Minimax Optimal Model Selection for Closed Loop Identification: a Data Driven Approach

Dalton Jones, MIT
daltonj@mit.edu

Munther A. Dahleh, MIT
dahleh@mit.edu

***Abstract*—This paper explores the non parametric estimation of the impulse response of a linear time invariant (LTI) system with finite amount of noisy input output data. The inputs to the system as well may be chosen almost arbitrarily and can be a function of past outputs and noise, a novel extension of previous work. The algorithm used to do this estimation must simultaneously select the best model class as well as the parameters in order to minimize some measure of estimation error. This methodology takes special care to address the issue of overfitting in model sets with too many parameters and tries to balance bias from model misspecification with the variance originating from the noise in order to optimally select a model order for estimation. We demonstrate that, under very loose assumptions, bounds can be put on the error obtained from our method and show how our algorithm can be used with several illustrative examples. We use methods from self normalized martingale theory in order to guarantee the asymptotic consistency of our estimator for finite impulse response (FIR) systems and demonstrate how it can be used to approximate the behavior of infinite impulse response (IIR) systems. Simulations explore the effectiveness of these methods as noise and filter length vary.**

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

I. Introduction

One of the most effective ways to understand the behavior of an LTI system is through the characteristics of the impulse response. As such, accurately estimating the parameters of this function allows the engineer to fully characterize a system of interest. Perhaps the most common way to estimate these parameters is through a least squares type estimate which transforms potentially noisy data into an accurate representation of the underlying impulse response. There is some ambiguity in how least squares is applied in this case since, in general, there can be an infinite number of impulse response coefficients. This approach has been studied through the lens of model selection because of the observation that, in general, with a small amount of noisy data, lower dimensional models result in better estimates (though they add bias) than their higher dimensional counterparts. In this way, model selection usually proceeds by minimizing some function that approximates estimation error. This criterion allows the engineer to compare between different model sets, each of which is given a likely estimation error through the machinery of the aforementioned approximation function.

In general, these criteria are designed assuming that the inputs are additive white Gaussian noise or otherwise nicely structured, in which case the outputs and the prediction errors can be described by known distributions (such as the squared distribution) as in [1]. To the authors knowledge, no such criteria exist to evaluate the optimality of different model classes when identifying a system using closed loop measurements. This paper develops just such a function, and demonstrates it's efficacy in model selection with closed loop inputs for system identification of finite impulse response systems. We show that should the inputs satisfy certain excitation conditions that our procedure is asymptotically consistent. We then present extensions of these results for different output noise characteristics and demonstrate the efficacy of our methodology for identifying stable IIR systems.

Model selection and regularization is a classical problem within many fields of statistics. In almost any estimator using a finite amount of data, there is a delicate balancing act to limit the number of parameters so as to not simply fit the model to noise while still retaining enough complexity to minimize bias injected into the model from misspecification or underparameterization. This bias variance trade-off is ubiquitous in statistical learning theory and can be dealt with in several different ways. The most common of these is through so-called regularization, specifically by adaptively penalizing the magnitude of learned parameters through some parameter λ. The paper [2], demonstrated that FIR models can be learned consistently using regularized least squares, despite the potential bias that regularization induces. In that work, regularization was done solely through the parameter λ and only used open loop operation of the system with random inputs, whereas in this paper, we may vary the size of the regression in an intelligent way while also adding the same kind of regularization through some parameter λ.

Our work builds on the work done in [1], [3]–[5] in which the authors developed an information theoretic criteria to compare different model classes by using output data to evaluate the accuracy of estimates originating from a nested set of models. This work is particularly compelling in that it provides confidence sets and statistical tests to determine the likelihood of each model producing the observed data. Moreover, these tests, for a fixed confidence level reduce to the classical Akiake Information Criterion (AIC) [6] which is among the first and most celebrated ways to select optimal model complexity. The results from the above work can then be combined with a zero-pole modeling procedure as in [7] to obtain a rational transfer function.

In addition, the estimation of impulse response parameters is considered in [8], [9] in conjunction with subspace techniques to obtain a parametrization of an LTI system through noisy input output data. Moreover, in [9] the authors developed a procedure to select the model order of an identified LTI system by minimizing a measure of the output error resulting from the estimated parameters and the observed measurements. Such model selection procedures are related to other subspace based procedures as [10], [11] which show the asymptotic consistency of using criteria related to the singular values of the system Hankel matrix to approximate an LTI system. The authors in [12] demonstrate a related asymptotically consistent two step procedure in which a large model is learned and then reduced. While each of these works results depend on the independence of inputs and noise, the application of least squares to system identification with correlated inputs and noise was considered in [13], [14] in which a certain self normalized martingale inequality was utilized to upper bound the error of the estimates and show that under certain conditions this procedure results in negligible estimation error. Such techniques will be instrumental in developing our own closed loop model selection criterion.

Similar to this work, closed loop system identification in finite time is considered in [15] in which the Markov parameters of the optimal one step ahead predictor are obtained. These estimates are then used to estimate a reduced order model in which the output error is bounded by some specified constant. Our work is additionally related to the adaptive learning and control schemes considered in [16], [17].

In this paper, we develop a minimax criterion that measures the worst case parameter error for a given model dimension. This criterion can then be used to compare and select the a minimax optimal model class for a given data set. We demonstrate that model selection using this method is asymptotically consistent for output error FIR models and then extend our results to estimate the impulse response of the optimal one step ahead predictor for an FIR system with colored noise. Finally we consider the application of these methods to general stable LTI systems.

The rest of the paper is as follows: in section II we cover the notation and preliminaries used to develop our results. Section III describes the formulation of our problem while section IV states the main results and gives several examples of when they might be used. Results of simulations are presented in section V and the paper concludes with section VI.

II. Notation and Preliminaries

We consider discrete time linear models of the form:

$$y_t = \sum_{i=1}^{n} g_i u_{t-i} + e_t \tag{1}$$

such that the model is causal, single input/single output and stable. We also use the notation $y_t = G(q)u(t) + e_t$ where q is the shift operator, $qu_t = u_{t+1}$. The inputs may be chosen arbitrarily while the noise term e_t is centered additive sub-Gaussian white noise with variance σ_e^2. Note that we do not make the standard system identification assumption that the noise e_t is independent of the input u_t, and in fact it is possible that these value are correlated (with past noise driving future adaptive inputs.) We do, however, stipulate that the input u_t must be independent of e_t or in other words that there is some delay in how the input u_t utilizes past noise terms.

A finite amount of data in the form of inputs $\{u_1, ..., u_t\}$ and outputs $\{y_1, ..., y_t\}$ are available. The objective of this paper is to develop a methodology to utilize this data to accurately estimate the true parametrization of the system

$$G = \begin{bmatrix} g_1 & \dots & g_n \end{bmatrix}^\top \tag{2}$$

despite the fact that outputs from the models are corrupted by noise terms and that inputs, noise and outputs may be correlated in some complex manner. For most of the paper, we assume that we are dealing with only finite impulse response systems, however, later we demonstrate that our method can be used to estimate the parameters of stable infinite impulse response systems to an arbitrary degree of accuracy.

Hence, the model we are working with above can be modified such that

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

where the term $n_{m,t} = \sum\limits_{m+1}^{n} G_i u_{t-i}$ represents the error associated with truncating at the mth parameter.

In this way, collecting terms we have that:

$$\begin{aligned} Y_t &= \begin{bmatrix} y_t & y_{t-1} & \dots & y_m \end{bmatrix}^\top \\ N_{m,t} &= \begin{bmatrix} n_{m,t} & n_{m,t-1} & \dots & n_{m,m} \end{bmatrix}^\top \\ U_{m,t} &= \begin{bmatrix} u_{t-1} & u_{t-2} & u_{t-3} & \dots \\ u_{t-2} & u_{t-3} & \dots & \ddots \\ \vdots & & \vdots & \ddots \\ u_{m-1} & u_{m-2} & \dots & u_0 \end{bmatrix} \\ E_t &= \begin{bmatrix} e_t & e_{t-1} & \dots & e_m \end{bmatrix}^\top \\ G_m &= \begin{bmatrix} G_1 & G_2 & \dots & G_m \end{bmatrix}^\top \end{aligned}$$

Now we can rewrite our system equation as:

$$Y_t = U_{m,t} G_m + N_{m,t} + E_t$$

We will for the most part drop the subscript m for notational simplicity but recall that the above expression depends explicitly on the choice of this parameter.

III. Motivation and Problem Formulation

As noted above, the outputs $\begin{bmatrix} y_t & \dots & y_1 \end{bmatrix}^\top$ can be written as the product between an $t \times n$ Toeplitz matrix of inputs U plus some error vector consisting of independent entries. Such that

$$Y = UG + E.$$

There exist many methods to estimate the value of G from observed input output data U and Y, but by far the most popular is the least squares estimator. This estimator projects the output onto the span of the input in such a way that reconstructs the behavior of G. In particular, the solution is given by

$$\hat{G} = (U^\top U)^{-1} U^\top Y.$$

This calculation necessarily requires knowledge of n, the length of G in order to be successfully executed, something that in general is not known a priori. Furthermore, note that since $Y = UG + E$ we can see from the above definition that $\hat{G} = G + (U^\top U)^{-1} U^\top E$ and hence the quality of the estimation can be determined through an analysis of the noise term $(U^\top U)^{-1} U^\top E$. In the usual case when U and E are independent, we could appeal to asymptotic results to show that, since the covariance between U and E is zero, this term goes to zero quickly and we are left with a consistent estimator of the true parameter. On the other hand, with limited noisy data, the term $(U^\top U)^{-1} U^\top E$ may incorporate too much of the noise into the parameter estimate. This problem becomes all the more apparent the closer the g_i get to zero and in fact, in this case, it is preferable to simply set this parameter equal to zero than let it depend on noisy data. Setting $g_i = 0$ is the same as constraining the space over which we can search for the parameters of G to be an $n-1$-dimensional subspace.

The problem becomes even worse when we relax the requirement that U and E are independent (perhaps because of some sort of closed loop relationship between outputs and inputs.) In this case, it's not clear from first glance that the error in the estimate $(U^\top U)^{-1} U^\top E$ even converge to 0 with more measurements.

A similar problem was considered in [14] where the authors demonstrate that as long as the error terms e_t are independent from past inputs and outputs and that they are subgaussian, one can obtain convergence rates for the regularized least square estimator:

$$\hat{G} = (U^\top U + \lambda I)^{-1} U^\top Y.$$

In particular, utilizing tools from the theory of self normalized martingale theory, one can show that this error term converges at a rate of $\mathcal{O}\left(1/\sqrt{T}\right)$. Thus for this work, we will utilize the regularized least squares estimator. It is worth noting that the regularization term itself creates some bias in the estimate, but also that as the amount of data increases, this bias shrinks to a negligible value when the input data is rich enough.

Note again that in order for the above estimation to have the desired convergence properties, the order of the filter n must be known. In the case that it is not, and a filter dimension $m < n$ is used to estimate the lower order system G_m the error term is given by:

$$(U_m^\top U_m + \lambda I)^{-1} U_m^\top E + (U_m^\top U_m + \lambda I)^{-1} U_m^\top N_m + ||G - G_m||_2$$

where the first term accounts for the variance in the estimate and the subsequent terms account for the bias originating from model misspecification. Generally speaking, the first term increases with m while the second decreases. Balancing these to minimize overall error by intelligently selecting m is the crux of this work.

In order to obtain a metric that can be used to optimally select the model order for an estimator of a filter, some information about the true model and how it was operated must be assumed. This information need not be very detailed but will allow practitioners a way to understand how increasing model order decreases truncation error.

To see why this is necessary, consider the contrived case of a system with arbitrary delay and gain. In particular, suppose that n is increasing and that $g_i = 0$ for $i < n$ and that $g_n = \alpha$ where α can be made arbitrarily large. In this case, without limiting α and or the value of n somehow, almost nothing of value can be said about the truncation error

$$(U_m^\top U_m + \lambda I)^{-1} U_m^\top N_m + ||G - G_m||_2$$

which in this case if $m < n$ can grow arbitrarily large. With this in mind, some structural assumptions should be made about the parameters $\{g_i\}$. In this work, we consider the following different scenarios but we note that our methodology could be extended to many other possible impulse response structures. The scenarios are as follows:

1) Either $|g_i| \leq G_{max},\ |u_i| \leq u_{max}, n \leq n_{max}$ or
2) $|g_i| \leq \alpha\beta^i,\ |u_i| \leq u_{max}$. Where $\alpha > 0$ and $\beta \in (0, 1)$.

Note that in the first case, most of these assumptions are satisfied with high probability for subgaussian random variables and internally stable feedback. The second assumption amounts to stability in the asymptotic case as $n \to \infty$. In order to provide motivation for these assumptions, consider the following negative result.

A. An impossibility result

I will try to keep this as succinct as possible. In the general case, estimating in closed loop with model misspecification (in terms of order) can lead to arbitrarily large estimation error from bias. In our case, suppose U_m is the Toeplitz matrix of inputs corresponding to the first m impulse response parameters and Z_m is the (possibly infinite) Toeplitz matrix corresponding to the other impulse response parameters. Then the error from omitted model order is

$$(U_m^\top U_m)^{-1} U_m^\top Z_m g_{n-m} + (U_m^\top U_m)^{-1} U_m^\top E$$

where the first term comes from the misspecification and the second from the noise in the model. If input is chosen through closed loop, its possible that the covariance of $\mathrm{Cov}(u_t, u_t - i)$ can be arbitrary and that the first term can also as a result be almost anything.

Example III.1. *As an example, consider the case where u_t is the output of a (noiseless) linear controller $K(z)$ such that the impulse response of K decays slower than α^k for $\alpha \in (0, 1)$. Suppose also that the impulse responses of the open loop system $g_i \geq \beta^i$ for some $\beta \in (0, 1)$. Both of these systems* *are stable, however, for any truncation m, we see that the error in the above expression can be* ***lower bounded*** *by*

$$(\alpha\beta)^m \frac{1-(\alpha\beta)^{n+1}}{1-(\alpha\beta)}$$

which, since these values are arbitrary, as $\alpha, \beta, n \to 1$, this error can be made infinitely large for any m. So in essence ***without some knowledge of the system, developing a non parametric estimation technique that grows*** *m* ***with data is impossible with general feedback without some knowledge of the system.*** *Note that this problem is solved with knowledge of β and α. Since we can then at least get a sense of how to grow m.*

IV. Model Selection With Correlated Data

Note that, in order to select the minimax optimal model order for the least squares estimator, we must have some function that approximates the parameter error $||G - \hat{G}_m||_2$. Note that this error can be decomposed using the triangle inequality into

$$||G - \hat{G}_m||_2 \leq ||G - G_m||_2 + ||G_m - \hat{G}_m||_2$$

where the first term represents pure truncation error, and the second term represents estimation error. Note that the estimation error contained in the second term will also has some bias due to the model misspecification as well. As remarked in the previous section, the truncation error term $||G - G_m||_2$ can be upper bounded by using prior knowledge of the system while the estimation error can be bounded using the following results. We first state a result from [14] that will then be used to upper bound estimation error originating from the noise in the system.

Theorem IV.1 (Abbasi-Yadkori, 2011)**.** *Suppose that $\mathcal{F}_t$ is a filtration and that $m_k \in \mathbb{R}^d$ is a stochastic process adapted to $\mathcal{F}$ while η_k is a real valued martingale difference process also adapted to $\mathcal{F}$. Suppose that η_k is R-conditionally subgaussian given $\mathcal{F}_{k-1}$. The martingale*

$$S_t = \sum_{i=1}^{t} \eta_i m_{i-1}$$

and the matrix

$$V_t = \sum_{i=0}^{t} m_i m_i^\top, \;\; \bar{V}_t = V_t + V$$

for some positive definite V have the property that with probability at least $1-\delta$

$$||S_t||^2_{\bar{V}_t^{-1}} \leq 2R^2 \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right).$$

The above result allows us to understand the rate of decay of the error in the estimate in terms of the behavior of the weighted empirical covariance matrix $U^\top U + \lambda I$. It will be used to show the following technical result:

Theorem IV.2. *The least squares solution $\hat{G}_m$ to the above problem satisfies with probability at least $1-\delta$:*

$$\mathrm{Tr}((G_m - \hat{G}_m)^\top(\lambda I + U^\top U)(G_m - \hat{G}_m)) \leq \beta_t$$

where β_t is given as

$$\left(\sqrt{\sigma_e^2 \log\left(\frac{\det(\lambda I + U^\top U)^{1/2}}{\delta\lambda^{1/2}}\right)} + ||G||_F^2\sqrt{\lambda} + b(t,m)\right).$$

$b(t,m)$ is a bias term that corresponds to one of the two assumptions made above regarding the structure of G.

Proof. First of all note that

$$\begin{aligned}\hat{G}_m &= (U^\top U + \lambda I)^{-1}U^\top(UG_m + N + E)\\ &= (U^\top U + \lambda I)^{-1}U^\top U G_m + (U^\top U + \lambda I)^{-1}(N+E)\\ &+ \lambda(U^\top U + \lambda I)^{-1}G_m - \lambda(U^\top U + \lambda I)^{-1}G_m\\ &= G_m + (U^\top U + \lambda I)^{-1}(N+E) - \lambda(U^\top U + \lambda I)^{-1}G_m\end{aligned}$$

Now let X be any matrix of compatible dimension, then we see that

$$\begin{aligned}&|Tr(X(\hat{G}_m - G_m)^\top)|\\ &\leq |Tr(X(U^\top U + \lambda I)^{-1}E|\\ &+ |Tr(X(U^\top U + \lambda I)^{-1}N|\\ &+ \lambda|Tr(X(U^\top U + \lambda I)^{-1}G_m^\top|\\ &\leq \sqrt{Tr(X(U^\top U + \lambda I)^{-1}X^\top)Tr(E(U^\top U + \lambda I)^{-1}E^\top)}\\ &+ \sqrt{Tr(X(U^\top U + \lambda I)^{-1}X^\top)Tr(N(U^\top U + \lambda I)^{-1}N^\top)}\\ &+ \lambda\sqrt{Tr(X(U^\top U + \lambda I)^{-1}X^\top)Tr(G_m(U^\top U + \lambda I)^{-1}G_m^\top)}\end{aligned}$$

Setting $X = (\hat{G}_m - G_m)(U^\top U + \lambda I)$ the above expression simplifies to

$$\begin{aligned}&\sqrt{Tr((\hat{G}_m - G_m)(U^\top U + \lambda I)(\hat{G}_m - G_m)^\top)}\\ &\leq \sqrt{Tr(E^\top U(U^\top U + \lambda I)^{-1}U^\top E)}\\ &+ \sqrt{Tr(N^\top U(U^\top U + \lambda I)^{-1}U^\top N)}\\ &+ \lambda||G_m||_F\end{aligned}$$

The first term in the above sum consists of conditionally subgaussian random variables, and we hence may bound this term (using results from self normalized martingale theory) by

$$\begin{aligned}&\sqrt{Tr(E^\top U(U^\top U + \lambda I)^{-1}U^\top E)}\\ &\leq \sqrt{\sigma_e^2 \log\left(\frac{det(U^\top U + \lambda I)^{1/2}}{\delta det(\lambda I)^{1/2}}\right)}.\end{aligned}$$

We can upper bound $\lambda||G_m||_F$ with $\lambda||G||_F$ giving the proper term in our bound, and then we are left with the bias from the truncation.

To simplify this term, note that we have

$$\begin{aligned}\sqrt{Tr(N^\top U(U^\top U + \lambda I)^{-1}U^\top N)} &\leq \frac{1}{\sqrt{\lambda}}||NU||_F\\ &\leq \frac{1}{\sqrt{\lambda}}u_{max}||N||_F\end{aligned}$$

where the term $||N||_F$ can generally be bounded by using some of the structural assumptions on G. Thus the term $b(t,m) = \frac{1}{\sqrt{\lambda}} u_{max}||N||_F$.

□

We can use the above result to yield a data driven criteria to balance bias and variance in our estimates when utilizing a finite data set. Note first that

$$\begin{aligned}&\sqrt{\sigma_{\min}(U^\top U + \lambda I)}||G_m - \hat{G}_m||_2 \\ &\leq \sqrt{\operatorname{Tr}\left((G_m - \hat{G}_m)(U^\top U + \lambda I)(G_m - \hat{G}_m)^\top\right)}\end{aligned}$$

and we also have by the triangle inequality that

$$||G - \hat{G}_m||_2 \leq ||G_m - \hat{G}_m||_2 + ||G - G_m||_2.$$

Putting these two equations together yields the following result:

Corollary IV.2.1. *The estimation error using the regularized least squares estimator with variable model order m can be upper bounded as*

$$\begin{aligned}||G - \hat{G}_m||_2 &\leq ||G - G_m||_2 + \frac{\left(||G||_F^2\sqrt{\lambda} + b(t,m)\right)}{\sqrt{\sigma_{\min}(U^\top U + \lambda I)}} \\ &+ \frac{\left(\sqrt{\sigma_e^2 \log\left(\frac{\det(U^\top U + \lambda I)^{1/2}}{\delta\lambda^{1/2}}\right)}\right)}{\sqrt{\sigma_{\min}(U^\top U + \lambda I)}}\end{aligned}$$

Proof. Note that the previous result along the preceding observation gives the inequality

$$\begin{aligned}&\sqrt{\sigma_{\min}(U^\top U + \lambda I)}||G_m - \hat{G}_m||_2 \\ &\leq \sqrt{\operatorname{Tr}\left((G_m - \hat{G}_m)(U^\top U + \lambda I)(G_m - \hat{G}_m)^\top\right)} \\ &\leq \left(\sqrt{\sigma_e^2 \log\left(\frac{\det(\lambda I + U^\top U)^{1/2}}{\delta\lambda^{1/2}}\right)} + ||G||_F^2\sqrt{\lambda} + b(t,m)\right).\end{aligned}$$

Dividing both sides of the inequality by $\sqrt{\sigma_{\min}(U^\top U + \lambda I)}$ and employing the triangle inequality gives the result. □

The decomposition of the error in the above corollary is suggestive of a particular bias/variance trade-off present in this class of problems. Namely, if we define:

$$f_{\text{bias}}(m, U, Y) = ||G - G_m||_2 + \frac{\left(||G||_F^2\sqrt{\lambda} + b(t,m)\right)}{\sqrt{\sigma_{\min}(U^\top U + \lambda I)}}$$

and

$$f_{\text{variance}}(m, U, Y) = \frac{\left(\sqrt{\sigma_e^2 \log\left(\frac{\det(U^\top U + \lambda I)^{1/2}}{\delta\lambda^{1/2}}\right)}\right)}{\sqrt{\sigma_{\min}(U^\top U + \lambda I)}}$$

we see that as m increases, generally speaking the bias function f_{bias} will decrease while the value of the variance function f_{variance} increases. Since the combination of the two upper bound the worst case parameter error present in the estimate, it suggests that the minimax optimal model selection should be made on the basis of the following optimization problem:

$$\hat{m}(U, Y) = \arg\max_m f_{\text{variance}}(m, U, Y) + f_{\text{bias}}(m, U, Y).$$

Note that the preceding results are stated in full generality without imposing any known structure on the coefficients of G. In this case, it's nearly impossible to upper bound $b(t,m)$ or $||G - G_m||$ and hence the above criterion is next to useless. If, on the other hand, we make some fairly general assumptions on the structure of G, we can obtain a very good closed loop model selection procedure for FIR systems. To see this, we give two examples:

Example IV.3. *Suppose that $|G_i| \leq G_{max},\ |u_i| \leq u_{max}, n \leq n_{max}$. A simple calculation shows that*

$$||G - G_m||_F \leq (\sqrt{n_{max} - m})G_{max}$$

and that

$$b(t,m) \leq t(n_{max} - m)u_{max}G_{max}$$

so that the optimal value of m given some fixed data length t can be calculated as the minimum of

$$\begin{aligned}&(\sqrt{n_{max} - m})G_{max} \\ &+ \frac{\left(\sqrt{\sigma_e^2 \log\left(\frac{\sqrt{\det(U^\top U + \lambda I)^{1/2}}}{\delta\lambda^{1/2}}\right)}\right)}{\sqrt{\sigma_{\min}(U^\top U + \lambda I)}} \\ &+ \frac{\left(G_{max}\left(\sqrt{n_{max}\lambda} + t(n_{max} - m)u_{max}^2\right)\right)}{\sqrt{\sigma_{\min}(U^\top U + \lambda I)}}\end{aligned}$$

.

The solution m^* to the above problem is going to depend on many different factors, but in particular the growth of the weighted empirical covariance matrix $(U^\top U + \lambda I)$ and the maximal values of G, u and n. Also it should be noted that as soon as $m > n$ the first term and the truncation bias term disappear since now we can simply set some parameters to zero and the true model is a part of the class. This reasoning will allow us to show that our methodology will give an asymptotically consistent estimator.

The above example was quite general, but still required some upper bound on the model order $n < n_{max}$. In the general case, if we dispense with this assumption, we may allow the model itself to grow to any arbitrary size (even infinite) and still apply the same criterion for model selection provided that certain stability conditions are satisfied.

Example IV.4. *Suppose that the impulse response parameters of and inputs into the system in question satisfy $|g_i| \leq \alpha\beta^i,\ |u_i| \leq u_{max}$. Where $\alpha > 0$ and $\beta \in (0,1)$. Note that bounding the size of the inputs is not an unreasonable assumption since, in reality, there are limits to the inputs that can be actuated and past a certain point, these size of the inputs must saturate.*

In this case we see that

$$||G - G_m||_F \leq \alpha\beta^m \sqrt{\frac{1}{1-\beta^2}}$$

and that the function

$$b(t,m) = tu_{max}\beta^m/(1-\beta).$$

Now, in order to find the optimal parameter m given the amount of data t we simply maximize the following with respect to m:

$$\begin{aligned} &\alpha\beta^m\sqrt{\frac{1}{1-\beta^2}} \\ &+ \frac{\left(\sqrt{\sigma_e^2 \log\left(\frac{\det(U^\top U+\lambda I)^{1/2}}{\delta\lambda^{1/2}}\right)}\right)}{\sqrt{\sigma_{\min}(U^\top U+\lambda I)}} \\ &+ \frac{\left(||G||_F^2\sqrt{\lambda} + tu_{max}^2\beta^m/(1-\beta)\right)}{\sqrt{\sigma_{\min}(U^\top U+\lambda I)}}. \end{aligned}$$

Note that in this optimization, because of the convergence of the upper bound on the sum of the impulse response parameters $\sum \beta^i$, one can dispense with any assumptions on the maximal order of the system n. Additionally, all stable LTI systems satisfy this property, so this is not a terribly stringent requirement. It is also clear from this example that this methodology could be easily extended to the estimation of a stable IIR filter as we will briefly discuss later in this section. The above example also clearly demonstrates a bias that decreases with m and a variance that is non-decreasing with m.

These two examples give a sense of how our closed loop minimax optimal order estimation can be used in practice. Experimental evidence of the effectiveness of this procedure will be given later in section V. Note also that in the above analysis and calculations, very little was assumed about the content of the inputs $\{u_i\}$. In fact, all that was required was that there is some delay in whatever feedback exists between y and u so that we maintain independence between e_t and the past. We can then, through the data collected and using some knowledge of the system, upper bound the worst case estimation error for each model order and select the optimal order parameter for the given data. In the subsequent results, more requirements will be put on the inputs similar to the standard "persistence of excitation" requirements common to system identification.

A. Consistency

One of the drawbacks of applying many traditional statistical techniques to finite datasets is that they rely on a sometimes infinite amount of data to converge. However, entertaining the fiction that *large enough* data sets allow us to apply asymptotic results is a cornerstone of statistics and by and large is very successful in providing meaningful predictions and results. In our case, as in [10] we would like to show that, in the limit of an infinite amount of data, that our algorithm coupled with least squares converges to the true system G. In order to show this, we need to make some assumptions regarding the structure of the data. Particularly, we need to guarantee that the spectrum of the inputs is sufficiently rich and grows in magnitude over time. In mathematical terms we can state this as:

Assumption IV.5 (Persistence of Excitation)**.** *For each m and some $t > t_m$ there exists $\alpha_m > 0$ and $\beta_m > 0$ such that*

$$\alpha_m tI \preceq U_{m,t}^\top U_{m,t} + \lambda I \preceq \beta_m tI. \tag{3}$$

Such inputs will be called persistently exciting at all orders. It is worth pointing out that if any additive white Gaussian noise is injected into the inputs independent of the noise e, then the inputs automatically satisfy this condition.

With this assumption in place we can now state our consistency result.

Theorem IV.6. *Suppose that G is a FIR system of order n, that $g_n \neq 0$, and that the inputs u satisfy assumption IV.5. Then as $t \to \infty$, the estimate $\hat{G}_{\hat{m}} \to G$.*

Proof. Note that for any $m < n$ we have

$$||G - G_m|| > c_m$$

for some constant $c_m > 0$ and that for $m \geq n$ we have both that

$$||G - G_m|| = 0$$

and that the truncation bias term $b(t,m) = 0$. Furthermore, we have by assumption IV.5 that the error due to variance from the noise e can be upper bounded by

$$\frac{\left(\sqrt{\sigma_e^2 \log\left(\frac{\det(U^\top U+\lambda I)^{1/2}}{\delta\lambda^{1/2}}\right)}\right)}{\sqrt{\sigma_{\min}(U^\top U+\lambda I)}} \leq \frac{\left(\sqrt{\sigma_e^2 \log\left(\frac{(\beta t)^{1/2}}{\delta\lambda^{1/2}}\right)}\right)}{\sqrt{\alpha t}}$$

which converges to zero as $t \to \infty$ for any m. Also we see that the term $||G||_2/\sqrt{\alpha t}$ must also converge to 0 for any m and hence, in the limit, the minimax optimal order selection procedure will select $m \geq n$ and the error will converge to 0 showing that our criterion coupled with regularized least squares yields a consistent estimator. □

Hence, the above result shows that, when the input is rich enough (which can be achieved even through a very small amount of noise added to the feedback) the order selection criteria yields a consistent estimator. Furthermore, even when $||G - G_m||$ and $b(t,m)$ are not directly available (which is usually the case since we don't know G a priori) but we can upper bound these values by functions $f_1(m,U,Y)$ and $f_2(m,U,Y)$. As long as f_1 and f_2 approach zero for $m \geq n$ and are greater than some constant for $m < n$ we still obtain a consistent estimator with exciting inputs. This can be summarized in as corollary IV.6.1.

Corollary IV.6.1. *Suppose that the inputs to a FIR system satisfy assumption IV.5 and that $||G - G_m|| \leq f_1(m,U,Y)$*

and $b(t,m) \leq f_1(m,U,Y)$. Then as long as for $m < n$, $f_i(m,U,Y) > c_m$ for all t and that for $m \geq n$, $f_i(m,U,Y) \to 0$ for $t \to \infty$, the minimax model selection procedure with regularized least squares is consistent.

B. Extensions to General Noise

Note that in the above formulation, we considered a FIR system with output error e. The structure of this model allowed us to decouple the noise from the input and obtain convergence results. Suppose now that the FIR model is now given as:

$$\begin{aligned} y_t &= \sum_{i=1}^{n} g_i u_{t-i} + \sum_{i=1}^{n} h_i e_{t-i} \\ &= G(q)u_t + H(q)e_t = G(q)u_t + v_t \end{aligned}$$

where e_t is a sequence of additive white Gaussian noise and $H(q)$ is monic and invertible without loss of generality.

Suppose further that there exists a state space representation of the system given by:

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

for some noise γ_t and η_t.

In this case, in closed loop, even if there is a delay between y_{t-1} and u_t we can see that the noise term v_t is no longer independent of u_t and we can no longer apply the machinery of self normalized martingale theory. However, we note that, following the analysis in [18], we can rewrite the system above in predictor form as:

$$y_t = H^{-1}(q)G(q) + (1 - H^{-1}(q))y_t + e_t.$$

We can make several observations about the above representation of the FIR model. First of all, H^{-1} is also a monic polynomial in q and thus the output terms that appear on the right side of the equation are all delayed. Second, it is well known that this description of the system corresponds to the Kalman filter of the state space realization of the original system. Since the matrix A corresponding to the state space dynamics of the original FIR system must be nilpotent, we can see that the coefficients of $1 - H^{-1}$ must be given by

$$\begin{bmatrix} CK & CAK & \dots & CA^{n-1}K & 0 & 0 & \dots \end{bmatrix}$$

where the coefficients vanish past order n. But then the predictor form of the model is now an output error FIR model with which we can apply the results from above. Namely, we can show that, if we consider the least squares problem generated by minimizing

$$\min_{K,R} \sum ||y_t - K^\top [u_{t-1}, ..., u_{t-m}]^\top - R^\top [e_{t-1}, ..., e_{t-m}]^\top||_2 \tag{4}$$

in terms of G and H of unknown dimension as long as we make some assumptions on the structure of the predictor form of the model the minimax optimal model selection criterion yields an accurate estimate for the parameters of the predictor model. We can use the same analysis as above to demonstrate that, since the predictor model is an FIR system as well, that the model selection method is asymptotically consistent. Note however that we get truncation error originating now from both $H^{-1}G$ and from $1 - H^{-1}$ and that there is still truncation bias now also dependent on the tails of $y_{t-m-1}, ..., y_{t-n}$. Finally, it is possible to obtain the parameters of the original system G by simply adding 1 to our estimate of $1 - H^{-1}$, inverting to get H and then multiplying the resulting transfer function by our estimate of $H^{-1}G$ to get an estimate of G. While this procedure does involve inversion, it stands to reason that the best estimate obtained through this procedure results from the best estimate of both $1 - H^{-1}$ and $H^{-1}G$ which can be calculated through our model selection procedure.

EDIT: go into the actual mechanics of long division here and show that first impulse response terms of G only depend on the first m values of estimated impulse response. Stands to reason that the best estimator of the latter will result in the best estimates if the former.

also oyou should write the fact that $1 - H^{-1}$ is a FIR as a lemma (optimal one step ahead predictor is FIR for FIR systems corresponding to the above state space description.)

C. Applications to IIR Models: LTI System Model Selection

The above analysis, particularly as it relates to consistency only applies to FIR systems. However, the methodology developed to identify the optimal model order using closed loop data works equally well with IIR systems as well as long as they satisfy some stability conditions. In particular, consider the stable IIR system given by the input output description

$$y_t = \sum_{i=1}^{\infty} g_i u_{t-i} + e_t$$

where $g_i \leq \alpha\beta^i$ for some fixed $\alpha > 0$ and $\beta \in (0,1)$. Note that this case accounts for most nondegenerate stable discrete time LTI systems with measurement error. Here we see that we can bound the truncation error as before using terms of the form $\alpha\beta^m \frac{1}{1-\beta}$. In which case all of the upper bounding work that was done in example IV.4 applied. We cannot say that our methodology for selecting the model order is consistent however we can connect our estimation of $\hat{m}$ to other model order selection techniques in [9].

Note that $\hat{m}$ in our case was selected to minimize the worst case estimation error possible given a finite noisy data set generated in closed loop and some knowledge about the underlying system. We could then use the estimates of the impulse response parameters along with a subspace based system identification technique, such as the Ho-Kalman algorithm [19] in order to generate a state space realization of the system. The question of what order to choose for the state space realization was addressed in [9] in which the model order again tries to balance the estimation error and truncation error for approximated system Hankel matrices. This amounts to selecting a d such that $||H_{d,d} - \hat{H}_{d,d}||_2 + ||H_{d,d} - H_{\infty,\infty}||_2$ is minimized for system Hankel operators of different sizes. In our case, we estimate $\hat{m}$ parameters using closed loop data, which minimizes the maximum error we might observe through least squares. Moreover, these parameters can be used to create an $(\hat{m}-1)/2 \times (\hat{m}-1)/2$ Hankel matrix $H_{(\hat{m}-1)/2,(\hat{m}-1)/2}$ which would minimize $||H_{(\hat{m}-1)/2,(\hat{m}-1)/2} - \hat{H}_{(\hat{m}-1)/2,(\hat{m}-1)/2}||_F$. Since

$$||H_{(\hat{m}-1)/2,(\hat{m}-1)/2} - \hat{H}_{(\hat{m}-1)/2,(\hat{m}-1)/2}||_2 \leq ||H_{(\hat{m}-1)/2,(\hat{m}-1)/2} - \hat{H}_{(\hat{m}-1)/2,(\hat{m}-1)/2}||_F,$$

using $d = \hat{m} - 1)/2$ could give an approximate solution to the model selection procedure developed in [9] which is the first such result of its kind for data generated in closed loop.

V. Simulations

In order to see how well the minimax model order estimation using closed loop measurements works in practice, we simulated an FIR system given by the transfer function

$$G(z) = \sum_{i=1}^{100} g_i z^{-i}$$

with 100 parameters that each satisfy the relationship that $g_i \leq 1.5 \times .9^i$ so that the parameters decay in a predictable way. With this, we can use the tools developed in example IV.4 to upper bound the estimation error resulting from both truncation bias and noise.

Furthermore, the inputs into the system were selected in closed loop. In particular, they were chosen as noisy outputs from another FIR model $K(z)$ such that $u_t = K(z)y_t + r_t$ for some white noise term r_t. In this case, we would see that there could be significant bias in a least squares estimate originating from correlations between e and u. The data generation process is visually given in figure 1.

To verify the efficacy of our order selection method, we generated data from the aforementioned closed loop system. Using this data, we compared the estimation error generated when using the order recommended by our method coupled with regularized least squares versus simply fixing a model order and calculating the error. It can be seen that, especially when there is limited data, our method significantly outperforms a static strategy. Furthermore, our algorithm in this case doesn't have any prior knowledge of n and simply calculates the minimax order estimation criterion using the singular values of the regularized covariance matrix $U^\top U + \lambda I$ and can thus run regardless of the amount of available data. The results of this experiment can be seen in figure 2.

Also in order to understand how the optimal model order changes with respect to the noise magnitude σ_e^2 we simulated the same system above with varying amounts of measurement noise to see that the number of parameters recommended by the method decreases as noise increases. This agrees with our intuition that by limiting the number of parameters we are limiting the potential of the model to fit to just noise.

VI. Conclusion

In this paper, we developed a new method with which to choose the order of an impulse response model given data collected in closed loop. We showed that, under certain structural assumptions on the true impulse response coefficients g_i, there exists a straightforward way to upper bound

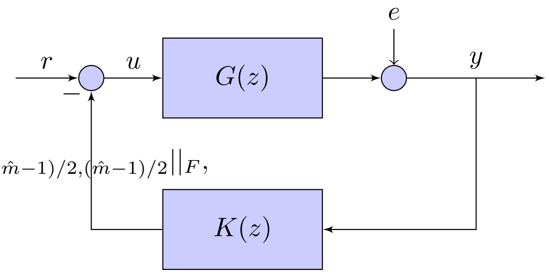

Fig. 1. Data is generated in closed loop with inputs being a combination of the outputs of FIR filter $K(z)$ and random input r, while the process transforming inputs u into outputs y is described by the FIR system $G(z)$. Note that this set up would not normally be amenable to traditional system identification using least squares in finite time as there are significant correlations between errors and noise.

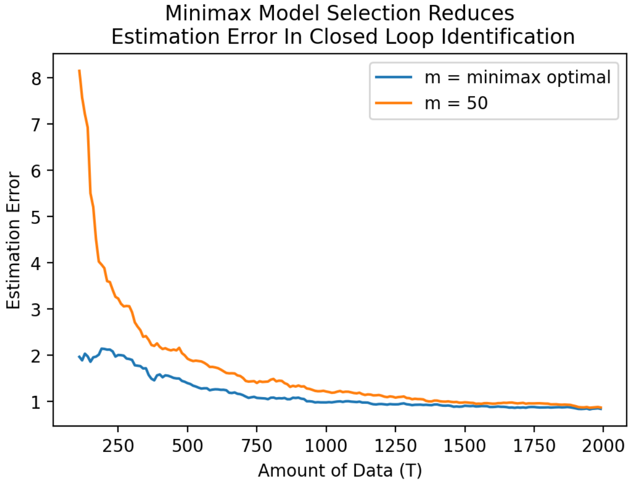

Fig. 2. This figure shows how as more measurement are collected, the minimax optimal model selection procedure consistently chooses model orders which yield less estimation error than a fixed model size (seen in orange.) This indicates that the criteria is serving as a proxy to minimize the worst case estimation error, yielding provably accurate approximations.

the worst case estimation error (hence the term minimax estimation error.) In the limit of infinite data, we show that, using sufficiently informative inputs and our closed loop model selection technique yields a consistent estimator. Furthermore, we extend these notions to the case of more exotic noise structures and to infinite impulse response systems. Finally, simulations show the effectiveness of these methods in using model order to balance estimation errors from both truncation and noise. Future work in this area should focus on developing a generalization of our method for model selection for stable LTI (IIR) models with additive white Gaussian process and measurement noise operating in closed loop.

References

[1] S. Beheshti and M. A. Dahleh, "Noisy data and impulse response estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 510–521, 2009.

[2] B. Lai, S. A. U. Islam, and D. S. Bernstein, "Regularization-induced bias and consistency in recursive least squares," *arXiv preprint arXiv:2106.08799*, 2021.

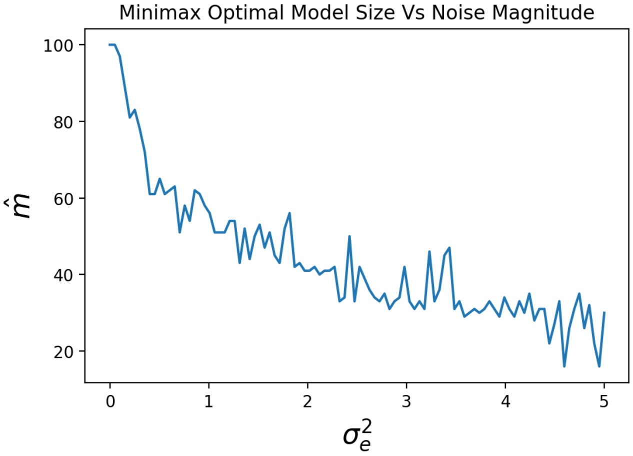

Fig. 3. This figure shows the relationship between an increasing amount of noise in the model and the optimal model order according to the minimax criterion developed in this paper. It can clearly be seen that, as the magnitude of the noise increases, the model order decreases in order to avoid overfitting to noise.

[3] S. Beheshti and M. A. Dahleh, “A new information theoretic approach to order estimation problem,” *IFAC Proceedings Volumes*, vol. 36, no. 16, pp. 765–770, 2003.

[4] ——, “A new minimum description length,” in *Proceedings of the American Control Conference*, vol. 2, 2003, pp. 1602–1607.

[5] ——, “Lti systems, additive noise, and order estimation,” in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 6. IEEE, 2003, pp. 6491–6496.

[6] H. Akaike, “A new look at the statistical model identification,” *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[7] H.-C. So and Y.-T. Chan, “Analysis of an lms algorithm for unbiased impulse response estimation,” *IEEE transactions on signal processing*, vol. 51, no. 7, pp. 2008–2013, 2003.

[8] 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.

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

[10] R. Shibata, “Selection of the order of an autoregressive model by akaike’s information criterion,” *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[11] D. Bauer, “Order estimation for subspace methods,” *Automatica*, vol. 37, no. 10, pp. 1561–1573, 2001.

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

[13] 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

[14] 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.

[15] 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.

[16] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, “Regret bounds for robust adaptive control of the linear quadratic regulator,” *arXiv preprint arXiv:1805.09388*, 2018.

[17] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, “Finite-time system identification and adaptive control in autoregressive exogenous systems,” in *Learning for Dynamics and Control*. PMLR, 2021, pp. 967–979.

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

[19] B. Ho and R. E. Kálmán, “Effective construction of linear state-variable models from input/output functions,” *at-Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.