Bias Removal in System Identification

Dalton Jones, MIT
daltonj@mit.edu

Munther A. Dahleh, MIT
dahleh@mit.edu

Peko Hosoi, MIT
peko@mit.edu

***Abstract*—Correlated data can cause significant bias in misspecified regression problems. This is a particular problem in identifying the parameters of linear time invariant (LTI) systems, when data collected over time can have unknown correlations resulting in unpredictably biased identification. Moreover, recent closed loop identification approaches often rely upon knowledge of the level of stability of the underlying system in order to design the regression problem. In this paper, we take a novel approach to minimizing bias present in estimates of the impulse response parameters for an unknown LTI system using correlated data. We derive an algorithm to adaptively minimize estimation bias using a subset of collected data. We additionally demonstrate several methods to entirely remove bias from an estimator using data collected in closed loop when additional system perturbation is allowed or when enough data has been collected initially. Our analysis reveals a novel method with which to combine potentially correlated and noisy data from different sources to synthesize an estimate of the underlying system. Finally we derive sufficient conditions for the asymptotic consistency of these methods.**

***Index Terms*—Closed Loop System Identification, Model Selection, LTI Systems**

I. Introduction

Control theory studies ways to best manipulate the behavior of dynamical systems. The generality of this field of study has found applications in everything from controlling self driving cars to modeling and developing optimal interventions for pandemics. In order to develop control algorithms, a good model of the system of interest must be derived. This is either done using physical laws or through collected data. The process of deriving models for dynamical systems using collected data falls in the subfield of system identification [1], [2].

Classically, it is assumed that systems could be perturbed with random iid inputs, then outputs from the system measured in order to derive an estimate of the underlying structure of the system. In this case, both asymptotic and non asymptotic guarantees can be made regarding the size of the estimation error. Problems in convergence and consistency of these methods arise however when correlations are present in the inputs, which can occur when inputs are generated through feedback. Furthermore, it may be the case that the engineer only has access to historical data and cannot actively manipulate the system of interest. If the data was generated using correlated inputs, most estimation procedures will result in a biased approximation of the system parameters. However, if one must make due with an existing data, it begs the question of how well one can do to derive an unbiased estimator. One can imagine another situation in which a system is costly to run in which case all existing data should be combined optimally with future system measurements. This paper addresses these two problems.

The task of identifying the parameters of a system using a finite data set was explored in [3]–[6]. The key similarity in all of these papers was that the inputs were assumed to be generated randomly and independently from noise and one another. This made the analysis of the least squares estimator used as the workhorse of identification much more straightforward, but it is not a luxury we allow ourselves. Indeed, while this scenario is reasonable in the case when there is a system one can randomly perturb, if in fact all that is available is a data set or multiple data sets of historical inputs and outputs to a system, it is unlikely that the inputs will be chosen randomly. This is particularly the case when on considers likely application of this work in economics, sociology, agriculture and other fields where experimentation may be difficult but there are historical timeseries data available. For this reason, we would like to gain a better understanding of closed loop system identification, in particular in the case where we only have access to a finite amount of data.

Closed loop system identification in general is a quite difficult problem and has been studied since the inception of the field of system identification. Classic results from [1] demonstrate that, if enough data is present, there is a delay in feedback, and an optimal solution to a certain statistical risk function can be found, this optimal solution must converge to the true system parameters. The difficulty becomes finding feasible ways to find this optimum, since in general we may be searching over a class of rational functions, a nonconvex problem.

The problem of finite time closed loop system identification has been explored in several works, particularly those of [7]–[11] in which the authors establish a way to estimate the parameters of an LTI system using closed loop inputs. In all of these cases, the derivation relies on the so called "innovation" form of the dynamics in order to transform the noise of the input output model to be white over time. Using this form, these papers then exploit some a priori knowledge of the system's stability, particularly some upper bound on $\rho(A)$ in order to determine the best number of parameters to estimate to obtain a certain convergence rate. This dependence on knowledge of the underlying system is dispensed with in this paper where we develop non parametric methods to derive consistent estimators of system parameters. This work also utilizes tools developed in [12], namely the theory of self normalized martingale inequalities in order to bound the growth of error in linear regression with dependent inputs and

noise.

In particular, including inputs generated in closed loop into the regression problem solving for the parameters of the system results in a bias known as omitted variable bias. Such a bias arise when non independent covariates are not accounted for in the regression problem. These problems have been studied in [13], [14] in which it is noted how insidious and hard to detect such errs can be. In the case of estimating LTI systems, such a bias is inevitable in the case that inputs are not independent and only a limited number of system parameters are estimated (usually to limit the effects of other noise terms). However, this bias can lead to inconsistent estimators and questionably accurate model approximations. The classic method of using instrumental variables [15], [16] to eliminate this bias also can't work in the case of system identification where the exact covariance structure relating inputs is unknown. Our work develops a new way to take inputs generated from some closed loop process and combine them to reduce or eliminate bias. The advantage we have in this paper is that some of the terms in the omitted variable bias are known (in the form of a matrix product consisting of past outputs) and hence this quantity can be manipulated in order to minimize or eliminate this bias term. This is not the case in the traditional setting of omitted variable bias where measurements of omitted variable may not even be available.

The rest of the paper is structured as follows. In section II, we introduce notation and discuss the models and results relevant to the paper. Section III introduces the specifics of the system identification problem and discusses the difficulties associated with omitted variable bias. Section IV presents the main results of the paper including an algorithm to empirically minimize bias in the estimation of impulse response parameters and a convergence analysis of this result and other similar methods. Finally, section V presents the results of several simulation studies that corroborate the claims made in our theoretical results and we conclude in section VI.

II. Notation and Preliminaries

In this paper, we consider systems of the form

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

where $B \in \mathbb{R}^{n\times 1}$, $C \in \mathbb{R}^{1\times n}$, and $A \in \mathbb{R}^{n\times n}$. We assume that A is Schur stable, or in other words that the spectral radius or largest eigenvalue $\rho(A) < 1$. Such a system can be described in terms of it's inputs and outputs by the formula

$$y_t = \sum_{i=0}^{t-1} G_i u_{t-i} + e_t$$

where $G_i = CA^iB$ are called the impulse response parameters or Markov parameters of the LTI system. We can generalize the above equation to the input output mapping using the Toeplitz matrix constructed from the inputs

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \end{pmatrix} = \begin{bmatrix} u_0 & 0 & \dots & 0 \\ u_1 & u_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_{t-1} & u_{t-2} & \dots & u_0 \end{bmatrix} \begin{pmatrix} G_0 \\ G_1 \\ \vdots \\ G_{t-1} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_t \end{pmatrix} \tag{2}$$

III. Motivation and Problem Formulation

The goal of system identification is to use measurements from a system to synthesize a model that represents the behavior of that system. In the case of linear systems, this problem is well studied in [1] and can be approached in many different ways. In our case, we first consider the case in which we want to estimate the impulse response parameters of an unknown stable linear time invariant system. This classic problem in system identification can be solved by first noting that the input output map can be described as in equation 2 with

$$Y = UG + E$$

where Y is a vector of outputs, E is a vector of measurement noise, G are the first T impulse response parameters of the system and U is the lower triangular Toeplitz matrix of inputs as in equation 2. In this case, we can solve for the least squares solution of G via $\hat{G} = \left(U^\top U\right)^{-1} U^\top Y$. Interestingly, while this estimator is actually unbiased, regardless of how the inputs u were generated (they could have come from some closed loop process or have some dependence on past outputs y) it is plagued by the presence of a lot of noise. Hence, the common approach to minimize the effect of the noise is simply to estimate the first r impulse response parameters. To do this, we rewrite equation 2 in the form

$$Y = U_r G_r + Z_r \delta_r + E$$

where U_r is the first r columns of U, Z_r are the remaining columns, G_r are the first r parameters of G and δ_r are the remaining parameters. In the case that the inputs u are chosen iid from some mean zero distribution, one can show that again the least squares solutions given by $\hat{G}_r = \left(U_r^\top U_r\right)^{-1} U_r^\top Y$ is still unbiased. This follows from results derived in [5].

However, if the inputs have some correlation, as often occurs in the case that the system is operating in closed loop, then the estimator becomes biased. In particular, the error is given by

$$G_r - \hat{G}_r = \left(U_r^\top U_r\right)^{-1} U_r^\top E + \left(U_r^\top U_r\right)^{-1} U_r^\top Z_r \delta_r.$$

Hence, even if the system is run as long as possible and the first term in the above equation approaches zero, the least squares solution will simply converge to a biased estimate of $\hat{G}_r$. The bias in the estimate, known as omitted variable bias, may degrade the approximation so much as to render any prediction made by the estimator almost meaningless, defeating the point of solving for $\hat{G}_r$ in the first place. If, on the other hand, there was an effective way of minimizing this bias without any knowledge of the system parameters G and particularly

δ_r, it would go a long way to making closed loop system identification more feasible. This is of particular importance because in many settings, all the data that is available is generated in some closed loop method where future inputs depend on previous outputs.

A. Error From Noise

It should be pointed out that there are two sources of error present in the above estimation set up. The omitted variable bias $\left(U_r^\top U_r\right)^{-1} U_r^\top Z_r \delta_r$ and the error from noise $\left(U_r^\top U_r\right)^{-1} U_r^\top E$. There is a slew of techniques that can be used to control the second source of error $\left(U_r^\top U_r\right)^{-1} U_r^\top E$. In particular, for the remainder of this paper, we assume that w_t is a mean zero Gaussian random variable independent of the previous trajectory of the system. Additionally, we assume that the inputs are "rich enough" in the sense that, despite their potential correlations, they satisfy the following matrix inequality:

$$cTI \preceq U_r^\top U_r.$$

Where $c > 0$ is some constant, T is the number of data points, and I is the identity matrix. In this case we can apply the results from [5], [12] in order to bound this error term. Particularly, we are able then to say that, given some finite set of data, that the least squares estimate is within some distance of the true solution with an arbitrarily high probability. We assume that the inputs satisfy these requirements for the remainder of the paper leaving us to focus on the first source of error, the omitted variable bias.

IV. Bias Reduction Methods

As discussed in the previous section, omitted variable bias can plague system identification algorithms when closed loop inputs are used. Removing or reducing such bias is imperative to obtain more accurate learned models. To see how this might be done, consider the following simple problem.

Example IV.1. *Suppose we have two inputs u_0, u_1 and two outputs y_1, y_2 then the input output relationship is given by*

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} u_0 & 0 \\ u_1 & u_0 \end{bmatrix} \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} + \begin{pmatrix} e_0 \\ e_1 \end{pmatrix}.$$

Now suppose that we are trying to estimate g_0. The least squares solution if we only use the first column of the matrix U is given by

$$\hat{g}_0 = \left(\begin{bmatrix} u_0 \\ u_1 \end{bmatrix}^\top \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}^\top \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

which has the omitted variable bias of $(u_0^2 + u_1^2)^{-1} u_0 u_1 g_1$ in addition to the error present through the noise process. Suppose that we now run the system again with the inputs $-u_0, u_1$ with corresponding outputs $\bar{y}_1, \bar{y}_2$. We then have the relationship between inputs and outputs given as

$$\begin{pmatrix} y_1 \\ y_2 \\ \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} = \begin{bmatrix} u_0 & 0 \\ u_1 & u_0 \\ -u_0 & 0 \\ u_1 & -u_0 \end{bmatrix} \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} + \begin{pmatrix} e_0 \\ e_1 \\ \bar{e}_0 \\ \bar{e}_1 \end{pmatrix}.$$

We can then perform least squares using this relationship to solve for $\hat{g}_0$, but we note now that the omitted variable bias is $(u_0^2 + u_1^2 + \bar{u}_0^2 + \bar{u}_1^2)^{-1}(u_0 u_1 g_1 - u_0 u_1 g_1) = 0$. And hence, at the expense of some added noise and an additional disturbance to the system, we now have an unbiased estimator of the parameter g_0. With this observation, we conclude that by combining data in such a way that the empirical covariances of the inputs cancel, one may reduce or eliminate the bias in estimates for impulse response parameters. Also note that in the above example, we did not need any knowledge of the parameter g_1. We will generalize this result to show that we may reduce bias without knowing the structure of δ_r.

A. Bias Reduction Algorithm

Suppose that there are infinitely many impulse response parameters G that describe our system of interest. If we fix $r << n$ to estimate the first r impulse response parameters of the system, we can rewrite the input output equation as

$$Y = U_r G_r + Z_r \delta_r + E$$

where U_r is the first r columns of U and Z_r are the remaining $n - r$ columns of U. In this case, we see that the least squares solution to estimate G_r is $(U_r^\top U_r)^{-1} U_r^\top Y$ which has error:

$$(U_r^\top U_r)^{-1} U_r^\top Z_r \delta_r + (U_r^\top U_r)^{-1} U_r^\top E.$$

Roughly speaking, in the case that U is generated in closed loop with delay, as long as $U_r^\top U_r$ grows quickly enough the second term goes to zero. If the inputs u are correlated over time however, the first term does not converge. The question then becomes, how do we deal with the omitted variable bias term $(U_r^\top U_r)^{-1} U_r^\top Z_r \delta_r$?. As we noted above, more data generated from the closed loop is not necessarily the answer since we just converge to a biased estimate of dubious quality. The goal would be to somehow minimize the bias $(U_r^\top U_r)^{-1} U_r^\top Z_r \delta_r$. One way to do this is to measure the response to the same system under inputs that are generated using different feedback laws.

In particular, suppose we collect data from the same system n_D times such that we have n_D datasets to learn from. Here we denote $u_k(t)$ to be the tth input of the kth data set with outputs similarly denoted. For each data set we have the familiar relation:

$$Y_k = U_{k,r} G_r + Z_{k,r} \delta_{k,r} + E$$

which we can use with the usual least squares algorithm. Supposing that we used all of the data, we could stack the vectors Y_k on top of one another in addition to stacking $U_{k,r}$ and $Z_{k,r}$ to get the usual least squares solution with bias error of the form $(U_r^\top U_r)^{-1} U_r^\top Z_r \delta_r$. But this can be rewritten as $(\frac{1}{n_D T} U_r^\top U_r)^{-1} \frac{1}{n_D T} \sum_k U_{k,r}^\top Z_{k_r} \delta_r$. Thus, roughly speaking,

if the feedback mechanisms are different enough, the mean empirical covariance across systems $\frac{1}{n_D}\sum_k U_{k,r}^\top Z_{k_r}$ could be small. This depends on the different feedback mechanisms present. It might also be that we don't want to use all of the data we have since it could be biased and we'd like to combine the data in a way that these biases cancel.

Thus, instead of taking all of the datasets, we might as well choose a subset of the datasets $S \subseteq [n_D]$ such that the term $||(\sum_{k\in S} U_{k,r}^\top U_{k,r})^{-1}\sum_{k\in S} U_{k,r}^\top Z_{k_r}||$ is minimized. This allows us to put together different correlated data in order to minimize the omitted variable bias. This informal discussion allows us to formulate the following algorithm.

Algorithm 1 Minimizing Omitted Variable Bias for LTI System Identification

- Select $0 < r << T$, $n_D > 0$, $T > 0$;
- Collect input output data of length T, n_D times from the system using arbitrary feedback;
- Solve for subset $S \subseteq [n_D]$ such that $||(\sum_{k\in S} U_{k,r}^\top U_{k,r})^{-1}\sum_{k\in S} U_{k,r}^\top Z_{k_r}||$ is minimized;
- Perform "least biased" regression using data from the subset S;

It is worth noting that the above minimization is a combinatorial optimization problem, and thus is unlikely to be solved efficiently, however even a greedy algorithm can decrease the bias.

Proposition IV.2. *Suppose that S is the optimal subset of data returned from algorithm 1. Then the bias term satisfies:*

$$\begin{aligned}&||(\sum_{k\in S} U_{k,r}^\top U_{k,r})^{-1}\sum_{k\in S} U_{k,r}^\top Z_{k_r}||\\ &\leq ||(\sum_{k=1}^{n_D} U_{k,r}^\top U_{k,r})^{-1}\sum_{k=1}^{n_D} U_{k,r}^\top Z_{k_r}||.\end{aligned}$$

Proof. The proof follows immediately from the structure of algorithm 1. □

This result suggests a way to reduce the bias of an estimation of $\hat{G}_r$ using only known information. Note that this result also could be used as a part of a model selection algorithm, in which, given a data set and a known noise level, one could select the subset S and the order r such that the estimation of the first r parameters from the subset of data S minimizes bias and error from noise (which can be upper bounded using the self normalized martingale theory of [12].)

B. Consistency of Bias Reduction Algorithm

One question we could ask about the above algorithm is when does it asymptotically converge to the correct values of the first r impulse parameters of the system G_r? The proposition describes sufficient conditions for this to occur.

Proposition IV.3. *Suppose that data of length T is collected from n_D runs of a system where T is fixed, $n_D \to \infty$ and we are trying to estimate the first r impulse response parameters G_r using the algorithm 1. Suppose also that $\alpha_{min} I|S| \preceq \sum_{k\in S} U_{r,k}^\top U_{r,k}$ for all k. Suppose $S(n_D)$ is the optimal subset of data given n_D runs of the system. Then we have the estimator $\hat{G}_r(S(n_D))$ is consistent if $|S(n_D)| \to \infty$ and*

$$\lim_{n_D\to\infty}\left(\sum_{k\in S(n_D)} U_{r,k}^\top U_{r,k}\right)^{-1}\left(\sum_{k\in S(n_D)} U_{r,k}^\top Z_{r,k}\right)\delta_r = 0. \tag{3}$$

Proof. Suppose that the condition in equation 3 holds, then in the limit, the only error term is given by

$$\left(\sum_{k\in S(n_D)} U_{r,k}^\top U_{r,k}\right)^{-1}\left(\sum_{k\in S(n_D)} U_{r,k}^\top E_k\right).$$

Using the fact that we can lower bound the singular values of $\sum_{k\in S(n_D)} U_{r,k}^\top U_{r,k}$, we see that the error is mean zero and has variance at most $\frac{1}{|S(n_D)|}\frac{1}{\alpha_{min}^2}\sigma_e^2$. Using Chebyshev's inequality shows that the estimator must thus converge in probability and we obtain a consistent estimator. □

Example IV.4. *Consider the case of example IV.1 in which $T = 2$, $r = 1$ and we're just trying to estimate G_0. Suppose we collect data from n_D runs of the system where each of the inputs are generated as follows:*

- $u_0 \sim \mathcal{N}(0, 1)$
- $u_1 = \psi u_0 + v$
- $\psi \sim \mathcal{N}(0, 1)$
- $v \sim \mathcal{N}(0, 1)$.

In this case, for individual runs with high probability there will be some omitted variable bias (since the probability that the product $u_0 u_1 = 0$ is 0.) However, if we combine the data sets using the class of feedback methods, we can see that in the limit, the omitted variable bias is given by:

$$\begin{aligned}&\left(\sum u_0(i)^2\right)^{-1}\sum u_0(i)\left(\psi(i)u_0(i) + v(i)\right)\\ &\to \frac{1}{n_D}\mathbb{E}_{\psi,u_0,v}\left(u_0(i)\left(\psi(i)u_0(i) + v(i)\right)\right)\\ &= 0.\end{aligned}$$

The second term in the above equation is actually a upper bound on the error one can get using the minimization algorithm 1 according to proposition IV.2. Hence, since this goes to zero, the bias of the least squares solution given by the algorithm must go to zero asymptotically and since we need infinitely many measurements for the bias terms to vanish, we can calculate the variance of the error terms which scale as $\frac{1}{S(n_D)\sigma_e^2}$ and apply Chebyshev's inequality once again to attain the desired result.

The above example illustrates an important question, namely, under what circumstances does there exist a set of inputs with bias that effectively cancels out. We begin with a straightforward result:

Lemma IV.5. *For a set of inputs $U_1, ..., U_N$, denote the autocovariance of the input U_j at τ as $K_j(\tau)$. Suppose that the inputs are chosen from some randomized underlying process such that $K_j(\tau) \sim K(\tau)$ are themselves random variables. Then if $\mathbb{E}_j K_j(\tau)$ have mean zero, then the estimator obtained by Algorithm 1 is asymptotically unbiased.*

Proof. The proof of this result follows from the structure of $U_r^\top Z_r$ which contains unnormalized entries of the empirical covariance matrix for some realization of inputs U_j. In this case, when we normalize by $\left(U_r^\top U_r\right)^{-1}$ we note that each entry of the matrix $U_r^\top Z_r$ is a realization of $\mathbb{E}(K(\tau)) + \eta$ where η is some finite variance mean zero noise. In this case, as we add multiple bias terms $U_r^\top Z_r$ from different systems together and multiply by $\left(U_r^\top U_r\right)^{-1}$ which is upper bounded by $\frac{1}{cT} I$ we obtain that, by the law of large numbers, the error terms η will be driven to zero by the normalization and that we are left with only the expectation $\mathbb{E}(K(\tau)) = 0$. Hence we have an asymptotically unbiased estimate. □

The above result shows the conditions under which a collection of inputs will work well with our algorithm. But it begs the question of whether or not commonly chosen input sequences actually fall into this category. Maybe more specifically, is there a set of controllers that yield closed loop inputs with the aforementioned property? It turns out that there is.

Corollary IV.5.1. *Consider a class of inputs originating from a discrete time linear system under a different linear feedback in which the feedback has the property that the markov parameters of the feedback are mean zero random variables. In this case, inputs and outputs collected from such a class of inputs have the property described in the previous lemma.*

Proof of this result is straightforward by expansion through the markov parameters of both systems then taking expectation. This result gives a sufficient condition under which a set of inputs will *provably* work well using algorithm 1. This is not to say that an arbitrary set of inputs can not be incrementally debiased using our method as well, but this describes a setting in which the success of our method is asymptotically guaranteed.

C. Bias Elimination Methods

The above result provides an effective empirical technique to reduce the bias when estimating the impulse response parameters of a linear system when the inputs to the system are generated in an arbitrary way. However, recall the example provided at the outset of this section completely eliminated the bias (at the potential expense of adding additional noise to the system.) This result raises the question, to what extent is it possible to remove bias from estimation using inputs generated in closed loop?

To answer this question, we consider the following scenario. Suppose that we have a single dataset made up of inputs $\{u_0, \dots, u_T\}$ and outputs $\{y_1, \dots, y_{T+1}\}$ each of length $T+1$. Then, as noted above, if we are trying to estimate the first r impulse response parameters G_r we obtain the omitted variable bias $(U_r^\top U_r)^{-1} U_r^\top Z_r \delta_r$. Note that the matrix $U_r^\top Z_r$ has a very particular structure given T and r.

Proposition IV.6. *The matrix $M_r = U_r^\top Z_r$ given a data set of length $T+1$ has the form such that*

$$M_r[i,j] = \sum_{i=0}^{T-r-j} U_i U_{i+r+j-i} \tag{4}$$

where matrix indexing starts with $M_r[0,0]$.

Proof. The proof of this result is immediate upon calculating the product. □

Since each of these matrices are similarly structured, for any general input sequence $\{u_i\}$ we know how the empirical covariance matrix $U_r^\top Z_r$ should behave. Now note that if we ran the system again, this time with the input

$$\begin{bmatrix} -u_0 & 0 & \dots & 0 & u_T \end{bmatrix}$$

. If we call the Toeplitz matrix associated with these inputs U_2 we see that, combining this data with the previous data as in the first example of the section yields a bias term

$$\Big(\sum_{k\in\{1,2\}} U_{k,r}^\top U_{k,r}\Big)^{-1} \sum_{k\in\{1,2\}} U_{k,r}^\top Z_{k_r} \delta_r.$$

But note that each of the terms $u_0 u_T$ in the original matrix $U_{1,r}^\top Z_{1,r}$ cancels using $U_{2,r}^\top Z_{2,r}$. Continuing in this fashion we can add the inputs

$$\begin{bmatrix} -u_0 & 0 & \dots & u_{T-1} & 0 \end{bmatrix}$$
$$\vdots$$
$$\begin{bmatrix} -u_0 & u_1 & 0 & \dots & 0 & 0 \end{bmatrix}$$

we eliminate all terms involving u_0 in the original matrix M_r while leaving the rest of the matrix unchanged. We can then do the same thing setting $u_0 = 0$ and adding new inputs of the form

$$\begin{bmatrix} 0 & -u_1 & 0 \dots & 0 & u_j & 0 & \dots \end{bmatrix}$$

to cancel all the terms in M_r involving u_1. We can continue adding inputs of this form until the matrix $\sum U_{k,r}^\top Z_{k_r} = 0$ (albeit, we are also adding a fair amount of noise during this procedure. However, the method described leads us to state the following result:

Proposition IV.7. *For any data set $\{u, y\}$ generated using closed loop inputs, there exists a sequence of inputs and outputs $\{u^{(i)}, y^{(i)}\}$ such that, the resulting least squares estimator for G_r obtained by appending all the data together is unbiased.*

Proof. The proof follows from the above discussion where you construct $T(T-1)/2$ input sequences in order to cancel terms involving each u_i in sequence. The result is that $\sum U_{k,r}^\top Z_{k_r} = 0$ which forces the bias to be zero in any realization of the

noise sequence, regardless of the initial dependence of u on y or e. □

Note that each of the inputs of the above process could (and should!) be paired with random white noise in order to eliminate the bias originating from the original correlated inputs while still minimizing the estimation error due to the measurement noise e.

The above results demonstrate that it is possible, even when starting with closed loop input output data, to construct a regression that is unbiased. However, since the method requires running the system many additional times, we may add some noise to the estimate. Additionally, it may be possible that it is impossible to run the system again. In this case we need to take a different approach to obtain an unbiased estimator of G_r. Suppose, for the remainder of these results that the inputs u are somehow correlated, but independent of the noise E.

Specifically, suppose we have data from n_D runs of the LTI system, each of length T with $n_D > T$. Suppose also that the inputs are rich enough, in the sense the collection of inputs span $\mathbb{R}^T$. Now, construct a matrix $\phi_u \in \mathbb{R}^{T\times T}$ such that the rows of ϕ_u are T linearly independent inputs $u^{(i)}$ from the set of collected inputs. In this case we have the following result:

Theorem IV.8. *Suppose that the above assumptions are satisfied. Then an unbiased estimator of G_r can be constructed using the matrix ϕ_u.*

Proof. We will prove this result in two parts. First, we note that the unit impulse input given by $u_0 = 1$ and $u_i = 0$ for all $i > 0$ has the property that, using the notation as before, $U_r^\top Z_r = 0$. This can be easily verified by using proposition IV.6 since the terms in $U_r^\top Z_r$ are sums of products of inputs with no cross terms u_i^2. Thus using the unit impulse input gives an unbiased estimator of G_r if we solve for $\hat{G}_r$ using the usual least squares solution. However, the inputs from the dataset are potentially derived from some closed loop process and thus using these to estimate G_r naturally results in nonzero omitted variable bias. Denote

$$u_{\text{imp}} = \begin{bmatrix}1 & 0 & \dots & 0\end{bmatrix}^\top.$$

Then, due to the linearity of the system, if we solve for $v = \phi_u^{-1}u_{\text{imp}}$, we see that the linear combination of inputs and outputs given by

$$\begin{aligned}\sum_{i=1}^{T} v_i Y^{(i)} &= \sum_{i=1}^{T} v_i U^{(i)} G + \sum_{i=1}^{T} v_i E^{(i)} \\ &= I_T G + \sum_{i=1}^{T} v_i e^{(i)}\end{aligned}$$

where I_T is the identity matrix of size $T \times T$. In this case, we see that using the synthetically constructed input $\sum_{i=1}^{T} v_i u^{(i)}$ with the corresponding output $\sum_{i=1}^{T} v_i y^{(i)}$, then solving for the least squares solution for G_r with this input and output gives

$$\hat{G}_r = G_r + \sum_{i=1}^{T} v_i E_r^{(i)}$$

where E_r is equal to the first r terms for each $E^{(i)}$. But clearly, since each $E^{(i)}$ consists of iid Gaussian noise, this process yields an unbiased estimator. □

The above proof demonstrates that, given a rich enough data set, we can effectively remove the bias from a least squares estimate of the first r impulse parameters by taking a specified linear combination of the inputs and outputs. This process yields the error term $\sum_{i=1}^{T} v_i E_r^{(i)}$ which is mean zero with variance $||v_r||_2^2\sigma_e^2$ where v_r is the first r terms of the vector v calculated in the proof of theorem IV.8. Hence, the error scales with the norm of $||v_r||_2^2$. Note that we can obtain the bound:

$$||v_r||_2^2 \leq ||\phi_u^{-1}||_2^2 \leq \frac{1}{\sigma_{min}^2(\phi_u)}$$

by construction. Thus, if we can bound $\sigma_{min}^2(\phi_u)$ from below, we will also be able to bound the error of our synthesized least squares estimator.

The above approach can be used to construct a consistent estimator of G_r in the case that the dataset is growing in a particular way. Specifically, suppose that we continue collecting inputs and outputs from the system, with zero initial state, such that for any given n_D input output datasets, with high probability there exists $\alpha\lfloor\frac{n_D}{T}\rfloor$ subsets each of size T each of which span $\mathbb{R}^T$. Suppose also that the matrix $\phi_u(i)$ has the property that $\sigma_{min}(\phi_u(i)) > c_{min}$ for some constant $c_{min} > 0$. Then we have the following consistency result:

Theorem IV.9. *Suppose the data collected satisfy the above stipulations that for any given n_D input output datasets, with high probability there exists $\lfloor\alpha\frac{n_D}{T}\rfloor$ subsets each of size T each of which span $\mathbb{R}^T$ for $\alpha \in (0, 1)$. Suppose also that the matrix $\phi_u(i)$ has the property that $\sigma_{min}(\phi_u(i)) > c_{min}$ for some constant $c_{min} > 0$. Then there exists a consistent estimator of G_r regardless of how the individual inputs u were generated.*

Proof. Given a data set such as the one described we can apply the result in theorem IV.8 to get $\lfloor\alpha\frac{n_D}{T}\rfloor$ synthetic inputs and outputs that satisfy the following relationship:

$$\bar{Y} = \bar{I}_{\lfloor\alpha\frac{n_D}{T}\rfloor}G + \sum_{j=1}^{\lfloor\alpha\frac{n_D}{T}\rfloor}\sum_{i=1}^{T} v(j)^{(i)}E(j)^{(i)}$$

where $v(j)^{(i)}$, $E(j)^{(i)}$ are the ith members of the jth dataset, $\bar{Y}$ is the column vector of appended outputs, and $\bar{I}_{\lfloor\alpha\frac{n_D}{T}\rfloor}$ is the $T\lfloor\alpha\frac{n_D}{T}\rfloor \times T$ matrix of stacked $T\times T$ identity matrices. In this case, however, as was discussed above, the least squares

estimate for G_r from this equation is unbiased with error upper bounded by

$$\frac{1}{\lfloor \alpha \frac{n_D}{T} \rfloor} \sum_{j=1}^{\lfloor \alpha \frac{n_D}{T} \rfloor} \sum_{i=1}^{T} v(j)^{(i)} E(j)^{(i)}.$$

This error is clearly mean zero and since each term $\sum_{i=1}^{T} v(j)^{(i)} E(j)^{(i)}$ has variance $||v(j)_r||_2^2 \sigma_e^2$ and that $||v(j)_r||_2^2 \leq \frac{1}{c_{min}^2}$ we obtain the following upper bound on the variance:

$$\text{Var}(\hat{G}_r) \leq \frac{1}{\lfloor \alpha \frac{n_D}{T} \rfloor} \frac{1}{c_{min}^2} \sigma_e^2.$$

A simple application of Chebyshev's inequality shows that as $n_D \to \infty$, $\hat{G}_r \to G_r$ in probability, and thus we constructed a consistent estimator of the first r impulse parameters of an LTI system using potentially closed loop inputs.

□

Note that the assumptions in the above theorem are not terribly onerous, if there is some level of randomization in how the feedback was selected or better yet that there is some small amount of random noise injected into the inputs. Then the result from [17] demonstrates that with high probability, there will exist subsets from our whole dataset that are linearly independent and which each have smallest singular value bounded below by a term that scales as $\sqrt{T} - \sqrt{T-1}$. So as long as T is finite, in this case we are guaranteed to obtain a consistent estimator.

V. Simulations

Several simulations were run in order to determine the effectiveness of the algorithms designed in the previous section. The first study simulated the case when there exist correlated inputs u_0, u_1 of length 2 to an LTI system. In particular, the inputs were generated such that the initial input u_0 is distributed as a zero mean, unit variance Gaussian and u_1 is generated as $u_1 = \frac{1}{2}u_0 + v$ where v is generated as an independent zero mean, unit variance Gaussian. Unbiasing inputs are then generated by putting the inputs $-u_0, u_1$ through the system and measuring the outputs. Errors from regression for G_0 only using the initial correlated data are clearly biased, with a mean around .2 while error using both the biased and unbiasing data have a mean error of 0, emphasizing the efficacy of the unbiasing algorithm mentioned in proposition IV.7. These results are shown in figure 1. Additionally, the convergence of the debiasing method is shown in figure 2, demonstrating that no matter how much correlated data is collected, it's resulting bias can effectively be unwound and result in an asymptotically consistent estimator when additional, specifically selected inputs are used.

Several experiments were conducted in order to test the effectiveness of algorithm 1. First, repeatedly generating biased datasets and then using algorithm 1 to debias the data in order to estimate the parameters of the system results in an unbiased estimator as seen in figure 3. Furthermore, this algorithm is

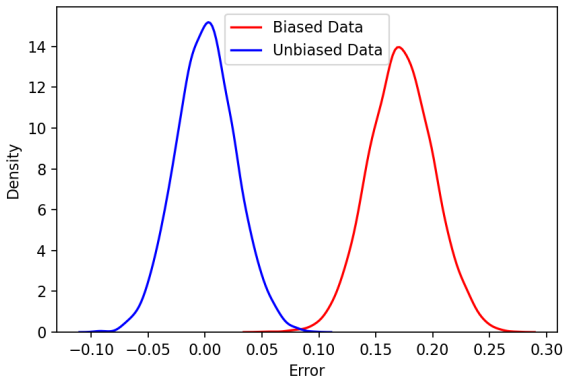

Fig. 1. For this experiment, we ran the system described in example IV.4 with ψ fixed at $\frac{1}{2}$. This input structure naturally leads to correlation between the inputs which, when only estimating g_0 therefore leads to omitted variable bias. By using the de biasing methodology of proposition IV.7, we empirically show here that while the original data is clearly biased, this data combined with new data that removes the omitted variable bias results in a mean zero (unbiased) estimator of the parameter g_0 as predicted by the theory.

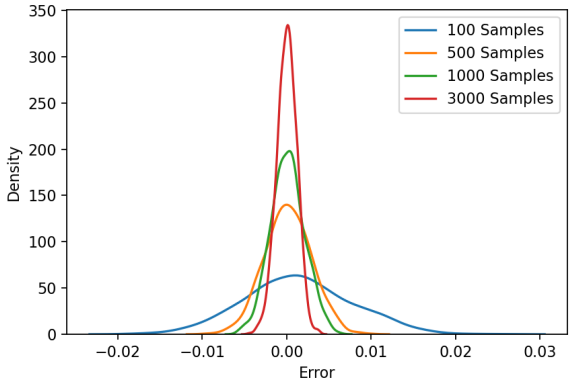

Fig. 2. Using the same methodology as in figure 1, we used the debiased method of proposition IV.7 using larger and larger datasets to demonstrate the asymptotic consistency of the resulting least squares estimator for g_0.

shown to be effective regardless of the size of the data set. In particular, as the number of runs of the system using correlated inputs increases, the algorithm always successfully removes bias, while the error from bias using the entire dataset can be significant. This can be seen in figure 4

VI. Conclusion

In this paper, we have considered new ways to combine measurements taken from the same system in an arbitrary fashion to generate unbiased estimates of system parameters. We have shown that under some reasonable assumptions, our methods result not only in less biased or unbiased estimates of the impulse response parameters of a system, but that in the limit they converge to the correct values. We demonstrate through simulations the efficacy of these methods and show how they could be used in practice. Most importantly, our results suggest new ways to combine measurements from

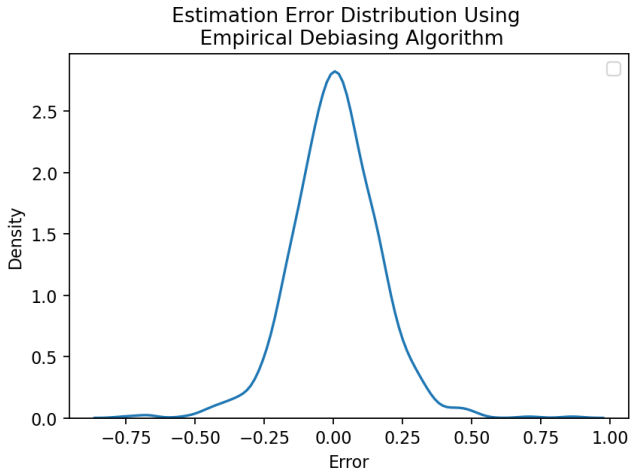

Fig. 3. In this plot, we show the distribution of the estimation errors estimating g_0 using a fixed dataset generated using correlated inputs as in figure 1 and algorithm 1 to empirically select the best subset of the data that will result in the least biased estimator. Clearly, the distribution is zero mean suggesting that the algorithm is effective in constructing an unbiased estimator.

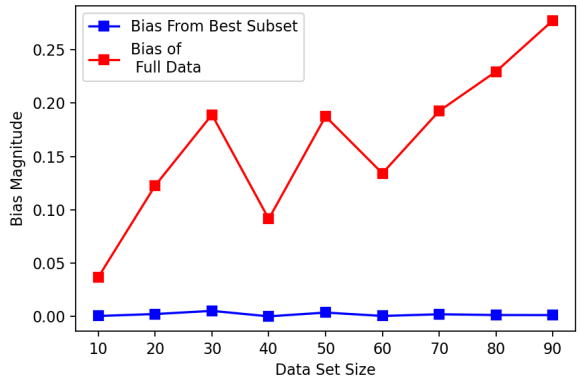

Fig. 4. This figure shows how large the bias can be for different sized data sets (where again the data is generated as in figure 1 when all the data is used (red) and when the best subset of that data is used (blue).

different, but related, systems to effectively perform system identification.

One exciting direction for future research is the possibility of using similar techniques and exploiting the characteristics of linear systems to take measurements from "similar" systems, that are all related via a low rank structure. In this case, we can do similar operations as in this paper, to combine observed data in order to learn the behavior of systems that have not been seen a priori as long as something about the low rank structure is known.

Note for Peko: I think that the way to frame this paper is to note that bias from the estimation process is endemic in these methods. Also, the benefit of a linear model is that it is incredibly interpretable (I can perhaps tell you exactly how interest rates matter for future inflation values if I do the "correct" regression.) The trick is that doing the "correct" regression from collected data is usually pretty impossible either do to correlations between inputs and noise (closed loop) or misspecification of the model structure. In either case, the algorithms presented in this paper eliminate the bias created by those two problems and allow one to more confidently estimate the "causal" effects of an input on an output. In this sense then, while we can get good *predictions* from a biased linear model, we throw interpretability out the window. In order to obtain a better understanding of the underlying "causality", unless we have perfect data and a perfect understanding of the relationship between inputs and outputs, we need to employ some method of bias reduction. I think the trick then is to a.) come up with some experiments to show how much coefficients can change with different data and varying model structure assumptions and b.) figure out a good conference or journal to submit this too.

References

[1] L. Ljung, "System identification," in *Signal analysis and prediction*. Springer, 1998, pp. 163–173.

[2] L. Ljung, R. Singh, and T. Chen, "Regularization features in the system identification toolbox," *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 745–750, 2015.

[3] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite time lti system identification." *J. Mach. Learn. Res.*, vol. 22, pp. 26–1, 2021.

[4] T. Sarkar and A. Rakhlin, "How fast can linear dynamical systems be learned?" *CoRR*, vol. abs/1812.01251, 2018. [Online]. Available: http://arxiv.org/abs/1812.01251

[5] S. Oymak and N. Ozay, "Non-asymptotic identification of lti systems from a single trajectory," in *2019 American control conference (ACC)*. IEEE, 2019, pp. 5655–5661.

[6] A. Goldenshluger and A. Zeevi, "Nonasymptotic bounds for autoregressive time series modeling," *Annals of statistics*, pp. 417–444, 2001.

[7] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems," *arXiv preprint arXiv:2003.11227*, 2020.

[8] ——, "Finite-time system identification and adaptive control in autoregressive exogenous systems," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 967–979.

[9] ——, "Explore more and improve regret in linear quadratic regulators," *arXiv preprint arXiv:2007.12291*, 2020.

[10] B. Lee and A. Lamperski, "Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 3419–3424.

[11] S. Dean, N. Matni, B. Recht, and V. Ye, "Robust guarantees for perception-based control," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 350–360.

[12] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in neural information processing systems*, vol. 24, pp. 2312–2320, 2011.

[13] K. A. Clarke, "The phantom menace: Omitted variable bias in econometric research," *Conflict management and peace science*, vol. 22, no. 4, pp. 341–352, 2005.

[14] P. P. Jovanis, J. Aguero-Valverde, K.-F. Wu, and V. Shankar, "Analysis of naturalistic driving event data: Omitted-variable bias and multilevel modeling approaches," *Transportation research record*, vol. 2236, no. 1, pp. 49–57, 2011.

[15] M. P. Murray, "The bad, the weak, and the ugly: Avoiding the pitfalls of instrumental variables estimation," *Available at SSRN 843185*, 2006.

[16] I. Andrews, J. H. Stock, and L. Sun, "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, vol. 11, pp. 727–753, 2019.

[17] M. Rudelson and R. Vershynin, "Smallest singular value of a random rectangular matrix," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 62, no. 12, pp. 1707–1739, 2009.