Proceedings of Machine Learning Research vol 117:1–28, 2020 31st International Conference on Algorithmic Learning Theory

Robust Guarantees for Learning an Autoregressive Filter

Holden Lee HOLDEN.LEE@DUKE.EDU
Duke University

Cyril Zhang CYRIL.ZHANG@PRINCETON.EDU
Princeton University

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

The optimal predictor for a known linear dynamical system (with hidden state and Gaussian noise) takes the form of an autoregressive linear filter, namely the Kalman filter. However, making optimal predictions in an unknown linear dynamical system is a more challenging problem that is fundamental to control theory and reinforcement learning. To this end, we take the approach of directly learning an autoregressive filter for time-series prediction under unknown dynamics. Our analysis differs from previous statistical analyses in that we regress not only on the inputs to the dynamical system, but also the outputs, which is essential to dealing with process noise. The main challenge is to estimate the filter under worst case input (in $\mathcal{H}_\infty$ norm), for which we use an L^∞-based objective rather than ordinary least-squares. For learning an autoregressive model, our algorithm has optimal sample complexity in terms of the rollout length, which does not seem to be attained by naive least-squares.

Keywords: control theory, time series, linear dynamical systems, autoregressive models, optimal filtering

1. Introduction

The problem of estimating the hidden state and outputs of a known linear dynamical system (LDS), given the inputs and observations, is a well-studied problem in control theory (Kamen and Su, 1999). When the process and observation noise are independent and mean-zero with known covariances, this problem is solved by the Kalman filter (Kalman, 1960; Anderson and Moore, 2012), which recursively propagates the optimal linear estimator for the hidden state. When the recursion for the estimator is unrolled, the Kalman filter is seen to be a linear autoregressive filter: it predicts the system's next output as a linear combination of the system's past ground-truth outputs.

However, when the LDS is *unknown*, optimal filtering is a much harder problem. Moreover, in the absence of model knowledge, learning to filter a LDS is generally required for controlling a LDS, a foundational problem in machine learning and control theory. One widely-used approach is to learn the dynamical matrices from data, after which one can simply apply the Kalman filter. Unfortunately, this approach runs into computational barriers: the usual formulation of this problem is nonconvex. System identification techniques provide various practical algorithms for this problem (Ljung, 1998). However, these algorithms, such as EM (Roweis and Ghahramani, 1999), lack rigorous end-to-end guarantees, and are often unstable or find suboptimal solutions in high dimensions.

In this work, we bypass the state-space representation of an LDS, and analyze the statistical guarantees of learning an autoregressive filter directly. This allows us to compete with the predictions

of the steady-state Kalman filter, without the computationally intractable task of explicitly identifying the system. We present a polynomial-time algorithm for learning an autoregressive filter for time-series prediction. The predictor has robust ($\mathcal{H}_\infty$) learning guarantees, which do not seem to be attained by naive least-squares.

1.1. Background

Our primary motivation is the following question: *can we learn to predict the observations as well as the Kalman filter (in $\mathcal{H}_\infty$ norm) without learning the system?* We consider the setting of a linear dynamical system with hidden state, defined by[1]

$$h(t) = Ah(t-1) + Bx(t-1) + \xi(t) \tag{1.1}$$
$$y(t) = Ch(t) + \eta(t), \tag{1.2}$$

where $x(t) \in \mathbb{R}^m$ are inputs, $h(t) \in \mathbb{R}^d$ are hidden states, $y(t) \in \mathbb{R}^n$ are outputs, $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times m}$, $C \in \mathbb{R}^{n\times d}$, and $\xi(t) \in \mathbb{R}^d$ and $\eta(t) \in \mathbb{R}^n$ are independent zero-mean noise (we consider the case when they are Gaussian). Crucially, only the $y(t)$, and not the $h(t)$, are observed. A major difficulty of learning the dynamics from data comes from the fact that the objective function $||y(t) - CA^t h(0) - \sum_{i=0}^{t-1} CA^i Bx(t-1-i)||^2$ is nonconvex in A, B, C. A classic approach is *subspace identification* (Ho and Kalman, 1966; Van Overschee and De Moor, 2012), for which statistical guarantees only exist in the asymptotic regime or under stringent assumptions. In the presence of noise, these methods are often used to initialize the EM algorithm (Roweis and Ghahramani, 1999), a classic heuristic for a non-convex objective.

Another classical model for dynamical systems is the autoregressive-moving average (ARMA) model (Hamilton, 1994; Box et al., 1994; Brockwell and Davis, 2009), which models latent perturbations using a *moving average* process. A central technique here is to recover an ARMA model by solving the Yule-Walker equations. However, to our knowledge, existing work on provably learning these models is limited to asymptotic guarantees.

1.2. Our results

We show that under certain stability conditions of the Kalman filter, we can bypass proper identification of the system, and still converge to the performance of the Kalman filter. We take an improper learning approach, reducing this problem to the general problem of *learning an autoregressive model*.

Our algorithm is based off a simple and familiar algorithm in time series analysis: using a sine-wave input design to fit an autoregressive model using least-squares. However, a key problem with the ordinary least-squares approach is that it does not provide learning guarantees under worst-case input (that we have not necessarily seen), i.e., in the $\mathcal{H}_\infty$ norm. Such worst-case bounds are important because in the usual control-theoretic framework, bounds under the $\mathcal{H}_\infty$ norm are used to obtain guarantees for robust control.

To obtain $\mathcal{H}_\infty$ bounds for learning an autoregressive model, we augment our algorithm with a L^∞ objective to learn a predictor that is robust in the $\mathcal{H}_\infty$ sense. When applied to the Kalman filter, our work gives (to our knowledge) the first non-asymptotic sample complexity guarantees for learning an optimal autoregressive filter for estimation in a LDS.

1. Note that $h(t)$ and $x(t)$ are often denoted by x_t and u_t respectively in the control theory literature; in this paper we follow the machine learning convention of using h to denote hidden state.

1.3. Related work

LDS without hidden state, and FIRs. The problem of learning unknown dynamical systems has attracted a lot of recent attention from the machine learning community, due to connections to reinforcement learning and recurrent neural networks. Much progress has been made on the simpler related problem of learning and control in a linear dynamical model with no hidden state. Such a model is defined by

$$h(t) = Ah(t-1) + Bx(t-1) + \xi(t), \tag{1.3}$$

where A, B, $x(t)$, $\xi(t)$ are as before, but $h(t)$ is now observed. Dean et al. (2017) consider the linear quadratic regulator (LQR)—the control problem for such a LDS—and prove that the least-squares estimator of the dynamics, given independent rollouts, is sample-efficient for this setting. Simchowitz et al. (2018) show that access to independent rollouts is unnecessary; the LDS can be identified with a single rollout, even when the system is only marginally stable. Sarkar and Rakhlin (2019) improve the analysis and extend it to explosive A.

An alternative approach to identifying A and B is to learn the system as a finite-impulse response (FIR) filter. This is because the problem of learning a FIR filter can be thought of as a relaxation of the problem of learning a LDS, by "unrolling" the LDS. Tu et al. (2017) use ordinary least-squares with design inputs to learn a FIR, and give near-optimal sample complexity bounds. Boczar et al. (2018) complete the "identify-then-control" pipeline by studying robust control for this estimated FIR filter.

Limitations of FIRs. FIR filters are an insufficiently expressive class of models for prediction in linear dynamical systems with a latent state $h(t)$. Firstly, there are unstable or marginally stable systems which can be written as autoregressive models with short filter length, but the infinite impulse response filter is not approximated by any finite truncation. Moreover, performance guarantees for prediction using FIR filters are given under observation noise, and become very poor under process noise. In these cases, the statistical guarantees of prediction using FIR filters can be suboptimal by an arbitrarily large factor, while an autoregressive model can make statistically optimal predictions. To illustrate this dramatic gap, we analyze a simple example in Section 2.2. Our approach fills a gap in the literature, by giving statistical guarantees similar to those obtained by Tu et al. (2017) for the more expressive and useful family of autoregressive models.

LDS with hidden state, and autoregressive models. In the setting of linear dynamical systems with a hidden state, several recent works analyze settings in which the dynamics can be identified. Hardt et al. (2016) show that under certain conditions on the characteristic polynomial of the system's transition matrix, gradient descent learns the parameters of a single-input single-output LDS. However, they only consider the setting of observation noise, and not process noise (i.e. $\xi(t) = 0$). In work concurrent to ours, Simchowitz et al. (2019), building on Oymak and Ozay (2018), consider the problem of learning an autoregressive filter, and for the case of a LDS, are able to recover matrices $\overline{A}$, $\overline{B}$, $\overline{C}$ which give an *equivalent realization* of the LDS. Although they allow for semi-parametric noise and marginally stable systems, their guarantees are for estimating the matrices in operator norm, rather than the system in the more stringent $\mathcal{H}_\infty$ norm. Tsiamis and Pappas (2019) give guarantees for a subspace identification algorithm to estimate the Kalman gain in operator norm.

By unrolling the Kalman filter (see Section 2.3), the problem of learning the Kalman filter can be recast as learning an autoregressive model. Several works have addressed learning an ARMA model

in the online learning (regret minimization) setting. We note however that the regret framework is different than what is required for control, as it ensures performance only on the data that is seen; the predictor is not required to perform well on worst-case input. Anava et al. (2013) give an algorithm that works in the presence of adversarial (as opposed to i.i.d. Gaussian) noise. However, the constraint on the ℓ_1-norm of the coefficients of the error terms, which they require for the dynamical stability of their estimator of residuals, is very stringent. Kozdoba et al. (2019) use online gradient descent to learn an autoregressive model. They show that whenever the original LDS is observable, the Kalman filter is strictly stable, and hence only a finite horizon is necessary. They do not require the original LDS to be stable. However, their regret bounds scale as the size of the outputs, which can potentially grow in time when the system is not strictly stable.

Finally, we note the approach of online spectral filtering for prediction in symmetric and asymmetric LDS's (Hazan et al., 2017, 2018). In these works, the process noise is only handled up to a multiplicative factor of the optimal filter with knowledge of the system. Intuitively, this "competitive ratio bound" arises because these works consider regressing only on one or a few past observations y_t (in a somewhat rigid manner), rather than having the freedom to imitate an optimal autoregressive filter.

2. Problem setting and preliminaries

We first state the general problem of learning an autoregressive model, and then in Section 2.3 describe the connection to linear dynamical systems. In Section 2.4 we introduce some concepts from control theory and use it to write error bounds in terms of control-theoretic norms of filters (Lemma 2.4).

2.1. Problem statement

A (single-input, single-output) *dynamical system* converts input signals $x(0), \dots, x(T-1) \in \mathbb{R}$ into output signals (random variables) $y(1), \dots, y(T) \in \mathbb{R}$. We will assume that the data are generated by an autoregressive model:

$$y(t+1) = g^* * x(t) + h^* * y(t) + \eta(t+1) = \sum_{k=0}^{\infty} g^*(k)x(t-k) + \sum_{k=0}^{\infty} h^*(k)y(t-k) + \eta(t+1), \tag{2.1}$$

where $\eta(t) \sim N(0, \sigma^2)$ is a time series of i.i.d. Gaussian noise, g, h, are supported on $\mathbb{N}_0$, and $x(t) = 0$ for $t < 0$ and $y(t) = 0$ for $t \leq 0$.

Problem 2.1 *Let $g^*, h^* \in \mathbb{R}^{\mathbb{N}_0}$ be filters. The learner is given black-box access to the system $\mathscr{L}$ which takes inputs $x \in \mathbb{R}^{\mathbb{N}_0}$ to outputs $y \in \mathbb{R}^{\mathbb{N}}$ by (2.1). During each rollout, the learner specifies an input design $\{x(0), \dots, x(T-1)\}$, and receives the corresponding output sequence. After collecting outputs from s rollouts, the learner returns filters g, h, which specify a map from input to output signals via (2.1).*

For an estimate g, h of g^*, h^*, define the error in the prediction (compared to the expected value of $y(t+1)$) to be

$$y_{\text{err}}(t+1) = (g - g^*) * x(t) + (h - h^*) * y(t). \tag{2.2}$$

The goal is to learn g, h such that the expected error in the prediction is a small fraction ε_1 of the input, plus a small fraction ε_2 of the elapsed time:

$$\mathbb{E}\left[\sum_{t=1}^{T} \|y_{\text{err}}(t)\|^2\right] \le \varepsilon_1 \sum_{t=1}^{T} \|x(t)\|^2 + \varepsilon_2 T. \tag{2.3}$$

2.2. Inadequacy of learning an FIR filter

We complete the discussion in Section 1.3, exhibiting a minimal example where the gap between FIR and autoregressive models can be made arbitrarily large. Consider the system

$$\begin{aligned} h(t) &= rh(t-1) + x(t-1) + \xi(t), \\ y(t) &= h(t) + \eta(t) \end{aligned}$$

where $0 < r < 1$ and $\xi(t), \eta(t) \sim N(0,1)$. Then we can calculate using formulas for the Kalman filter that the variance in the estimation of h and y are $\sigma_h^2 = \frac{r^2+\sqrt{r^4+4}}{2}$, and $\sigma_y^2 = \sigma_h^2 + 1$. The mean squared error in estimating y_t using the Kalman filter is σ_y^2, which remains finite as $r \to 1$. On the other hand, if we were to estimate y_t without using the previous observations $y_{t-1}, \ldots$, then the average estimation error is $1 + (1 + r^2 + r^4 + \cdots) = 1 + \frac{1}{1-r^2}$, which blows up as $r \to 1$. Hence the multiplicative factor between the error using a FIR filter, and using the optimal filter, goes to ∞ as $r \to 1$. In general, FIR methods suffer a multiplicative factor depending on $\|G\|_\infty$ (Tu et al., 2017, §3 (Process noise)), which is $\frac{1}{1-r^2}$ in this example.

2.3. Connection to the Kalman filter

Our work is motivated by optimal state estimation in LDS's with hidden state given by the dynamics (1.1)–(1.2). The Kalman filter gives the optimal linear estimator in the case that the parameters and noise covariances of the LDS are known and $h(0)$ is drawn from a distribution with known mean $h^-(0)$ and covariance. We can compute matrices $A_{KF}^{(t)}$, $B_{KF}^{(t)}$, and $C_{KF}^{(t)}$ such that the optimal linear estimate of the latent state $\widehat{h}(t)$ and the observation $\widehat{y}(t)$ are given by a time-varying LDS (taking the $y(t)$ as feedback) with those matrices:

$$h^-(t) = A_{KF}^{(t)} h^-(t-1) + B_{KF}^{(t)} \begin{pmatrix} x(t-1) \\ y(t-1) \end{pmatrix} \tag{2.4}$$

$$\widehat{y}(t) = C_{KF}^{(t)} h^-(t). \tag{2.5}$$

In the case where the initial and noise distributions are Gaussian, $h^-(t)$ and $\widehat{y}(t)$ are furthermore the maximum a posteriori estimators, and the actual hidden state $h(t)$ and the observation $y(t)$ are Gaussians when conditioned on $\mathcal{F}_{t-1}$ (the observations up to time $t-1$): $h(t)|\mathcal{F}_{t-1} \sim N(h^-(t), \Sigma_h^{(t)})$ and $y(t)|\mathcal{F}_{t-1} \sim N(\widehat{y}(t), \Sigma_y^{(t)})$ for some covariance matrices $\Sigma_h^{(t)}$, $\Sigma_y^{(t)}$.

If the original system is observable and the noise is iid, taking $t \to \infty$, the matrices $A_{KF}^{(t)}$, $B_{KF}^{(t)}$, and $C_{KF}^{(t)}$ approach certain fixed matrices A_{KF}, B_{KF}, and C_{KF}, and the covariance matrices $\Sigma_h^{(t)}$ and $\Sigma_y^{(t)}$ approach fixed matrices Σ_h and Σ_y (Harrison, 1997). Our goal is to learn this *steady-state*

Kalman filter without knowing parameters of the original LDS.[2] In the Gaussian case, at steady-state, the actual hidden state $h(t)$ and observation $y(t)$ will be distributed as $h(t)|\mathcal{F}_{t-1} \sim N(h^-(t), \Sigma_h)$ and $y(t)|\mathcal{F}_{t-1} \sim N(\widehat{y}(t), \Sigma_y)$.

From now on, we will assume that the initial and noise distributions are fixed Gaussians, as our results will compare against the steady-state Kalman filter and rely on the fact that $y(t)|\mathcal{F}_{t-1}$ is a mean-zero random variable. This may not be true for general noise distributions.

Denote $B_{KF} = (B_{KF,x}\ B_{KF,y})$, where $B_{KF,x}$ and $B_{KF,y}$ are the submatrices acting on $x(t)$ and $y(t)$, respectively. Consider for simplicity the case where the input and output dimensions are 1: if the hidden state has dimension d, then $A_{KF} \in \mathbb{R}^{d\times d}$, $B_{KF,x}, B_{KF,y} \in \mathbb{R}^{d\times 1}$, $C_{KF} \in \mathbb{R}^{1\times d}$, and we simply have $\Sigma_h = \sigma_h^2$ for some σ_h. We can then "unfold" the Kalman filter into an equivalent autoregressive model (2.1) by letting $g^*(t) = C_{KF}A_{KF}^t B_{KF,x}$ and $h^*(t) = C_{KF}A_{KF}^t B_{KF,y}$, and $\eta(t) \sim N(0, \sigma_h^2)$.[3] Note that the autoregressive model captures the law of the random process defined by the LDS (under what is observable at each time step, i.e., the filtration $\mathcal{F}_t$), without utilizing a hidden state.

In this setting, we again attempt to minimize the error between the prediction and the expected value when the dynamics are known, $y_{\text{err}}(t) = \widehat{y}(t) - \mathbb{E}[y(t)|\mathcal{F}_{t-1}]$.

2.4. Preliminaries on control theory

An impulse response function can be equivalently be represented as a power series.

Definition 2.2 *For a sequence $f \in \mathbb{R}^{\mathbb{Z}}$ define the transfer function of f by $F(z) = \sum_{k\in\mathbb{Z}} f(k)z^{-k}$. We will always denote the transfer function of a sequence in $\mathbb{R}^{\mathbb{Z}}$ by the corresponding capital letter.*

Note that if $y = f * x$, then as formal power series, $Y = FX$, and equality holds as functions for z such that $F(z)$, $X(z)$ converge absolutely. Translation corresponds to multiplication: the transfer function of $t \mapsto y(t+1)$ is $zY(z)$. Hence, letting N be the transfer function of η, we have from (2.1) that

$$zY = G^*X + H^*Y + zN \tag{2.6}$$

$$\implies (1 - z^{-1}H^*)Y = z^{-1}G^*X + N \tag{2.7}$$

$$Y = z^{-1}G^*H^*_{\text{unr}}X + H^*_{\text{unr}}N \tag{2.8}$$

$$\text{where } H^*_{\text{unr}}(z) := \frac{1}{1 - z^{-1}H^*(z)}. \tag{2.9}$$

Thus, we can rewrite (2.1) as

$$y(t+1) = h^*_{\text{unr}} * g^* * x(t) + h^*_{\text{unr}} * \eta(t+1), \tag{2.10}$$

where $h^*_{\text{unr}}(k)$, the "unrolled" filter, is such that $\sum_{k=0}^{\infty} h^*_{\text{unr}}(k)z^{-k} = H^*_{\text{unr}}(z)$.

2. Note that if the parameters of the LDS are unknown, then any A, B, C for which the law of the y_t in (1.1)–(1.2) is the same as the law of the actual y_t is an equivalent realization. Then the Kalman filters computed from these A, B, C will all give equivalent predictions, so we need not distinguish between them.
3. Note this is not to be confused with the $\eta(t)$ in (1.1)–(1.2): this $\eta(t)$ has larger variance because it also incorporates the uncertainty about the hidden state.

Definition 2.3 *The $\mathcal{H}_\infty$-norm of a filter is the L^∞-norm of the transfer function over the unit circle $\|z\|_2 = 1$:*

$$\|f\|_{\mathcal{H}_\infty} = \|F\|_\infty := \max_{\|z\|_2=1} F(z). \tag{2.11}$$

The $\mathcal{H}_2$-norm of a filter is the L^2-norm of the transfer function over the unit circle:

$$\|f\|_{\mathcal{H}_2} = \|F\|_2 := \left(\frac{1}{2\pi}\int_{|z|=1} |F(z)|^2\, dz\right)^{\frac{1}{2}}. \tag{2.12}$$

For the rest of the paper we will assume the system is stable, i.e., $\|H^*\|_\infty < 1$, so that $\|H^*_{\text{unr}}\| < \infty$.[4]

The $\mathcal{H}_2$-norm represents the steady state variance under iid Gaussian noise as input, and the $\mathcal{H}_\infty$-norm represents the maximum norm of the output when $\|x\|_2 = 1$:

$$\|f\|^2_{\mathcal{H}_2} = \mathbb{E}_{\forall s,\eta(s)\sim N(0,1)} |(f*\eta)(t)|^2 \tag{2.13}$$

$$\|f\|_{\mathcal{H}_\infty} = \sup_{\|x\|_2=1} \|f*x\|_2. \tag{2.14}$$

From (2.2) and (2.10),

$$y_{\text{err}}(t+1) = (g-g^*)*x(t) + (h-h^*)*(h^*_{\text{unr}}*g^**x)(t-1) + (h-h^*)*(h^*_{\text{unr}}*\eta)(t) \tag{2.15}$$

$$= [(g-g^*) + \delta_1 * (h-h^*)*h^*_{\text{unr}}*g^*]*x(t) + [(h-h^*)*h^*_{\text{unr}}]*\eta(t) \tag{2.16}$$

where $\delta_i(j) = \mathbb{1}_{i=j}$. Because η has mean 0,

$$\mathbb{E}_\eta\left[\sum_{t=1}^T \|y_{\text{err}}(t)\|^2\right]$$
$$= \mathbb{E}_\eta\left[\sum_{t=1}^T \|[(g-g^*) + \delta_1*(h-h^*)*h^*_{\text{unr}}*g^*]*x(t)\|^2\right] + \mathbb{E}_\eta\left[\sum_{t=1}^T \|[(h-h^*)*h^*_{\text{unr}}]*\eta(t)\|_2^2\right].$$

Hence from (2.13) and (2.14) we obtain the following, noting that the noise in Problem 2.1 is $N(0,\sigma^2)$.

Lemma 2.4 *Suppose that $\|H^*\|_\infty < 1$. Then in the setting of Problem 2.1,*

$$\mathbb{E}\left[\sum_{t=1}^T \|y_{\text{err}}(t)\|^2\right] \le \left\|(G-G^*) + z^{-1}(H-H^*)H^*_{\text{unr}}G^*\right\|_\infty^2 \|x\|^2 + \|(H-H^*)H^*_{\text{unr}}\|_2^2\, \sigma^2 T$$

We will approximate g^*, h^* with finite-length filters of length r, so we need to make sure r is large enough to capture most of the response. For this, we use the following definition and lemma from Tu et al. (2017) which gives a sufficient length in terms of the desired error and a $\mathcal{H}_\infty$ norm.

4. The $\|H^*\|_\infty < 1$ condition is necessary to do worst-case ($\mathcal{H}_\infty$-norm) estimation over a infinite time horizon, with only access to a finite rollout. This is because an input with infinite response can have arbitrarily small response over a finite horizon. This suggests that to solve the control problem over infinite time horizon of a non-stable system, one should look for weaker assumptions than learning in $\mathcal{H}_\infty$-error that still allow control.

Definition 2.5 (Sufficient length condition, (Tu et al., 2017, Definition 1)) *We say that a Laurent series F has stability radius $\rho \in (0,1)$ if F converges for $\{x \in \mathbb{C} : |x| > \rho\}$. Let F be stable with stability radius $\rho \in (0,1)$. Fix $\varepsilon > 0$. Define*

$$R(\varepsilon) = \left\lceil \inf_{\rho<\gamma<1} \frac{1}{1-\gamma} \ln\left(\frac{\|F(\gamma z)\|_\infty}{\varepsilon(1-\gamma)}\right)\right\rceil. \tag{2.17}$$

Note that this "sufficient length condition" is analogous to having a $\frac{1}{1-\rho(A)}$ dependence on the spectral norm of A, for learning a LDS. Indeed, a filter corresponding to a LDS will have stability radius $\rho(A)$.

Lemma 2.6 ((Tu et al., 2017, Lemma 4.1)) *Suppose F is stable with stability radius $\rho \in (0,1)$. Then $\|f_{\geq L}\|_1 := \sum_{k\geq L} |f(k)| \leq \max_{\rho<\gamma<1} \frac{\|F(\gamma z)\|_\infty \gamma^L}{1-\gamma}$. Hence, if $L \geq R(\varepsilon)$, then $\|f_{\geq L}\|_1 \leq \varepsilon$.*

3. Algorithm and main theorem

We motivate our main algorithm, Algorithm 1. The most natural algorithm is the following: let the inputs be sinusoids at equally spaced frequencies, and solve a least-squares problem for g, h. However, ordinary least-squares will only give g, h for which the estimation error is small for *random* input, while we desire g, h for which the estimation error is small for *worst-case* input; in other words, it gives an average-case ($\mathcal{H}_2$), rather than the worst-case ($\mathcal{H}_\infty$) bound that we desire. This is analogous to the difference between estimating a $r \times r$ matrix in Frobenius and operator norm; the Frobenius norm trivially bounds the operator norm, but the resulting bound is typically $\sqrt{r}$ from optimal. Hence, the sample complexity bound from ordinary least-squares does not have optimal dependence on r. Note that Boczar et al. (2018) solve the analogous problem for a FIR filter f^* with least-squares without suffering an extra $\sqrt{r}$ factor, because in that setting, the matrix M in the least-squares problem is a fixed matrix depending on the inputs, the error $f - f^*$ in the estimate is gaussian, and supremum bounds for Gaussians are applicable. Our setting is more challenging because the M's depend on noise in observations $y(t)$ that we have no control over.

The first step of our algorithm is still to solve a least-squares problem. We do this in two parts: first, solve for h_{LS} by regressing on zero input, and then using h_{LS}, solve for $g_{LS}^{(j)}$ separately for each frequency j. We do this to avoid the error in h_{LS}—larger by a factor $\sqrt{r}$ because it is r-dimensional—contributing to the error in the $g_{LS}^{(j)}$.

The final step is to combine the $g_{LS}^{(j)}$. Because the number of frequencies is larger than the length r of the filter (necessary to be able to interpolate to unseen frequencies), we cannot find a single g that matches each $g_{LS}^{(j)}$ on the jth frequency. Keeping in mind our $\mathcal{H}_\infty$ objective, we hence optimize a L^∞ problem over the frequencies to interpolate the $g_{LS}^{(j)}$.

Note that in our algorithm we can just take just $0 < j < \frac{cr}{2}$ for the $\sin$ signals because the signals for $j = 0, \frac{cr}{2}$ are trivial; we consider $0 \leq j \leq \frac{cr}{2}$ to make the notation in the proof simpler. For convenience of notation we re-index the time series to start at $t = -L$.

Our algorithm differs from the one in Simchowitz et al. (2019) in that their algorithm first does a pre-filtering step (ridge regression) on past *outputs*, and then does a linear regression on just the previous *inputs*, while our algorithm regresses on the previous inputs and observations together. Moreover, we use designed inputs in order to ensure estimation in $\mathcal{H}_\infty$ norm.

Algorithm 1 Learning an autoregressive model

1: INPUT: burn-in time L, number of rollouts of each frequency ℓ, filter length r, $c > 4\pi$.

2: Collect length $T = cr$ rollouts of the $\sim 2c\ell r$ input signals starting at $t = -L$,

$$x^{(\bullet,k)} = x^{(\bullet)} \equiv 0, \qquad 1 \le k \le c\ell r \tag{3.1}$$

$$x_{\cos}^{(j,k)}(t) = x_{\cos}^{(j)}(t) = \cos\left(\frac{2\pi jt}{cr}\right) \qquad 0 \le j \le \frac{cr}{2} \qquad 1 \le k \le \ell, \tag{3.2}$$

$$x_{\sin}^{(j,k)}(t) = x_{\sin}^{(j)}(t) = \sin\left(\frac{2\pi jt}{cr}\right) \qquad 0 \le j \le \frac{cr}{2} \qquad 1 \le k \le \ell. \tag{3.3}$$

Let the outputs be $y^{(\bullet,k)}$, $y_{\cos}^{(j,k)}$, and $y_{\sin}^{(j,k)}$. Let $M^{(\bullet,k)} \in \mathbb{R}^{r\times T}$, $M_{\cos,t}^{(j,k)} \in \mathbb{R}^{2r\times T}$, and $M_{\sin,t}^{(j,k)} \in \mathbb{R}^{2r\times T}$ be the matrices with columns (for $1 \le t \le T$)

$$M_t^{(\bullet,k)} = y^{(\bullet,k)}(t-1:t-r) \quad M_{\cos,t}^{(j,k)} = \begin{pmatrix} x_{\cos}^{(j)}(t-1:t-r) \\ y_{\cos}^{(j,k)}(t-1:t-r) \end{pmatrix} \quad M_{\sin,t}^{(j,k)} = \begin{pmatrix} x_{\sin}^{(j)}(t-1:t-r) \\ y_{\sin}^{(j,k)}(t-1:t-r) \end{pmatrix} \tag{3.4}$$

where $x(t-1:t-r)$ denotes $(x(t-1),\dots,x(t-r))^\top$.

3: Solve the following least-squares problem under zero noise. Here, $y^{(\bullet,k)}$ refers to the vector $y^{(\bullet,k)}(1:T)$.

$$h_{LS} = \mathrm{argmin}_h \sum_{k=1}^{c\ell r} \left\| M^{(\bullet,k)\top} h - y^{(\bullet,k)} \right\|^2. \tag{3.5}$$

4: Solve the following least-squares problems, for $0 \le j \le \frac{cr}{2}$:

$$g_{LS}^{(j)} = \mathrm{argmin}_g \sum_{k=1}^{\ell} \left[\left\| M_{\cos}^{(j,k)\top} \begin{pmatrix} g \\ h_{LS} \end{pmatrix} - y_{\cos}^{(j,k)} \right\|^2 + \left\| M_{\sin}^{(j,k)\top} \begin{pmatrix} g \\ h_{LS} \end{pmatrix} - y_{\sin}^{(j,k)} \right\|^2 \right] \tag{3.6}$$

5: Solve and return

$$\begin{pmatrix} g \\ h \end{pmatrix} = \mathrm{argmin}_{g,h} \max \Bigg\{ \frac{1}{r} \sum_{k=1}^{c\ell r} \left\| M^{(\bullet,k)\top} (h - h_{LS}) \right\|^2, \tag{3.7}$$

$$\max_j \sum_{k=1}^{\ell} \left[\left\| M_{\cos}^{(j,k)} \left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} \right] \right\|^2 + \left\| M_{\sin}^{(j,k)} \left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} \right] \right\|^2 \right] \Bigg\}.$$

Theorem 3.1 (Learning an autoregressive model) *There is C, C' such that the following holds. In the setting of Problem 2.1, suppose that $\|G^*\|_\infty < \infty$, $\|H^*\|_\infty < 1$, and Algorithm 1 is run with $c \ge 8\pi$, burn-in time $L \ge \max\left\{ R_{H^*_{\mathrm{unr}}}\left(\frac{\delta}{4KT\sqrt{c\ell r}}\right), R_{H^*_{\mathrm{unr}}G^*}\left(\frac{\delta}{4K\sqrt{c\ell rT}}\right)\right\}$ where $K =$*

$\left(1 + \sum_{t=0}^{T-2} |h^*(t)|\right)^2$, *rollout length* T, *and* $\ell \geq C'^2(r + \ln\left(\frac{1}{\delta}\right))$ *rollouts of each input. Let*

$$\varepsilon_1 := \frac{C}{\sqrt{c\ell T}} \left(\ln\left(\frac{c\ell r T}{\delta}\right)\right)^{\frac{3}{2}} (1 + \|H^*\|_\infty) \|H^*_{\text{unr}}\|_\infty, \quad \varepsilon_2 := \frac{C}{\sqrt{c\ell T}} \left(\ln\left(\frac{c\ell r T}{\delta}\right)\right)^2. \quad (3.8)$$

Then with probability $1 - \delta$*, the algorithm returns* g, h *such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \|y_{\text{err}}(t)\|^2\right] \leq \varepsilon_1^2 \|x\|_2^2 + \varepsilon_2^2 T. \quad (3.9)$$

To prove the theorem, we establish the bounds

$$\left\|(G - G^*) + z^{-1}(H - H^*)H^*_{\text{unr}}G^*\right\|_\infty \leq \varepsilon_1 \qquad \|(H - H^*)H^*_{\text{unr}}\|_2 \leq \sigma^{-1}\varepsilon_2 \quad (3.10)$$

and use Lemma 2.4. Note there is no dependence on σ in (3.9) for the following reason: smaller σ means worse estimation of $\|(H - H^*)H^*_{\text{unr}}\|_2$ (the response to $N(0,1)$ noise) by a factor of σ^{-1}, but when tested on rollouts with noise $N(0, \sigma^2)$, the error is not affected.

We expect the $O\left(\frac{1}{\sqrt{\ell T}}\right)$ dependence on ℓ, T, r to be optimal: there are $O(r)$ parameters, and we have access to $O(\ell T r)$ samples (including samples in the same rollout). We also conjecture that the $\|H^*_{\text{unr}}\|_\infty$ dependence is unavoidable.

As an immediate corollary, we obtain a theorem for learning the Kalman filter. For simplicity, we state the result when $h(0)$ has the steady-state distribution, to avoid burn-in time arguments.

Corollary 3.2 (Improperly learning the Kalman filter) *Consider the system (1.1)–(1.2). Let* A_{KF}*,* $B_{KF,x}$*,* $B_{KF,y}$*,* C_{KF} *be the steady-state Kalman filter matrices and* σ_y^2 *be the variance in the estimate of* y*, as defined in Section 2.3. Let* $G^*(z) = \sum_{t=0}^{\infty} C_{KF} A_{KF}^t B_{KF,x} z^{-t}$ *and* $H^*(z) = \sum_{t=0}^{\infty} C_{KF} A_{KF}^t B_{KF,y} z^{-t}$*. Suppose that* A_{KF} *has spectral radius* < 1*, and suppose the rollouts are started with* $h(0) \sim N(0, \sigma_h^2)$*. Algorithm 1 with parameters given in Theorem 3.1 returns predictions such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \|y_{\text{err}}(t)\|^2\right] \leq \varepsilon_1^2 \|x\|_2^2 + \varepsilon_2^2 T. \quad (3.11)$$

4. Proof sketch

It will be convenient to first prove the theorem in the case when the burn-in time is infinite. Note that by the stability assumption on H^*, for signals with finite $\|x\|_\infty$, the outputs will not diverge.

Theorem 4.1 *Theorem 3.1 holds in the setting when the burn-in time* L *is infinite.*

We break the proof of Theorem 3.1 into 4 parts. The first 3 parts will prove Theorem 4.1. The full proof is in Section A.

Step 1 (Concentration): If $\overline{y} = M^\top x$ and $y = \overline{y} + \eta$, then the error from the least-squares problem $\operatorname{argmin}_x \left\|M^\top x - y\right\|^2$ is $(MM^\top)^{-1}M\eta$. A simple way to bound this is to bound $MM^\top$ from below and $M\eta$ from above. When we have s samples, and $x \in \mathbb{R}^r$, we expect $\left\|(MM^\top)^{-1}\right\| \leq O\left(\frac{1}{s}\right)$ and $\|M\eta\| \leq O\left(\sqrt{rs}\right)$.

We show that the matrices $Q^{(\bullet)} := \sum_{k=1}^{c\ell r} M^{(\bullet,k)} M^{(\bullet,k)\top}$ and $Q^{(j)} := \sum_{k=1}^{\ell} (M_{\cos}^{(j,k)} M_{\cos}^{(j,k)\top} + M_{\sin}^{(j,k)} M_{\sin}^{(j,k)\top})$ in the least-squares problem (3.5) and (3.6) concentrate using matrix concentration bounds (Lemma A.2), and that the terms such as $\sum_{k=1}^{c\ell r} M^{(\bullet,k)}\eta^{(\bullet,k)}$ concentrate by martingale concentration (Lemma A.5). The main complication is to track how the error $h_{LS} - h^*$ propagates into $g_{LS}^{(j)} - g^*$ (see (A.11) and following computations).

Step 2 (Generalization): The bounds we obtain on $\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix}$ in the direction of the jth frequency (A.63) show that the actual solution (g^*, h^*) does well in the min-max problem (3.7). The solution (g, h) to (3.7) will only do better. By concentration, the matrices in the least-squares problem $Q^{(\bullet)}$, $Q^{(j)}$ and in the actual expected square loss are comparable. Because (g, h) does well in the min-max problem, it will do comparably well with respect to the actual expected loss, when the input is one of the frequencies that has been tested, $\frac{2\pi j}{cr}$.

In this step, we already have enough to bound ε_2, the error in estimation with pure noise and no input signal.

Step 3 (Interpolation): We've produced (g, h) that is close to the actual (g^*, h^*) when tested on each of the frequencies $\frac{2\pi k}{cr}$, but need to extend this bound to all frequencies. Considering transfer functions and clearing denominators, this reduces to a problem about polynomial interpolation. We use a theorem from approximation theory (Theorem A.6) that bounds the maximum of a polynomial p on the unit circle, given its value at $\geq \deg p$ equispaced points. Note that it is crucial here that the number of parameters in g, h is less than the number of frequencies tested.

Note that we needed to clear $1 - z^{-1}H^*$ from the denominator, so we lose a factor of $\|H_{\text{unr}}^*\|_\infty = \left\|1 - z^{-1}H^*\right\|_\infty$ here. We obtain a bound on ε_1, finishing the proof of Theorem 4.1.

Step 4 (Truncation): Finally, we show that with a large burn-in time, the distribution of y's will be almost indistinguishable from the steady-state distribution, and hence the algorithm still works.

5. Conclusion and further directions

In the regime where Theorem 3.1 applies, we expect the dependence on the number of samples, as well as on $\|H_{\text{unr}}^*\|_\infty$, to be optimal. However, note that our theorem requires at least $\Omega(r^2)$ rollouts. It is an interesting question whether the bounds hold for fewer rollouts, or even for one rollout with carefully designed inputs, analogous to results in the case of LDS without hidden state (Simchowitz et al., 2018). Another open question is to prove a lower bound for the number of samples, in terms of $\|H_{\text{unr}}^*\|_\infty$.

By improperly learning the Kalman filter as an autoregressive model, we incur sample complexity depending on $\sqrt{r}$ rather than $\sqrt{d}$, where d is the dimension of the hidden state; obtaining bounds depending on d seems to be a difficult problem. Another important question is learning the optimal filter for noise models besides iid Gaussians.

We expect that the theorem can be generalized in a straightforward manner to multiple-input, multiple-output systems.

Finally, one can complete the "identify-then-control" pipeline by using the estimates from our algorithm for robust control. Although estimation in $\mathcal{H}_\infty$ norm of non-strictly stable systems is not possible in our setup, non-stable systems often arise in practice, so it is of great interest to find a weaker guarantees for such systems that still allow for robust control.

References

Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 172–184, 2013.

Brian DO Anderson and John B Moore. *Optimal filtering*. Courier Corporation, 2012.

Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 61(23):5987–5999, 2013.

Ross Boczar, Nikolai Matni, and Benjamin Recht. Finite-data performance guarantees for the output-feedback control of an unknown system. *arXiv preprint arXiv:1803.09186*, 2018.

G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice-Hall, 3 edition, 1994.

P. Brockwell and R. Davis. *Time Series: Theory and Methods*. Springer, 2 edition, 2009.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

J. Hamilton. *Time Series Analysis*. Princeton Univ. Press, 1994.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

P Jeff Harrison. Convergence and the constant dynamic linear model. *Journal of Forecasting*, 16(5): 287–292, 1997.

Thomas P Hayes. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2005.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pages 1–2, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. *arXiv preprint arXiv:1802.03981*, 2018.

BL Ho and Rudolph E Kalman. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82.1:35–45, 1960.

Edward W Kamen and Jonathan K Su. *Introduction to optimal estimation*. Springer Science & Business Media, 1999.

Mark Kozdoba, Jakub Marecek, Tigran T. Tchrakian, and Shie Mannor. On-line learning of linear dynamical systems: Exponential forgetting in kalman filters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4098–4105, 2019.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Lennart Ljung. *System identification: Theory for the User*. Prentice Hall, Upper Saddle River, NJ, 2 edition, 1998.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345, 1999.

Tuhin Subhra Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Lloyd N Trefethen. *Approximation theory and approximation practice*, volume 128. Siam, 2013.

Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. *ArXiv*, abs/1903.09122, 2019.

Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Peter Van Overschee and BL De Moor. *Subspace Identification for Linear Systems*. Springer Science & Business Media, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Appendix A. Proof

A.1. Concentration

We first set up notation and make some preliminary observations. A table of notation is provided in Section B. Let $\mathbf{X}_{\cos}^{(j)} \in \mathbb{R}^{r \times T}$ be the matrix with columns $x_{\cos}^{(j,k)}(t-1:t-r)$, $1 \le t \le T$ and likewise define $\mathbf{X}_{\sin}^{(j)}, \mathbf{Y}_{\cos}^{(j,k)}, \mathbf{Y}_{\sin}^{(j,k)}$, so that $M_*^{(j,k)} = \begin{pmatrix} \mathbf{X}_*^{(j,k)} \\ \mathbf{Y}_*^{(j,k)} \end{pmatrix}$ for $* \in \{\cos, \sin\}$. Let $\Gamma^{(\bullet)} = \mathbb{E}_{\eta^{(\bullet,k)}} M^{(\bullet,k)} M^{(\bullet,k)\top}$, $\Gamma_{*,t}^{(j)} = \mathbb{E}_{\eta_*^{(j,k)}} M_{*,t}^{(j,k)} M_{*,t}^{(j,k)\top}$, $\Gamma_{X,*,t}^{(j)} = \mathbf{X}_{*,t}^{(j)} \mathbf{X}_{*,t}^{(j)\top}$ where $* \in \{\cos, \sin\}$, $\eta^{(\bullet,k)}, \eta_*^{(j,k)}$ is the noise in the various rollouts. We will also write η for the noise from a generic rollout (so $\eta^{(\bullet,k)}$, $\eta_*^{(j,k)}$ are independent copies of η).

Let $\Gamma^{(j)} = \Gamma_{\cos,t}^{(j)} + \Gamma_{\sin,t}^{(j)}$ and $\Gamma_X^{(j)} = \Gamma_{X,\cos,t}^{(j)} + \Gamma_{X,\sin,t}^{(j)}$. These matrices not depend on t, which can be seen as follows. Consider the system response to $x^{(j)}(t) = e^{\frac{2\pi i j t}{cr}}$. (Although we cannot put in complex values in the system, there is a well-defined response for complex inputs.) Let $M^{(j)}$ be the matrix with columns $M_t^{(j)} = \begin{pmatrix} x^{(j)}(t-1:t-r) \\ y^{(j)}(t-1:t-r) \end{pmatrix}$, where the $y^{(j)}$ is defined as in (2.1) except with noise equal to $\eta^{(j)}(t) = \eta_{\cos}^{(j)}(t) + i\eta_{\sin}^{(j)}(t)$, $\eta_{\cos}^{(j)}(t), \eta_{\sin}^{(j)}(t) \sim N(0, \sigma^2)$. Because $M_{t+s}^{(j)}$ has the same distribution as $e^{\frac{2\pi i s}{cr}} M_t^{(j)}$, the expression $\mathbb{E}[M_t^{(j)} M_t^{(j)\dagger} + M_t^{(-j)} M_t^{(-j)\dagger}]$ does not depend on t. Expanding, it equals $\frac{1}{2}\mathbb{E}[(M_{\cos,t}^{(j)} + iM_{\sin,t}^{(j)})(M_{\cos,t}^{(j)} - iM_{\sin,t}^{(j)})^\top + (M_{\cos,t}^{(j)} - iM_{\sin,t}^{(j)})(M_{\cos,t}^{(j)} + iM_{\sin,t}^{(j)})^\top] = \Gamma_{\cos,t}^{(j)} + \Gamma_{\sin,t}^{(j)}$. Similarly, $\Gamma_X^{(j)}$ is well-defined. Note that $\Gamma_X^{(j)} = \mathbf{X}_{\cos,t}^{(j)} \mathbf{X}_{\cos,t}^{(j)\top} + \mathbf{X}_{\sin,t}^{(j)} \mathbf{X}_{\sin,t}^{(j)\top}$ has rank ≤ 2, as the columns of $\mathbf{X}_{\cos,t}^{(j)}$ and $\mathbf{X}_{\sin,t}^{(j)}$ are spanned by $x^{(\pm j)}(r:1)$.

Let $\overline{y}(t)$ denote the expected value of $y(t)$ given $y(s), x(s)$ for $s < t$: $\overline{y}(t+1) = g^* * x(t) + h^* * y(t)$. Let $\overline{\overline{y}}(t)$ denote the expected value of $y(t)$ given only the inputs $x(s)$ for $s < t$.

We first compute the error $h_{LS} - h^*$ and $g_{LS}^{(j)} - g^*$, and then the error in the mean response which is given by $\left(x_{\cos}^{(j)}(r:1)^\top \, \overline{\overline{y}}_{\cos}^{(j)}(r:1)^\top\right) \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right]$, and the analogous expression for $\sin$. This is broken up into subexpressions that we apply concentration bounds to.

Computing $h_{LS} - h^*$. Let

$$Q^{(\bullet)} = \sum_{k=1}^{c\ell r} M^{(\bullet,k)} M^{(\bullet,k)\top} = \sum_{k=1}^{c\ell r} \sum_{t=1}^{T} M_t^{(\bullet,k)} M_t^{(\bullet,k)\top} \tag{A.1}$$

$$Q^{(j)} = \sum_{k=1}^{\ell} (M_{\cos}^{(j,k)} M_{\cos}^{(j,k)\top} + M_{\sin}^{(j,k)} M_{\sin}^{(j,k)\top}) = \sum_{k=1}^{\ell} \sum_{t=1}^{T} (M_{\cos,t}^{(j,k)} M_{\cos,t}^{(j,k)\top} + M_{\sin,t}^{(j,k)} M_{\sin,t}^{(j,k)\top}). \tag{A.2}$$

We calculate the least squares solution h_{LS} and the error $h_{LS} - h^*$, noting that $y^{(\bullet,k)} = \overline{y}^{(\bullet,k)} + \eta^{(\bullet,k)}$.

$$h_{LS} = Q^{(\bullet)-1} \sum_{k=1}^{c\ell r} M^{(\bullet,k)} y^{(\bullet,k)} \tag{A.3}$$

$$h^* = Q^{(\bullet)-1} \sum_{k=1}^{c\ell r} M^{(\bullet,k)} \overline{y}^{(\bullet,k)} \tag{A.4}$$

$$h_{LS} - h^* = Q^{(\bullet)-1} \sum_{k=1}^{c\ell r} M^{(\bullet,k)} \eta^{(\bullet,k)} \tag{A.5}$$

$$= \Gamma^{(\bullet)-\frac{1}{2}} \underbrace{(\Gamma^{(\bullet)-\frac{1}{2}} Q^{(\bullet)} \Gamma^{(\bullet)-\frac{1}{2}})^{-1}}_{(0\bullet)} \underbrace{\Gamma^{(\bullet)-\frac{1}{2}} \sum_{k=1}^{c\ell r} M^{(\bullet,k)} \eta^{(\bullet,k)}}_{(1\bullet)} \tag{A.6}$$

Computing $g_{LS}^{(j)}$. The least squares solution $g_{LS}^{(j)}$ is

$$g_{LS}^{(j)} = \frac{1}{\ell T} \Gamma_X^{(j)+} \left[\sum_{k=1}^{\ell} [\mathbf{X}_{\cos}^{(j)} (y_{\cos}^{(j,k)} - \mathbf{Y}_{\cos}^{(j,k)} h_{LS}) + \mathbf{X}_{\sin}^{(j)} (y_{\sin}^{(j,k)} - \mathbf{Y}_{\sin}^{(j,k)} h_{LS})] \right]. \tag{A.7}$$

Noting that $\overline{y}_{\cos}^{(j,k)} = \mathbf{Y}_{\cos}^{(j,k)\top} h^* + \mathbf{X}_{\cos}^{(j)\top} g^*$, we calculate

$$y_{\cos}^{(j,k)} - \mathbf{Y}_{\cos}^{(j,k)} h_{LS} = \eta_{\cos}^{(j,k)} + \overline{y}_{\cos}^{(j,k)} - \mathbf{Y}_{\cos}^{(j,k)\top} h^* - \mathbf{Y}_{\cos}^{(j,k)\top} (h_{LS} - h^*) \tag{A.8}$$

$$= \eta_{\cos}^{(j,k)} + \mathbf{X}_{\cos}^{(j)\top} g^* - \mathbf{Y}_{\cos}^{(j,k)\top} (h_{LS} - h^*). \tag{A.9}$$

The analogous equation for $\sin$ holds. Substituting (A.9) into (A.7), letting $P_X^{(j)}$ be the projection onto the column space of $\Gamma_X^{(j)}$, and noting $\frac{1}{\ell T} \Gamma^{(j)+} \sum_{k=1}^{\ell} (\mathbf{X}_{\cos}^{(j)} \mathbf{X}_{\cos}^{(j)\top} + \mathbf{X}_{\sin}^{(j)} \mathbf{X}_{\sin}^{(j)\top}) g^* = P_X^{(j)} g^*$, we get

$$g_{LS}^{(j)} = P_X^{(j)} g^* + \frac{1}{\ell T} \Gamma_X^{(j)+} \left[\sum_{k=1}^{cr} [\mathbf{X}_{\cos}^{(j)} (\eta_{\cos}^{(j,k)} - \mathbf{Y}_{\cos}^{(j,k)\top} (h_{LS} - h^*)) + \mathbf{X}_{\sin}^{(j)} (\eta_{\sin}^{(j,k)} - \mathbf{Y}_{\sin}^{(j,k)\top} (h_{LS} - h^*))] \right] \tag{A.10}$$

Computing $\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix}$ (projected). We now calculate the error in $\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix}$, projected with $P_X^{(j)}$. The projection is because we do not care about the absolute error (which can be large), we only care about the mean error on the inputs $x_{\cos}^{(j)}$ and $x_{\sin}^{(j)}$, which are in the column space of $\Gamma_X^{(j)}$.

$$\begin{pmatrix} P_X^{(j)} & O \\ O & I_r \end{pmatrix} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \tag{A.11}$$

$$= \begin{pmatrix} \frac{1}{\ell T} \Gamma_X^{(j)+} \sum_{k=1}^{\ell} [\mathbf{X}_{\cos}^{(j)} (\eta_{\cos}^{(j,k)} - \mathbf{Y}_{\cos}^{(j,k)\top} (h_{LS} - h^*)) + \mathbf{X}_{\sin}^{(j)} (\eta_{\sin}^{(j,k)} - \mathbf{Y}_{\sin}^{(j,k)\top} (h_{LS} - h^*))] \\ h_{LS} - h^* \end{pmatrix}. \tag{A.12}$$

Let $\mathbf{Y}_{\cos}^{(j)}$ be the matrix with the mean responses to $x_{\cos}^{(j)}$, $\mathbf{Y}_{\cos,t}^{(j)} = \overline{y}_{\cos}^{(j)}(t-1:t-r)$, and likewise for $\sin$.

Computing $\left(x_{\cos}^{(j)}(r:1)^\top \, \overline{\overline{y}}_{\cos}^{(j)}(r:1)^\top\right)\left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS}\end{pmatrix} - \begin{pmatrix} g^* \\ h^*\end{pmatrix}\right]$. Write $y_{\cos}^{(j,k)} = \overline{\overline{y}}_{\cos}^{(j)} + \zeta_{\cos}^{(j,k)}$ and $\mathbf{Y}_{\cos}^{(j,k)} = \overline{\overline{\mathbf{Y}}}_{\cos}^{(j)} + \mathbf{Z}_{\cos}^{(j,k)}$, where $\zeta_{\cos}^{(j,k)}$ is the noise term and $\mathbf{Z}_{\cos}^{(j,k)}$ has $\zeta^{(j,k)}(t-1:t-r)$ as columns. (Note that $\eta_{\cos}^{(j,k)}$ only includes the new noise at each time step, while $\zeta_{\cos}^{(j,k)}$ is the accummulated noise; $\overline{y}^{(j)}$ is the expected value given the previous observations, and $\overline{\overline{y}}^{(j)}$ is the mean given only the inputs.) Then by (A.12), because $P_X^{(j)} x_{\cos}^{(j)}(r:1) = x_{\cos}^{(j)}(r:1)$,

$$\left(x_{\cos}^{(j)}(r:1)^\top \, \overline{\overline{y}}_{\cos}^{(j)}(r:1)^\top\right)\left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS}\end{pmatrix} - \begin{pmatrix} g^* \\ h^*\end{pmatrix}\right] \tag{A.13}$$

$$= x_{\cos}^{(j)}(r:1)^\top \frac{1}{\ell T}\Gamma_X^{(j)+}\left[\sum_{k=1}^{\ell} \mathbf{X}_{\cos}^{(j,k)}\eta_{\cos}^{(j,k)} + \mathbf{X}_{\sin}^{(j)}\eta_{\sin}^{(j,k)}\right] \tag{A.14}$$

$$+ \left[x_{\cos}^{(j)}(r:1)^\top \frac{1}{\ell T}\Gamma_X^{(j)+}\left[-\sum_{k=1}^{\ell}(\mathbf{X}_{\cos}^{(j)}\mathbf{Y}_{\cos}^{(j,k)\top} + \mathbf{X}_{\sin}^{(j)}\mathbf{Y}_{\sin}^{(j,k)\top})\right] + \overline{\overline{y}}_{\cos}^{(j)}(r:1)^\top\right](h_{LS} - h^*) \tag{A.15}$$

$$= x_{\cos}^{(j)}(r:1)^\top \frac{1}{\ell T}\Gamma_X^{(j)+}\left[\sum_{k=1}^{\ell} \mathbf{X}_{\cos}^{(j,k)}\eta_{\cos}^{(j,k)} + \mathbf{X}_{\sin}^{(j)}\eta_{\sin}^{(j,k)}\right] \tag{A.16}$$

$$+ \left[x_{\cos}^{(j)}(r:1)^\top \frac{1}{\ell T}\Gamma_X^{(j)+}\left[-\sum_{k=1}^{\ell}(\mathbf{X}_{\cos}^{(j)}\mathbf{Z}_{\cos}^{(j,k)\top} + \mathbf{X}_{\sin}^{(j)}\mathbf{Z}_{\sin}^{(j,k)\top})\right]\right](h_{LS} - h^*) \tag{A.17}$$

(see explanation below)

$$= \frac{1}{\ell T} x_{\cos}^{(j)}(r:1)^\top \Gamma_X^{(j)+\frac{1}{2}}\Bigg[\underbrace{\Gamma_X^{(j)+\frac{1}{2}}\sum_{k=1}^{\ell}(\mathbf{X}_{\cos}^{(j)}\eta_{\cos}^{(j,k)} + \mathbf{X}_{\sin}^{(j)}\eta_{\sin}^{(j,k)})}_{(1)} \tag{A.18}$$

$$+ \underbrace{\Gamma_X^{(j)+\frac{1}{2}}\sum_{k=1}^{\ell}(\mathbf{X}_{\cos}^{(j)}\mathbf{Z}_{\cos}^{(j,k)} + \mathbf{X}_{\sin}^{(j)}\mathbf{Z}_{\sin}^{(j,k)})(h_{LS} - h^*)}_{(2)}\Bigg] \tag{A.19}$$

In (A.17) we used that $x_{\cos}^{(j)}(r:1)^\top\Gamma_X^{(j)+}\left[-\sum_{k=1}^{\ell}(\mathbf{X}_{\cos}^{(j)}\overline{\overline{\mathbf{Y}}}_{\cos}^{(j)} + \mathbf{X}_{\sin}^{(j)}\overline{\overline{\mathbf{Y}}}_{\sin}^{(j)})\right] + \overline{\overline{y}}_{\cos}^{(j)}(r:1)^\top = 0$. To see this, let A be the matrix sending $x_*^{(j)}(t-1:t-r) \mapsto \overline{\overline{y}}_*^{(j)}(t-1:t-r)$ for $* \in \{\cos, \sin\}$. Then this equals $x_{\cos}^{(j)}(r:1)^\top\Gamma_X^{(j)+}\left[-\sum_{k=1}^{\ell}\sum_{t=1}^{T}(\mathbf{X}_{\cos,t}^{(j)}\mathbf{X}_{\cos,t}^{(j)\top} + \mathbf{X}_{\sin,t}^{(j)}\mathbf{X}_{\sin,t}^{(j)\top})A^\top\right] + x_{\cos}^{(j)}(r:1)^\top A^\top = 0$.

Computing $\left\|\Gamma^{(j)\frac{1}{2}}\left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS}\end{pmatrix} - \begin{pmatrix} g^* \\ h^*\end{pmatrix}\right]\right\|$.

$$\left\|\Gamma^{(j)\frac{1}{2}}\left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS}\end{pmatrix} - \begin{pmatrix} g^* \\ h^*\end{pmatrix}\right]\right\| \tag{A.20}$$

$$= \left\| \begin{pmatrix} \Gamma_{\cos,r+1}^{(j)\frac{1}{2}} \\ \Gamma_{\sin,r+1}^{(j)\frac{1}{2}} \end{pmatrix} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\| \tag{A.21}$$

because $\Gamma^{(j)} = \Gamma_{\cos,t}^{(j)} + \Gamma_{\sin,t}^{(j)}$ for any t (A.22)

$$= \sqrt{\sum_{*\in\{\cos,\sin\}} \left\| \Gamma_{*,r+1}^{(j)\frac{1}{2}} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\|^2} \tag{A.23}$$

$$= \sqrt{\sum_{*\in\{\cos,\sin\}} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right]^\top \mathop{\mathbb{E}}_{\eta_*^{(j)}} \begin{pmatrix} x_*^{(j)}(r:1) \\ y_*^{(j)}(r:1) \end{pmatrix} \left(x_*^{(j)}(r:1)^\top \, y_*^{(j)}(r:1)^\top \right) \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right]} \tag{A.24}$$

Because $\zeta_*^{(j)}(r:1)$ has mean 0,

$$\mathop{\mathbb{E}}_{\eta_*^{(j)}} \begin{pmatrix} x_*^{(j)}(r:1) \\ y_*^{(j)}(r:1) \end{pmatrix} \left(x_*^{(j)}(r:1)^\top \, y_*^{(j)}(r:1)^\top \right) \tag{A.25}$$

$$= \mathop{\mathbb{E}}_{\eta_*^{(j)}} \left[\begin{pmatrix} x_*^{(j)}(r:1) \\ \overline{y}_*^{(j)}(r:1) \end{pmatrix} \left(x_*^{(j)}(r:1)^\top \, \overline{y}_*^{(j)}(r:1)^\top \right) + \begin{pmatrix} 0 \\ \zeta_*^{(j)}(r:1) \end{pmatrix} \left(0 \; \zeta_*^{(j)}(r:1)^\top \right) \right] \tag{A.26}$$

Hence,

$$\left\| \Gamma^{(j)\frac{1}{2}} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\| \tag{A.27}$$

$$= \sqrt{\sum_{*\in\{\cos,\sin\}} \left\| \left(x_*^{(j)}(r:1)^\top \, \overline{y}_*^{(j)}(r:1)^\top \right) \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\|^2 + \mathop{\mathbb{E}}_{\eta_*^{(j)}} \left\| \zeta_*^{(j)}(r:1)^\top (h_{LS} - h^*) \right\|^2}. \tag{A.28}$$

Prospectus. In Section A.1.1, we lower bound $(0\bullet)$ in (A.6), and in Section A.1.2 we upper bound $(1\bullet)$ in (A.6) and (1) in (A.18). In Section A.1.3 we bound (2) in (A.19) and put the bounds together to obtain (for some C_7),

$$\left\| \Gamma^{(j)\frac{1}{2}} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\| \le \frac{C_7}{\sqrt{\ell T}} \left(\ln \left(\frac{c\ell r T}{\delta} \right) \right)^2 \tag{A.29}$$

A.1.1. MATRIX CONCENTRATION

Lemma A.1 (Concentration of sample covariance) *There are universal constants C_1, C_2 such that the following hold. Let $v_t \sim N(0, \Sigma)$ be iid, and let $\Sigma_m = \frac{1}{m} \sum_{t=1}^m v_t v_t^\top \in \mathbb{R}^{m\times m}$. Then for $\varepsilon = C_1 \left(\sqrt{\frac{r+u}{m}} + \frac{r+u}{m} \right)$,*

$$\mathbb{P}\left((1-\varepsilon)\Sigma^{\frac{1}{2}} \preceq \Sigma_m \preceq (1+\varepsilon)\Sigma^{\frac{1}{2}} \right) \ge 1 - 2e^{-u}. \tag{A.30}$$

Moreover, when $\varepsilon \le 1$ *and* $m \ge \left(\frac{C_2}{\varepsilon}\right)^2 \left(r + \ln\left(\frac{2}{\delta}\right)\right)$, *then*

$$\mathbb{P}\left((1-\varepsilon)\Sigma^{\frac{1}{2}} \preceq \Sigma_m \preceq (1+\varepsilon)\Sigma^{\frac{1}{2}}\right) \ge 1-\delta. \tag{A.31}$$

Proof The first part follows from (Vershynin, 2018, 4.7.3) on $\Sigma^{+\frac{1}{2}} v_t \sim N(0, P)$ where P is the projection onto the column space of Σ. (A^+ denotes the pseudoinverse of A.)

To get the second part from the first part, note that when $m \ge r + \ln\left(\frac{2}{\delta}\right)$, we can bound $\varepsilon_1 := C_1\left(\sqrt{\frac{r+\ln\left(\frac{2}{\delta}\right)}{m}} + \frac{r+\ln\left(\frac{2}{\delta}\right)}{m}\right) \le C_2\sqrt{\frac{r+\ln\left(\frac{2}{\delta}\right)}{m}}$ for $C_2 = 2C_1$. This is $\le \varepsilon$ under the condition on m. Hence

$$\mathbb{P}\left((1-\varepsilon)\Sigma^{\frac{1}{2}} \preceq \Sigma_m \preceq (1+\varepsilon)\Sigma^{\frac{1}{2}}\right) \ge \mathbb{P}\left((1-\varepsilon_1)\Sigma^{\frac{1}{2}} \preceq \Sigma_m \preceq (1+\varepsilon_1)\Sigma^{\frac{1}{2}}\right) \ge 1-\delta. \tag{A.32}$$

■

Lemma A.2 (Bounding $(0\bullet)$ **in** (A.6)**, etc.)** *For* $\ell \ge \left(\frac{C_2}{\varepsilon}\right)^2 \left(r + \ln\left(\frac{2}{\delta}\right)\right)$,

$$\mathbb{P}\left((1-\varepsilon)c\ell r\Gamma^{(\bullet)} \preceq \sum_{k=1}^{c\ell r} M_t^{(\bullet,k)} M_t^{(\bullet,k)\top} \preceq (1+\varepsilon)c\ell r\Gamma^{(\bullet)}\right) \ge 1-\delta \tag{A.33}$$

$$\mathbb{P}\left((1-\varepsilon)\ell\Gamma_{\cos,t}^{(j)} \preceq \sum_{k=1}^{\ell} M_{\cos,t}^{(j,k)} M_{\cos,t}^{(j,k)\top} \preceq (1+\varepsilon)\ell\Gamma_{\cos,t}^{(j)}\right) \ge 1-\delta \tag{A.34}$$

$$\mathbb{P}\left((1-\varepsilon)\ell\Gamma_{\sin,t}^{(j)} \preceq \sum_{k=1}^{\ell} M_{\sin,t}^{(j,k)} M_{\sin,t}^{(j,k)\top} \preceq (1+\varepsilon)\ell\Gamma_{\sin,t}^{(j)}\right) \ge 1-\delta \tag{A.35}$$

$$\mathbb{P}\left((1-\varepsilon)c\ell Tr\Gamma^{(\bullet)} \preceq \underbrace{\sum_{t=1}^{T}\sum_{k=1}^{c\ell r} M_t^{(\bullet,k)} M_t^{(\bullet,k)\top}}_{Q^{(\bullet)}} \preceq (1+\varepsilon)c\ell r\Gamma^{(\bullet)}\right) \ge 1-T\delta \tag{A.36}$$

$$\mathbb{P}\left((1-\varepsilon)\ell T\Gamma^{(j)} \preceq \underbrace{\sum_{t=1}^{T}\sum_{k=1}^{\ell}\left[M_{\cos,t}^{(j,k)} M_{\cos,t}^{(j,k)\top} + M_{\sin,t}^{(j,k)} M_{\sin,t}^{(j,k)\top}\right]}_{Q^{(j)}} \preceq (1+\varepsilon)\ell T\Gamma^{(j)}\right) \ge 1-2T\delta. \tag{A.37}$$

Proof The first three inequalities follow from applying Lemma A.1 to $M_t^{(\bullet,k)} \sim N(0, \Gamma^{(\bullet)})$, $M_{\cos,t}^{(j,k)} \sim N(0, \Gamma_{\cos,t}^{(j)})$ and $M_{\sin,t}^{(j,k)} \sim N(0, \Gamma_{\sin,t}^{(j)})$. The last two inequalities follow from a union

bound. ■

Note that we used independence between rollouts to obtain concentration, and union-bound within the rollouts.

A.1.2. VECTOR CONCENTRATION

We use the following two lemmas.

Lemma A.3 (χ^2_d-tail bound, Laurent and Massart (2000)) *For $t \geq 0$,*

$$\mathbb{P}_{x\sim N(0,I_d)}\left(\|x\|^2 \geq (d + 2(\sqrt{dt} + t))\right) \leq e^{-t} \tag{A.38}$$

Thus letting $C(d,\delta) := \left(d + 2\left(\sqrt{d\ln\left(\frac{1}{\delta}\right)} + \ln\left(\frac{1}{\delta}\right)\right)\right)^{\frac{1}{2}}$, $\mathbb{P}_{x\sim N(0,I_d)}(\|x\| \geq C(d,\delta)) \leq \delta$.

Note that $C(d,\delta) = O\left(\sqrt{d} + \sqrt{\ln\left(\frac{1}{\delta}\right)}\right)$.

Lemma A.4 (Azuma's inequality for vectors, Hayes (2005)) *Let $X_t, t \geq 0$ be a discrete-time martingale taking values in a real Euclidean space. Suppose that $X_0 = 0$ and for all $n \geq 1$, $\|X_n - X_{n-1}\| \leq c$. Then*

$$\mathbb{P}(\|X_n\| \geq a) \leq 2e^{1-\frac{(\frac{a}{c}-1)^2}{2n}}. \tag{A.39}$$

Lemma A.5 (Bounding $(1\bullet)$, (1) in (A.6) and (A.18)) *The following hold:*

$$\mathbb{P}\left(\left\|\sum_{k=1}^{c\ell r}\Gamma^{(\bullet)-\frac{1}{2}}M^{(\bullet,k)}\eta^{(\bullet,k)}\right\| \geq 3C\left(r,\frac{\delta}{4c\ell rT}\right)C\left(1,\frac{\delta}{4c\ell rT}\right)\sqrt{c\ell rT\ln\left(\frac{4}{\delta}\right)}\right) \leq \delta \tag{A.40}$$

$$\mathbb{P}\left(\left\|\sum_{k=1}^{\ell}\Gamma_X^{(j)+\frac{1}{2}}[\mathbf{X}_{\cos}^{(j)}\eta_{\cos}^{(j,k)} + \mathbf{X}_{\sin,t}^{(j)}\eta_{\sin}^{(j,k)}]\right\| \geq 3C\left(1,\frac{\delta}{4\ell T}\right)\sqrt{2\ell T\ln\left(\frac{4}{\delta}\right)}\right) \leq \delta. \tag{A.41}$$

Proof Consider the $c\ell rT$ partial sums of $\sum_{k=1}^{c\ell r}\sum_{t=1}^{T}\mathbb{1}[(A_{t,k}\cup B_{\bullet,t,k})^c]\Gamma^{(\bullet)-\frac{1}{2}}M_t^{(\bullet,k)}\eta^{(\bullet,k)}(t)$ where the events are defined as

$$A_{\bullet,t,k} = \left\{\left\|\eta^{(\bullet,k)}(t)\right\| > C\left(1,\frac{\delta}{4c\ell rT}\right)\right\} \tag{A.42}$$

$$B_{\bullet,t,k} = \left\{\left\|\Gamma^{(\bullet)-\frac{1}{2}}M_t^{(\bullet,k)}\right\| > C\left(r,\frac{\delta}{4c\ell rT}\right)\right\}. \tag{A.43}$$

Note this is a martingale as $M_t^{(\bullet,k)}$ is determined by $\eta^{(\bullet,k)}(s)$ for $s < t$, so Lemma A.4 applies. Note that $\Gamma^{(\bullet)-\frac{1}{2}}M_t^{(\bullet,k)} \sim N(0, I_r)$. We have by Lemma A.3 that for $a = 3C\left(r,\frac{\delta}{4c\ell rT}\right)C\left(1,\frac{\delta}{4c\ell rT}\right)\sqrt{c\ell rT\ln\left(\frac{4}{\delta}\right)}$,

$\mathbb{P}(A_{\bullet,t,k}), \mathbb{P}(B_{\bullet,t,k}) \leq \frac{\delta}{4c\ell rT}$. Hence

$$\mathbb{P}\left(\left\|\sum_{k=1}^{c\ell r}\left[\Gamma^{(\bullet)-\frac{1}{2}}M^{(\bullet,k)}\eta^{(\bullet,k)}\right]\right\| \geq a\right) \tag{A.44}$$

$$\leq \sum_{k=1}^{c\ell r}\sum_{t=1}^{T}[\mathbb{P}(A_{\bullet,t,k}) + \mathbb{P}(B_{\bullet,t,k})] + \mathbb{P}\left(\left\|\sum_{k=1}^{c\ell r}\sum_{t=1}^{T}\left[\mathbb{1}[(A_{\bullet,t,k}\cup B_{\bullet,t,k})^c]\Gamma^{(\bullet)+\frac{1}{2}}M_t^{(\bullet,k)}\eta^{(\bullet,k)}(t)\right]\right\| \geq a\right) \tag{A.45}$$

$$\leq 2c\ell rT\frac{\delta}{4c\ell rT} + \frac{\delta}{2} = \delta \tag{A.46}$$

where in the last inequality, we use Lemma A.5 and note that the definition of a implies because the following implications hold:

$$a \geq C\left(r, \frac{\delta}{4c\ell rT}\right)C\left(1, \frac{\delta}{4c\ell rT}\right)\left(1+\sqrt{2c\ell rT\left(1+\ln\left(\frac{4}{\delta}\right)\right)}\right) \tag{A.47}$$

$$\frac{\delta}{2} \geq 2e^{1-\frac{\left(\frac{a}{C\left(r,\frac{\delta}{4c\ell rT}\right)C\left(1,\frac{\delta}{4c\ell rT}\right)}-1\right)^2}{2c\ell rT}}. \tag{A.48}$$

Similarly, consider the $2\ell T$ partial sums of $\sum_{k=1}^{\ell}\sum_{t=1}^{T}\sum_{*\in\{\cos,\sin\}}\mathbb{1}\left[A_{*,t,k}^c\right]\Gamma_X^{(j)+\frac{1}{2}}\mathbf{X}_{*,t}^{(j)}\eta_*^{(j,k)}(t)$ where $A_{*,t,k} = \left\{\left\|\eta^{(\bullet,k)}(t)\right\| > C(1, \frac{\delta}{4rT})\right\}$. By Lemma A.4 and Lemma A.3, for $a = 3C\left(1, \frac{\delta}{4\ell T}\right)\sqrt{2\ell T\ln\left(\frac{4}{\delta}\right)}$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{\ell}\sum_{t=1}^{T}\sum_{*\in\{\cos,\sin\}}\left[\Gamma_X^{(j)+\frac{1}{2}}\mathbf{X}_{*,t}^{(j,k)}\eta_*^{(j,k)}(t)\right]\right\| \geq a\right) \tag{A.49}$$

$$\leq \sum_{k=1}^{\ell}\sum_{t=1}^{T}\mathbb{P}(A_{*,t,k}) + \sum_{k=1}^{\ell}\sum_{t=1}^{T}\sum_{*\in\{\cos,\sin\}}\mathbb{1}\left[A_{*,t,k}^c\right]\Gamma_X^{(j)+\frac{1}{2}}\mathbf{X}_{*,t}^{(j)}\eta_*^{(j,k)}(t) \tag{A.50}$$

$$\leq 2\ell T\left(\frac{\delta}{4\ell T}\right) + \frac{\delta}{2} = \delta. \tag{A.51}$$

■

A.1.3. PUTTING IT TOGETHER

Recall we are trying to bound $\left(x_{\cos}^{(j)}(r:1)^\top \overline{\overline{y}}_{\cos}^{(j)}(r:1)^\top\right)\left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS}\end{pmatrix} - \begin{pmatrix} g^* \\ h^*\end{pmatrix}\right]$ by bounding (A.18)–(A.19). There are constants $C_3, C_4, \ldots$ so that the following hold.

Bounding (1) in (A.18). By (A.41) in Lemma A.5,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{\ell}\Gamma_X^{(j)+\frac{1}{2}}[\mathbf{X}_{\cos}^{(j)}\eta_{\cos}^{(j,k)} + \mathbf{X}_{\sin,t}^{(j)}\eta_{\sin}^{(j,k)}]\right\| \geq 3C\left(1, \frac{\delta}{4\ell T}\right)\sqrt{2\ell T\ln\left(\frac{4}{\delta}\right)}\right) \leq \delta. \tag{A.52}$$

Bounding (2) in (A.19). By Lemma A.2 with $\delta \leftarrow\!\shortmid \frac{\delta}{T}$ and $\varepsilon \leftarrow\!\shortmid \frac{1}{2}$ and Lemma A.5, for $\ell \geq 4C_2^2\left(r + \ln\left(\frac{2T}{\delta}\right)\right)$, with probability $1-2\delta$,

$$\left\|\Gamma^{(\bullet)\frac{1}{2}}(h_{LS}-h^*)\right\| \leq \left\|(\Gamma^{(\bullet)-\frac{1}{2}}Q^{(\bullet)}\Gamma^{(\bullet)-\frac{1}{2}})^{-1}\right\| \left\|\Gamma^{(\bullet)-\frac{1}{2}}\sum_{k=1}^{c\ell r} M^{(\bullet,k)}\eta^{(\bullet,k)}\right\| \qquad \text{by (A.6)} \tag{A.53}$$

$$\leq \frac{2}{c\ell rT}\cdot 3C\left(r, \frac{\delta}{4c\ell rT}\right) C\left(1, \frac{\delta}{4c\ell rT}\right)\sqrt{c\ell rT\ln\left(\frac{4}{\delta}\right)} \tag{A.54}$$

$$\leq \frac{C_3}{\sqrt{c\ell T}}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^{\frac{3}{2}} \tag{A.55}$$

Note that $\zeta_*^{(j,k)}$ is distributed the same as $y^{(\bullet,k)}$. Hence $\sum_{k=1}^{\ell}\Gamma^{(\bullet)-\frac{1}{2}}\zeta_{\cos}^{(j,k)}(t-1:t-r) \sim N(0,\ell I_d)$, so by Lemma A.3, for each t,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{\ell}\Gamma^{(\bullet)-\frac{1}{2}}\zeta_{\cos}^{(j,k)}(t-1:t-r)\right\| \geq \sqrt{\ell}C\left(r,\frac{\delta}{2T}\right)\right) \leq \frac{\delta}{2T} \tag{A.56}$$

and similarly for $\sin$. Thus,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{\ell}\begin{pmatrix}\mathbf{Z}_{\cos}^{(j,k)\top}\\ \mathbf{Z}_{\sin}^{(j,k)\top}\end{pmatrix}\Gamma^{(\bullet)-\frac{1}{2}}\right\| \geq \sqrt{2T\ell}C\left(r,\frac{\delta}{2T}\right)\right) \leq \delta \tag{A.57}$$

Thus with probability $\geq 1-3\delta$,

$$\left\|\sum_{k=1}^{\ell}\begin{pmatrix}\mathbf{Z}_{\cos}^{(j,k)\top}\\ \mathbf{Z}_{\sin}^{(j,k)\top}\end{pmatrix}(h_{LS}-h^*)\right\| = \left\|\sum_{k=1}^{\ell}\begin{pmatrix}\mathbf{Z}_{\cos}^{(j,k)\top}\\ \mathbf{Z}_{\sin}^{(j,k)\top}\end{pmatrix}\Gamma^{(\bullet)-\frac{1}{2}}\right\|\left\|\Gamma^{(\bullet)\frac{1}{2}}(h_{LS}-h^*)\right\| \tag{A.58}$$

$$\leq \sqrt{2T\ell}C\left(r,\frac{\delta}{2T}\right)\frac{C_3}{\sqrt{c\ell T}}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^{\frac{3}{2}} \tag{A.59}$$

$$\leq C_4\sqrt{\frac{r}{c}}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^2. \tag{A.60}$$

For $\Sigma \succeq 0$, $v^\top(vv^\top+\Sigma)^+v \leq 1$[5], so each column of $\Gamma^{(j)+\frac{1}{2}}(\mathbf{X}_{\cos}^{(j)}\ \mathbf{X}_{\sin}^{(j)})$ has norm ≤ 1. Then,

$$\left\|\Gamma_X^{(j)+\frac{1}{2}}(\mathbf{X}_{\cos}^{(j)}\ \mathbf{X}_{\sin}^{(j)})\sum_{k=1}^{\ell}\begin{pmatrix}\mathbf{Z}_{\cos}^{(j,k)\top}\\ \mathbf{Z}_{\sin}^{(j,k)\top}\end{pmatrix}(h_{LS}-h^*)\right\| \leq \frac{C_4\sqrt{2Tr}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^2}{\sqrt{c}}. \tag{A.61}$$

Bounding $\left(x_{\cos}^{(j)}(r:1)^\top\ \overline{\overline{y}}_{\cos}^{(j)}(r:1)^\top\right)\left[\begin{pmatrix}g_{LS}^{(j)}\\ h_{LS}\end{pmatrix}-\begin{pmatrix}g^*\\ h^*\end{pmatrix}\right]$. Combining (A.52) and (A.61), with probability $1-4\delta$,

$$\frac{1}{\ell T}[(1)+(2)] \leq C_5\left(\frac{1}{\ell T}\right)\left(\sqrt{\ell T}\ln\left(\frac{\ell T}{\delta}\right)+\sqrt{\frac{Tr}{c}}\left(\ln\left(\frac{\ell rT}{\delta}\right)\right)^2\right) \leq \frac{C_6}{\sqrt{\ell T}}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^2 \tag{A.62}$$

5. By the Sherman-Morrison formula, $v^\top(vv^\top+\Sigma)^{-1}v = v^\top\Sigma^{-1}v - \frac{(v^\top\Sigma^{-1}v)^2}{1+v^\top\Sigma^{-1}v} \leq v^\top\Sigma^{-1}v$.

because $\ell \geq \frac{r}{c}$. Thus by (A.18)–(A.19),

$$\left\| \left(x_{\cos}^{(j)}(r:1)^\top \; \overline{y}_{\cos}^{(j)}(r:1)^\top \right) \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\| \leq \left\| x_{\cos}^{(j)}(r:1)^\top \Gamma^{(j)+\frac{1}{2}} \right\| \frac{C_6}{\sqrt{\ell T}} \left(\ln \left(\frac{c\ell r T}{\delta} \right) \right)^2 \tag{A.63}$$

$$\leq \frac{C_6}{\sqrt{\ell T}} \left(\ln \left(\frac{c\ell r T}{\delta} \right) \right)^2 \tag{A.64}$$

The analogous bound holds for $\sin$.

Bounding $\left\| \Gamma^{(j)\frac{1}{2}} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\|$. First, note that $\zeta_*^{(j)}(r:1) \sim N(0, \Gamma^{(\bullet)})$ so

$$\mathop{\mathbb{E}}_{\eta_*^{(j)}} \left\| \zeta_*^{(j)}(r:1)^\top (h_{LS} - h^*) \right\|^2 \leq \left\| \Gamma^{(\bullet)\frac{1}{2}} (h_{LS} - h^*) \right\|^2 \tag{A.65}$$

$$\leq \frac{C_3}{\sqrt{c\ell T}} \left(\ln \left(\frac{c\ell r T}{\delta} \right) \right)^{\frac{3}{2}} \tag{A.66}$$

provided that (A.55) holds. Now replace $\delta \leftarrow\!\shortmid \frac{\delta}{8}$. By (A.28), (A.64), and (A.66), with probability $1 - \delta$,

$$\left\| \Gamma^{(j)\frac{1}{2}} \left[\begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right] \right\| \leq \frac{C_7}{\sqrt{\ell T}} \left(\ln \left(\frac{c\ell r T}{\delta} \right) \right)^2. \tag{A.67}$$

A.2. Generalization

We now compute the performance of g^*, h^* on the minimax problem. Let

$$L^{(\bullet)}(h) = \frac{1}{r} \sum_{k=1}^{c\ell r} \left\| M^{(\bullet,k)\top} (h - h_{LS}) \right\|^2 \tag{A.68}$$

$$L^{(j)}(g,h) = \sum_{k=1}^{\ell} \left[\left\| M_{\cos}^{(j,k)\top} \begin{pmatrix} g \\ h \end{pmatrix} - y_{\cos}^{(j,k)} \right\|^2 + \left\| M_{\sin}^{(j,k)\top} \begin{pmatrix} g \\ h \end{pmatrix} - y_{\sin}^{(j,k)} \right\|^2 \right]. \tag{A.69}$$

Note that

$$L^{(\bullet)}(h) - L^{(\bullet)}(h_{LS}) = \frac{1}{r} (h - h_{LS})^\top Q^{(\bullet)} (h - h_{LS}) \tag{A.70}$$

$$L^{(j)}(g,h) - L^{(j)}(g_{LS}, h_{LS}) = \left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} \right]^\top Q^{(j)} \left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g_{LS}^{(j)} \\ h_{LS} \end{pmatrix} \right]. \tag{A.71}$$

We have that with probability $\geq 1 - \delta$, by (A.33) in Lemma A.2 and (A.55),

$$L^{(\bullet)}(h^*) - L^{(j)}(h_{LS}) \leq \frac{1}{r} \left\| \Gamma^{(\bullet)-\frac{1}{2}} Q^{(\bullet)} \Gamma^{(\bullet)-\frac{1}{2}} \right\| \left\| \Gamma^{(\bullet)\frac{1}{2}} (h^* - h_{LS}) \right\|^2 \tag{A.72}$$

$$\leq C_8 \frac{1}{r} (c\ell r T) \frac{1}{c\ell T} \left(\ln \left(\frac{c\ell r T}{\delta} \right) \right)^3 = C_8 \left(\ln \left(\frac{c\ell r T}{\delta} \right) \right)^3 \tag{A.73}$$

By (A.37) in Lemma A.2 and (A.67),

$$L^{(j)}(g^*,h^*)-L^{(j)}(g_{LS}^{(j)},h_{LS}) \le \left\|\Gamma^{(j)+\frac{1}{2}}Q^{(j)}\Gamma^{(j)+\frac{1}{2}}\right\| \left\|\Gamma^{(j)\frac{1}{2}}\left[\begin{pmatrix}g^*\\h^*\end{pmatrix}-\begin{pmatrix}g_{LS}^{(j)}\\h_{LS}\end{pmatrix}\right]\right\|^2 \tag{A.74}$$

$$\le C_8\ell T\frac{1}{\ell T}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^3 = C_8\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^3. \tag{A.75}$$

Because $\begin{pmatrix}g\\h\end{pmatrix}$ is the argmin of (3.7), we have $\left\|Q^{(\bullet)\frac{1}{2}}(h-h_{LS})\right\|_2^2 = r[L^{(\bullet)}(h)-L^{(\bullet)}(h_{LS})] \le C_8 r\left(\ln\left(\frac{\ell rT}{\delta}\right)\right)^3$. Hence

$$\left\|\Gamma^{(\bullet)\frac{1}{2}}(h-h^*)\right\|^2 \le 2\left(\left\|\Gamma^{(\bullet)\frac{1}{2}}(h-h_{LS})\right\|^2+\left\|\Gamma^{(\bullet)\frac{1}{2}}(h_{LS}-h^*)\right\|^2\right) \tag{A.76}$$

$$\le 2\left(\left\|Q^{(\bullet)-\frac{1}{2}}\Gamma^{(\bullet)}Q^{(\bullet)-\frac{1}{2}}\right\|\left\|Q^{(\bullet)\frac{1}{2}}(h-h_{LS})\right\|_2^2+\left\|\Gamma^{(\bullet)\frac{1}{2}}(h_{LS}-h^*)\right\|^2\right) \tag{A.77}$$

$$\le C_9\left(\frac{1}{c\ell rT}r\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^3+\frac{1}{c\ell T}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^3\right) \quad \text{by (A.55)} \tag{A.78}$$

$$\le \frac{C_9}{c\ell T}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^3 \tag{A.79}$$

and similarly

$$\left\|\Gamma^{(j)\frac{1}{2}}\left(\begin{pmatrix}g\\h\end{pmatrix}-\begin{pmatrix}g^*\\h^*\end{pmatrix}\right)\right\|_2^2 \le \frac{C_{10}}{\ell T}\left(\ln\left(\frac{\ell rT}{\delta}\right)\right)^4. \tag{A.80}$$

Now $\left\|\Gamma^{(\bullet)\frac{1}{2}}(h-h^*)\right\|_2^2$ represents the mean square estimation error when the input is 0 and the noise is $N(0,\sigma^2)$, so

$$\sigma\left\|(H-H^*)H_{\text{unr}}^*\right\|_2 = \left\|\Gamma^{(\bullet)\frac{1}{2}}(h-h^*)\right\| \le \frac{C_9}{\sqrt{c\ell T}}\left(\ln\left(\frac{c\ell rT}{\delta}\right)\right)^{\frac{3}{2}}. \tag{A.81}$$

This establishes one-half of Theorem 4.1.

We can decompose

$$M_{\cos,t}^{(j,k)} = \begin{pmatrix}x^{(j,k)}(t-1:t-r)\\ \overline{y}_{\cos}^{(j,k)}(t-1:t-r)\end{pmatrix}+\begin{pmatrix}0\\ \zeta_{\cos}^{(j,k)}(t-1:t-r)\end{pmatrix} \tag{A.82}$$

and similarly for sin. Define $M_t^{(j,k)}$ as follows: letting $y(t)$ be the response to $x(t)=e^{\frac{2\pi ijt}{cr}}$, $j\le\frac{cr}{2}$, with noise $\eta^{(j)}(t)=\eta_{\cos}^{(j)}(t)+i\eta_{\sin}^{(j)}(t)$, let $M_t^{(j,k)}=\begin{pmatrix}x(t-1:t-r)\\ y(t-1:t-r)\end{pmatrix}$. We can decompose the

mean response $\mathbb{E}M_t^{(j,k)} = \mathbb{E}[M_{\cos,t}^{(j,k)} + iM_{\sin,t}^{(j,k)}]$. We obtain an upper bound on the difference in the square mean response:

$$\left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix}\right]^\top (\mathbb{E}M_t^{(j,k)})(\mathbb{E}M_t^{(j,k)})^\top \left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix}\right] \tag{A.83}$$

$$= \left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix}\right]^\top \mathbb{E}[M_{\cos,t}^{(j,k)} - iM_{\sin,t}^{(j,k)}]\mathbb{E}[M_{\cos,t}^{(j,k)} + iM_{\sin,t}^{(j,k)}]^\top \left[\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix}\right] \tag{A.84}$$

$$\leq \left\| \Gamma^{(j)\frac{1}{2}} \left(\begin{pmatrix} g \\ h \end{pmatrix} - \begin{pmatrix} g^* \\ h^* \end{pmatrix} \right) \right\|_2^2 \tag{A.85}$$

$$\leq \frac{C_{10}}{\ell T} \left(\ln\left(\frac{c\ell r T}{\delta} \right) \right)^4 \tag{A.86}$$

using (A.80). Since the square mean response is exactly $\left|[(G - G^*) + (H - H^*)H_{\text{unr}}^* G^*](e^{\frac{2\pi ij}{cr}})\right|^2$, we get

$$\left|[(G - G^*) + z^{-1}(H - H^*)H_{\text{unr}}^* G](e^{\frac{2\pi ij}{cr}})\right| \leq \frac{C_{10}}{\sqrt{\ell T}} \left(\ln\left(\frac{c\ell r T}{\delta} \right) \right)^2. \tag{A.87}$$

Note the same inequality holds for j replaced by $r - j$ and $M_{\cos,t}^{(j,k)} + iM_{\sin,t}^{(j,k)}$ replaced by $M_{\cos,t}^{(j,k)} - iM_{\sin,t}^{(j,k)}$, so (A.87) holds for all $j \in \mathbb{Z}$.

A.3. Interpolation

Lemma A.6 *Let $Q(z) := \sum_{k=0}^{r-1} a_k z^k$, where $a_k \in \mathbb{C}$.*

1. *(Trefethen, 2013, Theorem 15.2) For any $N \geq r$, $\|Q\|_\infty \leq \left(\frac{2}{\pi}\ln(r+1) + 1\right) \max_{j=0,\ldots,N-1} |Q(e^{\frac{2\pi ij}{N}})|$.*
2. *(Bhaskar et al., 2013) For any $N \geq 4\pi r$, $\|Q\|_\infty \leq \left(1 + \frac{4\pi r}{N}\right) \max_{j=0,\ldots,N-1} |Q(e^{\frac{2\pi ij}{N}})|$.*

From (A.87) we get that for $\varepsilon = \frac{C_{10}}{\sqrt{\ell T}} \left(\ln\left(\frac{\ell r T}{\delta}\right)\right)^2$, $\omega = e^{\frac{2\pi i}{cr}}$, $j \in \mathbb{Z}$, that

$$\left|[(G - G^*) + z^{-1}(H - H^*)G^* H_{\text{unr}}^*](\omega^j)\right| \leq \varepsilon \tag{A.88}$$

$$\implies \left|[(G - G^*)(1 - z^{-1}H^*) + z^{-1}(H - H^*)G^*](\omega^j)\right| \leq \varepsilon \left|1 - \omega^{-j}H^*(\omega^j)\right| \leq \varepsilon(1 + \|H^*\|_\infty). \tag{A.89}$$

Suppose $c > 8\pi$. By Lemma A.6, since $(G - G^*)(1 - z^{-1}H^*) + z^{-1}(H - H^*)G^*$ has degree $\leq 2r$ in z^{-1},

$$\left\|(G - G^*)(1 - z^{-1}H) + z^{-1}(H - H^*)G^*\right\|_\infty \leq \varepsilon \left(1 + \frac{8\pi}{c}\right)(1 + \|H^*\|_\infty) \tag{A.90}$$

$$\implies \left\|(G - G^*) + z^{-1}(H - H^*)G^* H_{\text{unr}}^*\right\|_\infty \leq \varepsilon \left(1 + \frac{8\pi}{c}\right)(1 + \|H^*\|_\infty)\|H_{\text{unr}}^*\|_\infty. \tag{A.91}$$

This finishes the proof of Theorem 4.1.

A.4. Truncation error

We need the following lemma.

Lemma A.7 *Let* $F(z) = \sum_{t=0}^{\infty} f(t)z^{-t}$, $f(0) = 1$ *and* $G(z) = \frac{1}{F(z)} = \sum_{t=0}^{\infty} g(t)z^{-t}$. *Let* $K = \left(\sum_{t=0}^{r-1} |g(t)|\right)^2$. *Then letting* $f(t) = 0$ *for* $t < 0$, $\sum_{t=1}^{r} f(t-r:t-1)f(t-r:t-1)^\top \succeq \frac{1}{K^2} I_r$.

Proof For any power series $F(z) = \sum_{t=0}^{\infty} f(t)z^{-t}$, define $Z_F \in \mathbb{R}^{d\times d}$ by $(Z_F)_{i,j} = f(i-j)$. (Here, $f(i) = 0$ for $i < 0$.) Note that $Z_F Z_G = Z_{FG}$. Let $A = Z_F Z_F^\top = \sum_{t=1}^{r} f(t-r:t-1)f(t-r:t-1)^\top$. From $F(z)G(z) = 1$ we get $Z_G Z_F = I_d$, hence $Z_G A Z_G^\top = Z_G Z_F Z_F^\top Z_G^\top = I_d$. Because Z_G is invertible, we have $A \succeq \lambda I_d$ iff $Z_G(A - \lambda I_d)Z_G^\top \succeq 0$. Now $Z_G(A - \lambda I_d)Z_G^\top = I - \lambda Z_G Z_G^\top$. Letting $B = I - \lambda SS^\top$, we have

$$B_{ii} - \sum_{j\neq i} B_{ij} = 1 - \lambda \sum_{j,k} S_{ik}S_{jk} \geq 1 - \lambda K^2 \geq 0. \tag{A.92}$$

Thus by Gerschgorin's Disk Theorem, all eigenvalues of B are ≥ 0. ∎

Proof [Proof of Theorem 3.1] The proof of Theorem 3.1 relies on the following simple fact: If D_1, D_2 are two distributions on Ω with TV-distance $\leq \delta$, and $\mathcal{A}$ is any algorithm with input space Ω, then $\mathcal{A}(x), x \sim D_1$ and $\mathcal{A}(x), x \sim D_2$ also have TV-distance $\leq \delta$.

Consider Algorithm 1 run with signals x_∞ stretching back to $-\infty$ and signals $x_{\geq -L}$ only stretching back to $-L$. Consider the distributions they induce on $y(1:T)$. Suppose we choose L so that the TV-distance between those distributions is $\leq \delta' := \frac{\delta}{8c\ell r}$. Because there are $< 4c\ell r$ independent rollouts, the total TV-distance is $\leq \frac{\delta}{2}$. Then we can apply Theorem 4.1 with $\delta \leftarrow\!\shortmid \frac{\delta}{2}$ to get the desired result.

Let y_∞ and y_{fin} be the output signals given input signals x_∞ and $x_{\geq -L}$, and noise η_∞ and $\eta_{\geq -L}$. We have (using the shorthand $f_P := f\mathbb{1}_P$)

$$\begin{aligned}
y_\infty(t+1) &= h^*_{\text{unr}} * g^* * x_\infty(t) + h^*_{\text{unr}} * \eta_\infty(t+1) && \text{(A.93)}\\
y_{\text{fin}}(t+1) &= h^*_{\text{unr}} * g^* * (x_\infty \mathbb{1}_{\geq -L})(t) + [h^*_{\text{unr}} * (\eta_\infty \mathbb{1}_{\geq -L})](t+1) && \text{(A.94)}\\
&= [(h^*_{\text{unr}} * g^*)_{\leq L+t} * x_\infty](t) + (h^*_{\text{unr},\leq L+t+1} * \eta_\infty)(t+1) && \text{(A.95)}\\
y_\infty(t+1) - y_{\text{fin}}(t+1) &= [(h^*_{\text{unr}} * g^*)_{>L+t} * x_\infty](t) + (h^*_{\text{unr},>L+t+2} * \eta_\infty)(t+1) && \text{(A.96)}
\end{aligned}$$

To calculate the TV distance between the distributions of $y_\infty(1:T)$ and $y_{\text{fin}}(1:T)$, we need to bound the difference between the means and covariances.

Bounding difference in means. Note for $t \geq 0$, by the assumption $L \geq R_{H^*_{\text{unr}}G^*}(\varepsilon_2) - 1$ and Lemma 2.6, we have

$$[(h^*_{\text{unr}} * g^*)_{>L+t} * x_\infty](t) \leq \|(h^*_{\text{unr}} * g^*)_{\geq L+1}\|_1 \leq \varepsilon_2 \tag{A.97}$$

so $\|\mathbb{E}(y_\infty - y_{\text{fin}})(1:T)\| \leq \varepsilon_2\sqrt{T}$.

Bounding difference in covariances. Because $\mathbb{E}[\eta_\infty(i)\eta_\infty(j)] = \mathbb{1}_{i=j}$,

$$\begin{aligned}\mathrm{Cov}[y_\infty(1:T)]_{i,j} &= \mathbb{E}[y_\infty(i)y_\infty(j)] && \text{(A.98)}\\ &= \mathbb{E}[(h^*_{\text{unr}} * \eta_\infty)(i)(h^*_{\text{unr}} * \eta_\infty)(j)] && \text{(A.99)}\\ &= \mathbb{E}\left[\sum_{k=-\infty}^{\min\{i,j\}} h^*_{\text{unr}}(i-k)h^*_{\text{unr}}(j-k)\right] && \text{(A.100)}\end{aligned}$$

so

$$\mathrm{Cov}[y_\infty(1:T)] = \sum_{j=1}^{\infty} h^*_{\text{unr}}(j-T:j-1)h^*_{\text{unr}}(j-T:j-1)^\top. \tag{A.101}$$

Similarly

$$\mathrm{Cov}[y_{\text{fin}}(1:T)] = \sum_{j=1}^{\infty} h^*_{\text{unr},\le L+t+1}(j-T:j-1)h^*_{\text{unr},\le L+t+1}(j-T:j-1)^\top \tag{A.102}$$

Let $K = \left(1 + \sum_{t=0}^{T-2} |h^*(t)|\right)^2$. When $L + t + 2 \ge T$, by Lemma A.7 we can lower-bound this by

$$\mathrm{Cov}[y_{\text{fin}}(1:T)] \succeq \sum_{j=1}^{L+t+2} h^*_{\text{unr}}(j-T:j-1)h^*_{\text{unr}}(j-T:j-1)^\top \succeq \frac{1}{K^2} I_T \tag{A.103}$$

Also,

$$\mathrm{Cov}[y_\infty(1:T)] - \mathrm{Cov}[y_{\text{fin}}(1:T)] \preceq \sum_{j=L+T+2}^{\infty} h^*_{\text{unr}}(j-T+1:j)h^*_{\text{unr}}(j-T+1:j)^\top \tag{A.104}$$

$$\preceq \left(\sum_{j=L+T+2}^{\infty} \|h^*_{\text{unr}}(j-T+1:j)\|^2\right) I_T \tag{A.105}$$

$$\preceq T\left(\sum_{j=L+2}^{\infty} h^*_{\text{unr}}(j)^2\right) I_T \tag{A.106}$$

$$\preceq T\left(\sum_{j=L+2}^{\infty} |h^*_{\text{unr}}(j)|\right)^2 I_T \le T\varepsilon_1^2 I_T \tag{A.107}$$

where in the last inequality we used the assumption $L \ge R_{H^*_{\text{unr}}}(\varepsilon_1) - 2$ (for the ε_1 we will choose) and Lemma 2.6.

Bounding TV distance. For a random variable let $\mathcal{D}(X)$ denote its distribution. We apply the following formula for KL-divergence,

$$d_{KL}(N(\mu_1,\Sigma_1)||N(\mu_2,\Sigma_2)) = \frac{1}{2}\left[\ln\frac{|\Sigma_1|}{|\Sigma_2|} - d + \mathrm{Tr}(\Sigma_1^{-1}\Sigma_2) + (\mu_1-\mu_2)^\top\Sigma_1^{-1}(\mu_1-\mu_2)\right], \tag{A.108}$$

for $\mathcal{D}(y_{\text{fin}}(1:T)) = N(\mu_1, \Sigma_1)$ and $\mathcal{D}(y_\infty(1:T)) = N(\mu_2, \Sigma_2)$. Here, $\Sigma_1 \succeq \frac{1}{K^2} I_T$ and $\Sigma_2 - \Sigma_1 \preceq T\varepsilon_1^2 I_T$, so

$$d_{KL}(\mathcal{D}(y_{\text{fin}}(1:T))||\mathcal{D}(y_\infty(1:T))) \leq \frac{1}{2}\left[T\ln\left(\frac{1/K^2}{1/K^2+\varepsilon}\right) - T + T(1+K^2T\varepsilon^2) + K^2T\varepsilon_2^2\right] \tag{A.109}$$

$$\leq \frac{1}{2}(K^2T^2\varepsilon_1^2 + K^2T\varepsilon_2^2). \tag{A.110}$$

Now choose $\varepsilon_1 = \sqrt{\frac{\delta'^2}{2T^2K^2}}$ and $\varepsilon_2 = \sqrt{\frac{\delta'^2}{2TK^2}}$ to get this is $\leq \frac{\delta'^2}{2}$. Then by Pinsker's inequality,

$$d_{TV}(\mathcal{D}(y_{\text{fin}}(1:T)), \mathcal{D}(y_\infty(1:T))) \leq \sqrt{\frac{1}{2}\cdot\frac{\delta'^2}{2}} = \frac{\delta'}{2}. \tag{A.111}$$

This gives the desired result, noting that the assumption $L \geq \max\left\{R_{H^*_{\text{unr}}}\left(\frac{\delta}{4KT\sqrt{c\ell r}}\right), R_{H^*_{\text{unr}}G^*}\left(\frac{\delta}{4K\sqrt{c\ell rT}}\right)\right\}$ does indeed imply that the inequalities for L are indeed satisfied for the values of ε_1, ε_2 and $\delta' = \frac{\delta}{8c\ell r}$ we chose. Thus the TV-distance between the $y(1:T)$ of all the rollouts is at most $\frac{\delta}{2}$, as needed. ■

Appendix B. Notation

Notation	Definition
$H^*_{\text{unr}}(z)$	$\frac{1}{1-z^{-1}H^*(z)}$
$x^{(\bullet,k)} = x^{(\bullet)}$	$\mathbf{0}$ (the zero signal)
$x^{(j,k)}_{\cos} = x^{(j)}_{\cos}$	$t \mapsto \cos\left(\frac{2\pi jt}{cr}\right)$
$x^{(j,k)}_{\sin} = x^{(j)}_{\sin}$	$t \mapsto \sin\left(\frac{2\pi jt}{cr}\right)$
$y^{(\bullet,k)}, y^{(j,k)}_*$ $(* = \cos, \sin)$	Outputs for the above inputs
$\overline{y}^{(\bullet,k)}, \overline{y}^{(j,k)}_*$	Expected value given $y(s), x(s)$ for $s < t$
$\eta^{(\bullet,k)}, \eta^{(j,k)}_*$	$N(0,\sigma^2)$ noise in the rollouts; $y^{(j,k)}_* = \overline{y}^{(j,k)}_* + \eta^{(j,k)}_*$
$\overline{\overline{y}}^{(j)}_*$	Expected value given only x
$\zeta^{(j,k)}_*$	Accumulated noise for the inputs, $y^{(j,k)}_* = \overline{\overline{y}}^{(j,k)}_* + \zeta^{(j,k)}_*$
$M^{(\bullet,k)}$	Matrix with columns $y^{(\bullet,k)}(t-1:t-r)$, $1 \le t \le T$
$M^{(j,k)}_{*,t}$	Matrix with columns $\begin{pmatrix} x^{(j)}_*(t-1:t-r) \\ y^{(j,k)}_*(t-1:t-r) \end{pmatrix}$, $1 \le t \le T$
$\mathbf{X}^{(j)}_*$	Matrix with columns $x^{(j)}_*(t-1:t-r)$, $1 \le t \le T$
$\mathbf{Y}^{(j,k)}_*$	Matrix with columns $y^{(j,k)}_*(t-1:t-r)$, $1 \le t \le T$
$\mathbf{Z}^{(j,k)}_*$	Matrix with columns $\zeta^{(j,k)}_*(t-1:t-r)$, $1 \le t \le T$
$x^{(j)}$	$t \mapsto e^{\frac{2\pi ijt}{cr}}$
$\eta^{(j)}$	$\eta^{(j)}(t) = \eta^{(j)}_{\cos}(t) + i\eta^{(j)}_{\sin}(t)$, $\eta^{(j)}_{\cos}(t), \eta^{(j)}_{\sin}(t) \sim N(0,\sigma^2)$
$M^{(j)}$	Matrix with columns $\begin{pmatrix} x^{(j)}(t-1:t-r) \\ y^{(j)}(t-1:t-r) \end{pmatrix}$, $1 \le t \le T$
h_{LS}	Solution to (3.5)
$g^{(j)}_{LS}$	Solution to (3.6)
g, h	Solution to (3.7)
$\Gamma^{(\bullet)}$	$\mathbb{E}_{\eta^{(\bullet,k)}} M^{(\bullet,k)} M^{(\bullet,k)\top}$
$\Gamma^{(j)}_{*,t}$	$\mathbb{E}_{\eta^{(j,k)}_*} M^{(j,k)}_{*,t} M^{(j,k)\top}_{*,t}$
$\Gamma^{(j)}_{X,*,t}$	$\mathbf{X}^{(j)}_{*,t} \mathbf{X}^{(j)\top}_{*,t}$
$\Gamma^{(j)}$	$\Gamma^{(j)}_{\cos,t} + \Gamma^{(j)}_{\sin,t}$
$\Gamma^{(j)}_X$	$\Gamma^{(j)}_{X,\cos,t} + \Gamma^{(j)}_{X,\sin,t}$
$Q^{(\bullet)}$	$\sum_{k=1}^{c\ell r} M^{(\bullet,k)} M^{(\bullet,k)\top}$
$Q^{(j)}$	$\sum_{k=1}^{\ell} (M^{(j,k)}_{\cos} M^{(j,k)\top}_{\cos} + M^{(j,k)}_{\sin} M^{(j,k)\top}_{\sin})$
$P^{(j)}_X$	Projection onto column space of $\Gamma^{(j)}_X$
$L^{(\bullet)}(h)$	$\frac{1}{r}\sum_{k=1}^{c\ell r} \left\| M^{(\bullet,k)\top}(h - h_{LS}) \right\|^2$
$L^{(j)}(g,h)$	$\sum_{k=1}^{\ell} \left[\left\| M^{(j,k)\top}_{\cos} \begin{pmatrix} g \\ h \end{pmatrix} - y^{(j,k)}_{\cos} \right\|^2 + \left\| M^{(j,k)\top}_{\sin} \begin{pmatrix} g \\ h \end{pmatrix} - y^{(j,k)}_{\sin} \right\|^2 \right]$