA model (18) derived above and provide the first rate-optimal non-asymptotic prediction error bounds for ARMA models.

Corollary 2. *Consider the predictor model form of the ARMA(p, q) model* (18)*, defined by* (1) *and* (20)*, as well as the parameter class* M*, under Assumptions 1 to 5. Fix any* $\gamma \in (0, 1/2)$ *and let* $\widehat{\theta}$ *satisfy* (4)*. Then, there exist* T_1*,* T_2*,* T_3*, defined as in Theorem 1, and a universal constant* $c > 0$ *such that if* $T \geq \max\{T_1, T_2, T_3\}$*, we have:*

$$\mathbf{E}\Big[\frac{1}{T}\sum_{t=0}^{T-1}(f_t(\bar{X}_t, \widehat{\theta}) - f_t(\bar{X}_t, \theta_\star))^2\Big] \leq \frac{cd_\theta\sigma_w^2}{T} + \frac{B}{T^{\gamma+1}},$$

where $d_\theta = p + q$ *and* $B = 2L_1^2\,\mathrm{B}_\theta{}^2 + 16$*.*

The proof of Corollary 2 follows directly from Theorem 1, given the predictor model form of model (18).

Note that the above corollary applies to a particular class of ARMA models that satisfy Assumptions 1 to 5. Assumptions 1 to 4 are relatively benign for this example, and hold as long as the noise sequence $\{W_t\}_{t=0}^{T-1}$ is bounded and the system (18) is stable. For Assumption 1, see e.g. [23] for the case of $B_{\theta_\star}(\lambda) = 1$. Assumption 2 is true by construction of the regression functions (20) corresponding to model (18) as well as the hypothesis of bounded noise. Assumption 3 can be verified via arguments entirely analogous to those in [37] as long as $\{W_t\}_{t=0}^{T-1}$ is bounded and the system (18) is stable. Sufficient conditions for guaranteeing Assumption 4, related to the roots of the polynomials $A_{\theta_\star}(\lambda)$ and $B_{\theta_\star}(\lambda)$, can be found in [38]. Assumption 5 restricts our result to a specific class of quadratically identifiable ARMA models (see (3)). As previously explained in Section II, this assumption is weaker than the corresponding assumption made in the asymptotic literature

for the quadratic prediction error method [24]. Exploring potential relaxations of the identifiability condition (3) is an interesting problem for future work.

REFERENCES

[1] L. Ljung, *System Identification: Theory for the User*. Pearson Education, 1998.

[2] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. Springer, 2009.

[3] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Transactions on Automatic Control*, vol. 47, no. 8, pp. 1329–1334, 2002.

[4] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[5] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

[6] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*. PMLR, 2018, pp. 439–473.

[7] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[8] M. Jansson and B. Wahlberg, "On consistency of subspace methods for system identification," *Automatica*, vol. 34, no. 12, pp. 1507–1519, 1998.

[9] A. Chiuso and G. Picci, "The asymptotic variance of subspace estimates," *Journal of Econometrics*, vol. 118, no. 1-2, pp. 257–291, 2004.

[10] S. J. Qin, "An overview of subspace identification," *Computers & chemical engineering*, vol. 30, no. 10-12, pp. 1502–1513, 2006.

[11] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3648–3654.

[12] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite-sample perspective," *IEEE Control Systems Magazine*, vol. 43, no. 6, pp. 67–97, 2023.

[13] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, "A tutorial on the non-asymptotic theory of system identification," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 8921–8939.

[14] Y. Sattar and S. Oymak, "Non-asymptotic and accurate learning of nonlinear dynamical systems," *Journal of Machine Learning Research*, vol. 23, no. 140, pp. 1–49, 2022.

[15] S. Kowshik, D. Nagaraj, P. Jain, and P. Netrapalli, "Near-optimal offline and streaming algorithms for learning non-linear dynamical systems," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8518–8531, 2021.

[16] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *The Annals of Probability*, pp. 94–116, 1994.

[17] M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-iid processes," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[18] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, "Ergodic mirror descent," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1549–1578, 2012.

[19] V. Kuznetsov and M. Mohri, "Generalization bounds for non-stationary mixing processes," *Machine Learning*, vol. 106, no. 1, pp. 93–117, 2017.

[20] A. Roy, K. Balasubramanian, and M. A. Erdogdu, "On empirical risk minimization with dependent and heavy-tailed data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8913–8926, 2021.

[21] A. Sancetta, "Estimation in reproducing kernel hilbert spaces with dependent data," *IEEE Transactions on Information Theory*, vol. 67, no. 3, pp. 1782–1795, 2020.

[22] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica, Journal of the Econometric Society*, pp. 173–220, 1943.

[23] I. Ziemann and S. Tu, "Learning with little mixing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4626–4637, 2022.

[24] L. Ljung and P. E. Caines, "Asymptotic normality of prediction error estimators for approximate system models," *Stochastics*, vol. 3, no. 1-4, pp. 29–46, 1980.

[25] I. Ziemann, S. Tu, G. J. Pappas, and N. Matni, "Sharp rates in dependent learning theory: Avoiding sample size deflation for the square loss," *arXiv preprint arXiv:2402.05928*, 2024.

[26] T. Liang, A. Rakhlin, and K. Sridharan, "Learning with square loss: Localization through offset rademacher complexity," in *Conference on Learning Theory*. PMLR, 2015, pp. 1260–1285.

[27] C. Stamouli, I. Ziemann, and G. J. Pappas, "Rate-optimal non-asymptotics for the quadratic prediction error method," *arXiv preprint arXiv:2404.07937*, 2024.

[28] P.-M. Samson, "Concentration of measure inequalities for markov chains and ϕ-mixing processes," *The Annals of Probability*, vol. 28, no. 1, pp. 416–461, 2000.

[29] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[30] I. M. Ziemann, H. Sandberg, and N. Matni, "Single trajectory nonparametric learning of nonlinear dynamics," in *conference on Learning Theory*. PMLR, 2022, pp. 3333–3364.

[31] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, "A tutorial on the non-asymptotic theory of system identification," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 8921–8939.

[32] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees," *Journal of Machine Learning Research*, vol. 23, no. 32, pp. 1–30, 2022.

[33] D. Foster, T. Sarkar, and A. Rakhlin, "Learning nonlinear dynamical systems from a single trajectory," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 851–861.

[34] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[35] T. Liang, A. Rakhlin, and K. Sridharan, "Learning with square loss: Localization through offset rademacher complexity," in *Conference on Learning Theory*. PMLR, 2015, pp. 1260–1285.

[36] M. Davis, *Stochastic modelling and control*. Springer Science & Business Media, 2013.

[37] P. Caines, "Prediction error identification methods for stationary stochastic processes," *IEEE Transactions on Automatic Control*, vol. 21, no. 4, pp. 500–505, 1976.

[38] A. Klein and P. Spreij, "On fisher's information matrix of an armax process and sylvester's resultant matrices," *Linear Algebra and its Applications*, vol. 237, pp. 579–590, 1996.