Data-Driven Prediction with Stochastic Data: Confidence Regions and Minimum Mean-Squared Error Estimates

Mingzhou Yin, Andrea Iannelli, and Roy S. Smith

Abstract—Recently, direct data-driven prediction has found important applications for controlling unknown systems, particularly in predictive control. Such an approach provides exact prediction using behavioral system theory when noise-free data are available. For stochastic data, although approximate predictors exist based on different statistical criteria, they fail to provide statistical guarantees of prediction accuracy. In this paper, confidence regions are provided for these stochastic predictors based on the prediction error distribution. Leveraging this, an optimal predictor which achieves minimum mean-squared prediction error is also proposed to enhance prediction accuracy. These results depend on some true model parameters, but they can also be replaced with an approximate data-driven formulation in practice. Numerical results show that the derived confidence region is valid and smaller prediction errors are observed for the proposed minimum mean-squared error estimate, even with the approximate data-driven formulation.

I. INTRODUCTION

In dynamical system analysis, one of the fundamental problems is to predict system responses from given inputs and initial conditions. Conventionally, this is done by simulating a model of the system, derived from first principles and/or experimental data. However, increasing complexity of systems poses challenges to the modeling process. Direct approaches have therefore been widely pursued to obtain reliable predictions of system responses without an explicit model [1]. In what follows, the term ‘data-driven’ refers to such direct approaches.

A seminal result, known as the Willems’ fundamental lemma [2], shows that data-driven prediction can be conducted by linearly combining historical trajectory data with persistently exciting inputs for linear systems. A more general version of the lemma was recently given in [3]. This result enables model-based control design techniques to be adopted with direct data-driven formulations. This framework is especially suitable for predictive control, where multiple data-driven algorithms have been developed, including subspace predictive control [4], data-enabled predictive control [5], and behavioral input-output parametrization [6]. Successful applications have been described in [7], [8].

Recently, the extension of the fundamental lemma to stochastic data from a system identification point of view has been drawing increasing interest [9]. Such work includes model predictive control based on the prediction error method [10], maximum likelihood signal matrix model [11], [12], and a Wasserstein distance minimization approach [7].

With stochastic data, both the historical trajectories and the prediction conditions are uncertain, which makes it difficult to obtain statistical guarantees of the predictors. This limits the application of data-driven predictors to control design, particularly when robustness requirements and safety constraints exist. As a result, to the best of our knowledge, existing work on robust data-driven control [13], [14], [15] is restricted to bounded noise models with often loose prediction error bounds.

In this paper, a statistical framework on the accuracy of the predicted response is established under the assumption of Gaussian output noise. With this framework, confidence regions are available for a general form of stochastic data-driven predictors. The confidence region depends on the extended observability matrix of the system, but it can also be approximated through a data-driven formulation of model properties without direct knowledge of model parameters. The validity of the derived confidence regions is verified by numerical examples.

In addition, this statistical framework allows computation of the mean-squared error (MSE) of the predictor. In this way, a stochastic data-driven predictor is designed to be optimal for prediction accuracy in terms of minimizing the MSE. This optimal algorithm can be obtained in practice with a data-driven model characterization. It is shown numerically that the proposed minimum MSE predictor, even with a data-driven approximation of the MSE, obtains smaller prediction errors than existing stochastic predictors.

Notation. A Gaussian distribution with mean $\mu$ and covariance $\Sigma$ is indicated by $\mathcal{N}(\mu, \Sigma)$. The expectation and the covariance of a random vector $x$ are denoted by $\mathbb{E}(x)$ and $\text{cov}(x)$ respectively. For a vector $x$ and a positive definite matrix $Q$, the weighted Euclidean norm $(x^T Q x)^{1/2}$ is denoted by $\|x\|_Q$. For a matrix $X$, the vectorization operator $\text{vec}(X)$ stacks its columns in a single vector; $X^\top$ indicates the Moore-Penrose pseudoinverse. For a sequence of matrices $X_1, \ldots, X_n$, we denote $[X_1^\top \ldots X_n^\top]^\top$ by $\text{col}(X_1, \ldots, X_n)$.

II. THE DATA-DRIVEN PREDICTION PROBLEM

A. Problem Statement

Consider a discrete-time linear time-invariant (LTI) system with output noise, given by

$$
\begin{aligned}
x_{t+1} &= Ax_t + Bu_t, \\
y_t &= Cx_t + Du_t + w_t,
\end{aligned}
$$

This work was supported by the Swiss National Science Foundation under Grant 200021_178890.

The authors are with the Automatic Control Laboratory, Swiss Federal Institute of Technology (ETH Zurich), Physikstrasse 3, 8092 Zurich, Switzerland, {myin,iannelli,smith}@control.ee.ethz.ch.
where \( x_t \in \mathbb{R}^{n_x} \), \( u_t \in \mathbb{R}^{n_u} \), \( y_t \in \mathbb{R}^{n_y} \), \( w_t \in \mathbb{R}^{n_w} \) are the states, inputs, outputs, and output noise respectively. In this paper, we assume that the system is observable with observability index (lag) \( l \).

In data-driven prediction, the model parameters \( A, B, C, D \) are unknown, but \( M \) length-\( L \) input-output trajectories

\[
z^d = \text{col}\left( u^d_i, \cdots, u^d_{i+L-1}, y^d_i, \cdots, y^d_{i+L-1} \right) \in \mathbb{R}^{L(n_u+n_y)} \tag{2}
\]

where \( i = 0, \cdots, M-1 \) have been collected. The matrix that concatenates these trajectories

\[
Z = \begin{bmatrix}
  z^d_0 \\
  \vdots \\
  z^d_{M-1}
\end{bmatrix} \in \mathbb{R}^{L(n_u+n_y) \times M} \tag{3}
\]

is termed the signal matrix [11]. Depending on the construction, we can choose either \( t_{i+1} = t_i + 1 \) for a mosaic Hankel signal matrix, or \( t_{i+1} = t_i + L \) for a Page signal matrix [16]. The trajectories can also come from independent experiments [17].

The problem is then to predict output trajectory \( y = \text{col}(y_0, \cdots, y_{L-1}) \) from any given input trajectory \( u = \text{col}(u_0, \cdots, u_{L-1}) \) using only the collected historical trajectories. To obtain a unique output trajectory, the initial condition is also fixed by measuring the immediate past input-output trajectory \( u_{n_u} = \text{col}(u_0, \cdots, u_{L-1}) \) and \( y_{n_u} = \text{col}(y_0, \cdots, y_{L-1}) \), where \( L_0 = L' \geq l \). In other words, the data-driven prediction problem aims to find an input-output mapping in the following form:

\[
y = F_Z(u; u_{n_u}, y_{n_u}). \tag{4}
\]

### B. Noise-Free Data-Driven Prediction

In the noise-free case, the following lemma based on [3] provides a condition for the existence of an exact mapping.

**Lemma 1:** If \( w_t = 0 \), the exact mapping in the form of (4) exists if \( \text{rank}(Z) = n_u L + n_y \).

**Proof:** According to Corollary 19 in [3], if \( \text{rank}(Z) = n_u L + n_y \), for all \( (u_{n_u}, y_{n_u}, y) \), there exists \( g \in \mathbb{R}^M \) such that \( \text{col}(u_{n_u}, y_{n_u}, y) = Z g \). Note that the observability index \( l \) satisfies \( n_l \geq n_y \). The dimension of \( \text{col}(u_{n_u}, y_{n_u}, y) \) then satisfies \( n_l L + n_y L_0 \geq \text{rank}(Z) \), so \( y \) can be uniquely determined by \( (u_{n_u}, y_{n_u}, y) \).

Define a partition of \( Z \) as

\[
Z = \begin{bmatrix}
  U_p & U_f & Y_p & Y_f
\end{bmatrix},
\]

where \( U_p \in \mathbb{R}^{n_u L \times M}, U_f \in \mathbb{R}^{n_y L' \times M}, Y_p \in \mathbb{R}^{n_u L_0 \times M}, Y_f \in \mathbb{R}^{n_y L' \times M}. \) Following the proof of Lemma 1, the mapping can be obtained by first solving the linear system

\[
\text{col}(U_p, U_f, Y_p) g = \text{col}(u_{n_u}, u, y_{n_u}),
\]

and then applying \( y = Y_f g \). Although any solution to (6) is applicable (Proposition 1 in [18]), the pseudo-inverse solution is the most commonly used. So the solution to the noise-free data-driven prediction problem is,

\[
F_Z(\cdot) = Y_f g_{\text{pinv}}, \quad g_{\text{pinv}} = \left[ \begin{bmatrix} U_p & U_f & Y_p \end{bmatrix} \right]^+ [u_{n_u}, u, y_{n_u}] \tag{7}
\]

### C. Data-Driven Prediction with Stochastic Data

When the output noise \( w_t \) is no longer zero but a realization of a stochastic process, Lemma 1 no longer holds and the mapping (1) can only be estimated approximately. The output noise leads to uncertainties in both the output signal matrix \( \text{col}(Y_p, Y_f) \) and the output initial condition \( y_{n_u} \). In this paper, the distribution of \( w_t \) is assumed to be zero-mean Gaussian. Then, the distributions of \( y_{n_u} \) and \( \text{col}(Y_p, Y_f) \) are also Gaussian. In what follows, the distributions are denoted by

\[
y_{n_u} \sim \mathcal{N}(y_{n_u}^0, \Sigma_{y_{n_u}}),
\]

\[
\text{vec}(Y_p) \sim \mathcal{N}(\text{vec}(Y_p^0), \Sigma_Y),
\]

where \( y_{n_u}^0, Y_p^0, \) and \( Y_f^0 \) are noise-free versions of \( y_{n_u}, Y_p, \) and \( Y_f \) respectively, and \( y_{n_u} \) is uncorrelated with \( \text{col}(Y_p, Y_f) \).

Under this assumption, for a given \( g \), the distribution of

\[
Y_p g \sim \mathcal{N}\left( \text{vec}\left( Y_p^0 g \right), \Sigma_p \right),
\]

\[
Y_f g \sim \mathcal{N}\left( \text{vec}\left( Y_f^0 g \right), \Sigma_f \right),
\]

where \( \Sigma_p = \sigma_y^2 \Sigma_{y_{n_u}} \Sigma_{y_{n_u}}^T \Sigma_{y_{n_u}} \), and \( \Sigma_f = \sigma_y^2 \Sigma_{y_{n_u}} \Sigma_{y_{n_u}}^T \Sigma_{y_{n_u}} \).

A special case of the noise model is when the noise is i.i.d. with \( w_t \sim \mathcal{N}(0, \sigma^2) \), and the signal matrix \( Z \) is constructed as a Page matrix or from independent trajectories. In this case, we have \( \Sigma_{y_{n_u}} = \sigma^2 \Sigma_{y_{n_u}} \), \( \Sigma_Y = \sigma^2 \Sigma_{y_{n_u}} \), and thus \( \Sigma_p = \sigma^2 \| g \|^2 I_{n_u L_0} \).

Different algorithms have been developed under this noise model, most of which share the following form:

\[
F_Z(\cdot) = Y_f g,
\]

\[
\begin{bmatrix}
  U_p \\
  U_f \\
  Y_p \\
  Y_f
\end{bmatrix} g = \begin{bmatrix}
  u_{n_u} \\
  u \\
  y_{n_u}
\end{bmatrix}, \tag{11b}
\]

The slack variable \( \delta \) is introduced to compensate for the error in both \( Y_p \) and \( y_{n_u} \). The algorithms then propose different strategies for balancing the magnitude of \( g \) and the slack variable \( \delta \). The algorithms are summarized as follows.

**Subspace predictor** [10], [19]: The solution of the algorithm is exactly the same as that for the noise-free case (7). However, the interpretation here is different. It corresponds to the least-squares estimate of a linear mapping:

\[
F_Z(\cdot) = F_Z \text{col}(u_{n_u}, u, y_{n_u}), \tag{12}
\]

where

\[
F_Z = \arg\min_F \| Y_f - F \text{col}(U_p, U_f, Y_p) \|^2_F. \tag{13}
\]

This coincides with finding the vector \( g \) that maximizes likelihood estimation to find the vector \( g \) that
maximizes the conditional probability of \( \text{col}(\delta, Yf, g) \) given \( g \):

\[
\min_{g, \delta} \log \det \left( \Sigma_g + \begin{bmatrix} \Sigma_{yini} & 0 \\ 0 & 0 \end{bmatrix} \right) + \delta^T \left( \Sigma_p + \Sigma_{yini} \right)^{-1} \delta,
\]

subject to (11b). When \( \Sigma_{yini} = \sigma^2 I_{n_L} \) and \( \Sigma_g = \sigma^2 \|g\|^2_2 I_{n_L} \), an approximate quadratic program of (14) has been derived as

\[
\min_{g, \delta} \|\delta\|^2_2 + n_y \left( L' \sigma^2 / \|g\|^2_2 + L \sigma^2 \right) \|g\|^2_2
\]

s.t. (11b).

Wasserstein distance minimization [7]: this algorithm finds the vector \( g \) that minimizes the Wasserstein distance between the stochastic distribution of \( y_{ini} \) and that of \( Y_p g \):

\[
\min_{g, \delta} \|\delta\|^2_2 + \tr \left( \Sigma_{yini} + \Sigma_p - 2 \left( \Sigma_{yini} \Sigma_p \right)^{1/2} \right),
\]

subject to (11b). When \( \Sigma_{yini} = \sigma^2 I_{n_L} \) and \( \Sigma_g = \sigma^2 \|g\|^2_2 I_{n_L} \), an approximate quadratic program of (16) has been derived as

\[
\min_{g, \delta} \|\delta\|^2_2 + n_y L_0 \sigma^2 \|g\|^2_2
\]

s.t. (11b).

It is noted that the algorithms (12), (15), and (17) can be expressed in the following unified form:

\[
\mathcal{F}_Z(\cdot) = Yf \arg \min_g \|\delta\|^2_2 + \lambda \|g\|^2_2
\]

s.t. (11b),

where \( \lambda \to 0 \) for (12), \( \lambda = n_y \left( L' \sigma^2 / \|g\|^2_2 + L \sigma^2 \right) \) for (15), and \( \lambda = n_y L_0 \sigma^2 \) for (17). With an abuse of notation, \( \arg \min_g \) denotes the optimal solution of \( g \) for the program depending on both \( g \) and \( \delta \). The optimization problem in (18) is a strongly convex quadratic program with only equality constraints. It admits a closed-form solution that is linear with respect to \( \text{col}(u_{ini}, u, y_{ini}) \).

### III. CONFIDENCE REGION ANALYSIS

In this section, confidence regions are established for the stochastic data-driven prediction algorithms discussed in Section II-C. The result first exploits information from the underlying state-space model. Then, a data-driven approximation of the model information is proposed.

#### A. Derivation of the Confidence Region

For any stochastic data-driven predictor in the form of (11), the output estimate (4) differs from the true output \( y_0 \) due to the following two sources of error: 1) the output part of the signal matrix \( Yf \) is noisy, 2) the predictor estimates a trajectory whose output initial condition is \( Y_p g \), which differs from the trajectory to be predicted whose output initial condition is \( y_{ini} \). By characterizing the distributions of these two sources of error for a particular estimate of \( g \) and \( \delta \), we obtain the following confidence region for stochastic data-driven prediction.

**Theorem 1:** Consider a stochastic data-driven predictor \( y = \mathcal{F}_Z(u, u_{ini}, y_{ini}) = Yf, g \) satisfying (11). The true output \( y_0 \) is in the following ellipsoid set w.p. \( p \):

\[
\mathcal{E} = \left\{ \tilde{y} \mid (y - \tilde{y} - \Gamma \delta)^\top \Sigma^{-1} (y - \tilde{y} - \Gamma \delta) \leq \mu_p \right\},
\]

where

\[
\Sigma = \left[ -\Gamma \ \Sigma_{ini} \left( \Gamma^\top \right)^\top + \Gamma \Sigma_{yini} \Gamma^\top \right],
\]

and \( \mu_p \) satisfies \( F_{\chi^2(d)}(\mu_p) = p \), where \( F_{\chi^2(d)}(\cdot) \) is the cumulative distribution function of the \( \chi^2 \)-distribution with \( d \) degrees of freedom.

**Proof:** Let the stochastic noise in \( Y_p, Yf \) and \( y_{ini} \) be \( E_p, E_f \), and \( \epsilon_{ini} \) respectively, i.e.,

\[
E_p = Y_p - y_0^p, \ \ E_f = y_f^0 - y_{f, ini}, \ \ \epsilon_{ini} = y_{ini} - y_0^{ini}.
\]

The estimation error can be decomposed as follows, according to the two aforementioned sources of error

\[
y - y_0 = E_f g + y^-, \quad (23)
\]

where \( y^- \) is the error due to the discrepancy \( (y_f^0 g - y_0^{ini}) \) in the output initial condition. The initial condition error \( y^- \) can be seen as the free response from initial condition \( u_{ini} = 0 \), \( y_0 = y_f^0 g - y_0^{ini} \). From (11b) and (22), we have

\[
\begin{align*}
Y_0 g &= y_{ini} + \delta - E_p g, \quad y_{ini} = y_0^{ini} - \epsilon_{ini}, \\
y^- &= (y_{ini} - \delta - E_p g) - (y_{ini} - \epsilon_{ini}) = \delta + \epsilon_{ini} - E_p g.
\end{align*}
\]

Let the state of the trajectory at time \(-L_0 = x^-\). Then we have

\[
y_{ini}^- = \begin{bmatrix} C \\ \vdots \\ C A^{L_0-1} \end{bmatrix} x^- = \begin{bmatrix} C A^{L_0} \\ \vdots \\ C A^{L_0-1} \end{bmatrix} x^-.
\]

Since \( L_0 \geq 1 \), col \( (C, \ldots, C A^{L_0-1}) \) has full column rank. Thus, we have \( x^- = \text{col}(C, \ldots, C A^{L_0-1}) y_{ini}^- \). This directly leads to \( y^- = \Gamma y_{ini}^- \). From (23)-(25), the estimation error is then

\[
y - y_0 = E_f g + \Gamma \delta + \epsilon_{ini} - E_p g.
\]

Recall that \( \epsilon_{ini} \sim N(0, \Sigma_{yini}) \), col \( (E_p, E_f) g \) \( g \sim N(0, \Sigma_g) \), and they are uncorrelated. The distribution of \( (y - y_0) \) given \( g \) and \( \delta \) is Gaussian with

\[
\mathbb{E}(y - y_0) = \Gamma \delta;
\]

\[
\text{cov}(y - y_0) = \text{E} \left( \begin{bmatrix} -\Gamma & \Sigma_{ini} \end{bmatrix} \begin{bmatrix} E_p \\ E_f \end{bmatrix} g + \Gamma \epsilon_{ini} \right)
\]

\[
= \begin{bmatrix} -\Gamma I_{n_L} & \Sigma_g \end{bmatrix} \begin{bmatrix} E_p \\ E_f \end{bmatrix} g + \Gamma \Sigma_{yini} \Gamma^\top = \Sigma.
\]

Therefore, \( (y - y_0 - \Gamma \delta)^\top \Sigma^{-1} (y - y_0 - \Gamma \delta) \) is subject to the \( \chi^2 \)-distribution with \( d \) degrees of freedom. This directly leads to (19).
Remark 1: Theorem 1 still holds when the system is not observable by replacing $A$, $C$, and $l$ with those for the observable part of the system.

Remark 2: The derivation is inspired by the prediction error bound presented in Section IV.C of [14]. However, the results of [14] consider a bounded non-stochastic noise model and provide a deterministic but admittedly non-tight bound on $|y - y_0|$. 

B. Data-Driven Formulation of System Parameter $\Gamma$

The confidence region given in Theorem 1 is not available in practice since $\Gamma$ is dependent on the unknown model parameters $A$ and $C$. However, this system parameter matrix can be alternatively formulated by another data-driven prediction scheme offline.

As can be seen from the proof of Theorem 1, the matrix $\Gamma$ can be considered as a linear data-driven predictor with $u = 0$ and $u_{mi} = 0$. Supposing we have a noise-free signal matrix, the following lemma gives a data-driven version of Theorem 1 without knowledge of $A$ and $C$.

Lemma 2: Let $\bar{Z} = \text{col}(\bar{U}_p, \bar{U}_f, \bar{Y}_p, \bar{Y}_f)$ be a noise-free signal matrix with rank($\bar{Z}$) = $n_L + n_x$. If $\Gamma$ is replaced by $\Gamma_Z = \bar{Y}_f P$, where $P$ is the last $n_L b$ columns of col$(\bar{U}_p, \bar{U}_f, \bar{Y}_p)$, then Theorem 1 holds.

Proof: According to Lemma 1, for any output initial condition $y_{ini}$, $y = \Gamma_Z y_{ini}$ is the unique free response with $u_{mi} = 0$. So we have $y^- = \Gamma_Z y_{ini}$. The rest of the proof of Theorem 1 remains the same.

Remark 3: In general, $\Gamma_Z \neq \Gamma$. This is because when $n_L b > n_x$, the valid $\Gamma$ in the proof of Theorem 1 is not unique. The pseudo-inverse solution (20) gives only one possibility.

In practice, the noisy signal matrix $Z$ can be used to find an approximation of the data-driven system parameter $\Gamma_Z$. Recall that the estimated mappings in the form of (18) admit linear solutions. So they can be employed to find an estimate of the linear mapping $\Gamma_Z$ by setting $u = 0$, $u_{mi} = 0$. The closed-form solution is given by

$$\hat{\Gamma}_Z = Y_f \left( F^{-1} - F^{-1} U (UF^{-1}U^T)^{-1} UF^{-1} \right) Y_p^T,$$

(29)

where $F = \lambda \mathbb{1}_M + Y_p^T Y_p$ and $U = \text{col}(U_p, U_f)$ as derived in [11]. The hyperparameter $\lambda$ can be selected as approaching 0 (subspace predictor), $n_L b \sigma^2$ (signal matrix model), or $n_L b \sigma^2_Y$ (Wasserstein distance minimization), and the corresponding $\hat{\Gamma}_Z$ estimates are denoted by $\hat{\Gamma}_{Z, \text{sub}}$, $\hat{\Gamma}_{Z, \text{MM}}$, and $\hat{\Gamma}_{Z, \text{WP}}$ respectively. Note that in this estimation, the output initial condition $y_{ini}$ is known exactly without noise. This leads to a slight change in the hyperparameter of the signal matrix model solution. When $\Gamma$ is replaced by $\hat{\Gamma}_Z$, Theorem 1 only holds approximately. The validity of the approximation will be investigated in Section V.

IV. Minimum Mean-Squared Error Algorithm

In the proof of Theorem 1, the distribution of the estimation error has been derived in order to quantify the confidence region for a given estimate of $g$ and $\delta$ with the algorithms discussed in Section II-C. In this section, this distribution is used to propose an optimal predictor in the form of (11), which directly targets maximum prediction accuracy, instead of the statistical properties as in Section II-C. This algorithm finds $g$ and $\delta$ in the mapping by minimizing the expected estimation error subject to (28), which leads to the following proposition.

Proposition 1: The minimum MSE estimate of the mapping in the form of (11) is given by

$$\mathcal{F}_Z(\cdot) = Y_f \arg\min_g \delta^T \delta + \text{tr} \left( \left[ -\Gamma \cdot I_{n_L'} \right] \Sigma_g \left[ I_{n_L'} \right] \right)$$

s.t. (11b).

(30)

Proof: From (28), we have

$$\text{MSE}(y - y_0) = \mathbb{E} (y - y_0)^T (y - y_0)$$

$$= \text{tr} \left( \text{cov}(y - y_0) + \mathbb{E} (y - y_0) \mathbb{E} (y - y_0)^T \right)$$

$$= \text{tr} \left( \Sigma + \Gamma \delta^T \Gamma \delta \right) = \text{tr} \left( \Sigma + \delta^T \Gamma^T \Gamma \delta \right).$$

(31)

where the third equality comes from (28). From the definition of $\Sigma$ in (21), it is observed that since $\Gamma \Sigma y_{ini} \Gamma^T$ does not depend on the optimization variables $g$ and $\delta$, minimizing the MSE is equivalent to the optimization problem in (30).

If we assume that $\Sigma = \sigma^2 u_{ini}^T L_M$, (30) becomes a quadratic program

$$\mathcal{F}_Z(\cdot) = Y_f \arg\min_g \| \delta \|^2 + \lambda_{\text{MSE}} \| g \|^2$$

s.t. (11b).

(32)

where $Q = \Gamma^T \Gamma$ and $\lambda_{\text{MSE}} = \sigma^2 n_L' + \sigma^2 \text{tr}(Q)$. This optimization problem is very similar to the unified form (18) for existing algorithms, except that the Euclidean norm of $\delta$ is now weighted by $Q$. The solution (32) is also linear with respect to col$(u_{mi}, u, y_{ini})$.

The implications of Proposition 1 are twofold. On the one hand, it provides the optimal solution to the data-driven prediction problem with output noise in terms of minimizing the MSE. Although the optimal solution relies on the unknown extended observability matrix to formulate $\Gamma$, it can be used with a preliminary model or a model set via minimax approaches.

On the other hand, similar to establishing the confidence region, the parameter $\Gamma$ used in the minimum MSE solution (30) can be replaced by the data-driven estimate $\hat{\Gamma}_Z$ (29) derived from the same signal matrix for an approximate solution.

V. NUMERICAL EXAMPLES

In this section, numerical tests are conducted to illustrate the validity of the derived confidence region and the effectiveness of the proposed minimum-MSE algorithm. In the examples, stochastic data with i.i.d. noise are collected from one single experiment and used in $Z$ with a Page matrix construction. Unit Gaussian input sequences are used to generate the data.
First, we consider a simple two-dimensional example for illustration purposes. The prediction problem is to find the first two points \((L' = 2)\) in the step response of the following fourth-order system

\[
G_1(z) = \frac{0.1059(0.1z^4 + z^3 + 0.5z^2)}{z^4 - 2.2z^3 + 2.42z^2 - 1.87z + 0.7225}.
\]  
(33)

The prediction conditions are \(u_{\text{ini}} = 0, y_{\text{ini}} = 0\), and \(u = [1 \ 1]^T\). The following parameters are used: \(L = 10, L_0 = 8, M = 80\), and noise level \(\sigma^2 = 0.1\). A confidence level of \(p = 0.90\) is used in the following figures.

Figure 1 compares the confidence regions obtained using model-based \(\Gamma (20)\) (CR-MB), data-driven \(\hat{\Gamma}_{\text{Sub}}\) (CR-Sub), \(\hat{\Gamma}_{\text{SMM}}\) (CR-SMM), and \(\hat{\Gamma}_{\text{WD}}\) (CR-WD). The confidence regions are tested on the minimum-MSE predictor with data-driven \(\hat{\Gamma}_{\text{SMM}}\) (MSE-SMM), with the approximate data-driven confidence regions. Two confidence levels \(p = 0.95\) and \(p = 0.99\) are selected. The noise level is selected as \(\sigma^2 = 0.1\). The rows in Table I correspond to different predictors, whereas the columns correspond to different formulations of the confidence region. It can be seen from the table that the empirical confidence levels match the targeted \(p\)-value well with the model-based \(\Gamma\) (CR-MB) for all three predictors, where Theorem 1 is satisfied exactly. With the data-driven estimates \(\hat{\Gamma}\), the confidence regions become marginally more conservative as the empirical confidence levels are slightly larger in Table I. The results of the three data-driven estimates \((\text{CR-Sub, CR-SMM, CR-WD})\) are similar, which indicates that the confidence region is not very sensitive to the choice of \(\hat{\Gamma}\) estimation method.

Table II compares the empirical MSE of the predictors in the Monte Carlo simulations to the MSE estimated by (31) with the approximate data-driven confidence regions.

<table>
<thead>
<tr>
<th>(p = 0.95)</th>
<th>CR-MB</th>
<th>CR-Sub</th>
<th>CR-SMM</th>
<th>CR-WD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub</td>
<td>97.1%</td>
<td>98.7%</td>
<td>98.4%</td>
<td>98.7%</td>
</tr>
<tr>
<td>SMM</td>
<td>96.8%</td>
<td>97.4%</td>
<td>97.3%</td>
<td>97.3%</td>
</tr>
<tr>
<td>MSE-SMM</td>
<td>95.2%</td>
<td>96.4%</td>
<td>96.2%</td>
<td>96.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(p = 0.99)</th>
<th>CR-MB</th>
<th>CR-Sub</th>
<th>CR-SMM</th>
<th>CR-WD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub</td>
<td>99.3%</td>
<td>100%</td>
<td>99.8%</td>
<td>99.9%</td>
</tr>
<tr>
<td>SMM</td>
<td>99.2%</td>
<td>99.7%</td>
<td>99.7%</td>
<td>99.7%</td>
</tr>
<tr>
<td>MSE-SMM</td>
<td>99.0%</td>
<td>99.3%</td>
<td>99.2%</td>
<td>99.3%</td>
</tr>
</tbody>
</table>

**Fig. 2.** Comparison of different stochastic data-driven predictors in terms of the confidence regions \((p = 0.90)\) with model-based \(\Gamma\) (CR-MB).

To quantitatively assess the derived confidence region and the minimum-MSE prediction algorithm, the following campaign of 1000 Monte Carlo simulations is set up. A bank of 1000 single-input, single-output systems are randomly generated by the \texttt{drsa} command in MATLAB with random numbers of states between 3 and 8. These random systems are normalized to have an \(\mathcal{H}_2\)-gain of 1. The prediction problem uses the following parameters: \(L = 20, L_0 = 8, L' = 12,\) and \(M = 320\). The input \(u\) and the initial condition \((u_{\text{ini}}, y_{\text{ini}})\) are selected randomly with a unit Gaussian distribution.

Then, the sizes of the confidence regions are analyzed for different stochastic data-driven predictors. The following predictors are compared: 1) subspace predictor \((\text{Sub})\), 2) signal matrix model \((\text{SMM})\), 3) Wasserstein distance minimization \((\text{WD})\), and 4) minimum-MSE predictor using model-based \(\Gamma (20)\) (MSE-MB), data-driven \(\hat{\Gamma}_{\text{Sub}}\) (MSE-Sub), \(\hat{\Gamma}_{\text{SMM}}\) (MSE-SMM), and \(\hat{\Gamma}_{\text{WD}}\) (MSE-WD). Figure 2 shows the confidence regions of these stochastic predictors with model-based \(\Gamma\) (CR-MB). As can be seen from the figure, the existing algorithms \((\text{Sub, SMM, and WD})\) have larger confidence regions compared to the minimum-MSE algorithms \((\text{MSE-MB and MSE-SMM})\). This illustrates the effectiveness of the proposed algorithm in improving prediction accuracy. In this example, the confidence regions of MSE-Sub and MSE-WD are very close to that of MSE-SMM, so they are omitted in Figure 2.
The empirical MSE is computed as $\text{MSE}_{\text{emp}}(y - y_0) = \frac{1}{N_s} \sum_{i=1}^{N_s} \| y_i - y_{0,i} \|_2^2$, where $y_i$ and $y_{0,i}$ are the predicted and the true responses of the $i$-th simulation respectively, and $N_s = 1000$. Two different noise levels of $\sigma^2 = 0.1$ and $\sigma^2 = 1$ are considered. Similar to the observation from Table I, the estimated MSE is shown to be more conservative compared to the empirical ones for all three predictors. It is also observed that the region CR-SMM is the less conservative among those tested here. However, the estimated MSE can correctly predict the relative error magnitudes of different predictors. This illustrates that the estimated MSE can be a good indicator of prediction accuracy. Only three representative predictors are shown in Table I and Table II for clarity. The results of the other algorithms are similar.

Finally, we compare the prediction accuracy of the predictors by the empirical MSE, under three different noise levels $\sigma^2 = 0.1$, $\sigma^2 = 0.5$, and $\sigma^2 = 1$. The results are shown in Table III. For all three noise levels, the minimum-MSE predictor with model-based $\Gamma$ (MSE-MB) achieves the minimum empirical MSE. This is expected as MSE-MB exactly optimizes for this objective as demonstrated in Proposition 1. However, the model-based $\Gamma$ is not available in practice. Among the other practical algorithms, MSE-SMM has the smallest empirical MSE, with slightly better performance than the direct signal matrix model approach (SMM). This result shows numerically that, with approximate data-driven formulations of $\Gamma_\sigma$, the proposed minimum-MSE predictor still obtains a more accurate prediction than the existing algorithms.

### VI. Conclusions

In this paper, the prediction error of data-driven predictors with stochastic data is characterized statistically. The framework provides ellipsoidal confidence regions for various predictors. It also offers an optimal predictor that minimizes the mean-squared prediction error directly. In practice, both the confidence region and the minimum-MSE predictor can be implemented with data-driven approximations that show good accuracy numerically.

Both the derived confidence region and the minimum-MSE predictor can contribute to more reliable and effective applications of stochastic data-driven predictors to predictive control design with robustness guarantees on the satisfaction of safety/critical constraints.

### REFERENCES


