A Proofs for model equivalence

In this section, we prove a generalization of Theorem 4.1 for both LDSs with observed inputs and LDSs with hidden inputs.

A.1 Preliminaries

Sum of ARMA processes It is known that the sum of ARMA processes is still an ARMA process.

Lemma A.1 (Main Theorem in [Granger and Morris, 1976])**.** *The sum of two independent stationary series generated by ARMA(p, m) and ARMA(q, n) is generated by ARMA(x, y), where $x \leq p+q$ and $y \leq \max(p+n, q+m)$.*

In shorthand notation, $ARMA(p,m) + ARMA(q,n) = ARMA(p+q, \max(p+n, q+m))$.

When two ARMAX processes share the same exogenous input series, the dependency on exogenous input is additive, and the above can be extended to $\text{ARMAX}(p,m,r) + \text{ARMAX}(q,n,s) = \text{ARMAX}(p+q, \max(p+n, q+m), \max(r,s))$.

Jordan canonical form and canonical basis Every square real matrix is similar to a complex block diagonal matrix known as its Jordan canonical form (JCF). In the special case for diagonalizable matrices, JCF is the same as the diagonal form. Based on JCF, there exists a canonical basis $\{e_i\}$ consisting only of eigenvectors and generalized eigenvectors of A. A vector v is a generalized eigenvector of rank μ with corresponding eigenvalue λ if $(\lambda I - A)^\mu v = 0$ and $(\lambda I - A)^{\mu-1} v \neq 0$.

Relating the canonical basis to the characteristic polynomial, the characteristic polynomial can be completely factored into linear factors $\chi_A(\lambda) = (\lambda - \lambda_1)^{\mu_1}(\lambda - \lambda_2)^{\mu_2} \cdots (\lambda - \lambda_r)^{\mu_r}$ over $\mathbb{C}$. The complex roots $\lambda_1, \cdots, \lambda_r$ are eigenvalues of A. For each eigenvalue λ_i, there exist μ_i linearly independent generalized eigenvectors v such that $(\lambda_i I - A)^{\mu_i} v = 0$.

A.2 General model equivalence theorem

Now we state Theorem A.1, a more detailed version of Theorem 4.1.

Theorem A.1. *For any linear dynamical system with parameters $\Theta = (A, B, C, D)$, hidden dimension n, inputs $x_t \in \mathbb{R}^k$, and outputs $y_t \in \mathbb{R}^m$, the outputs y_t satisfy*

$$\chi_A^\dagger(L) y_t = \chi_A^\dagger(L)\xi_t + \Gamma(L) x_t, \tag{5}$$

where L is the lag operator, $\chi_A^\dagger(L) = L^n \chi_A(L^{-1})$ is the reciprocal polynomial of the characteristic polynomial of A, and $\Gamma(L)$ is an m-by-k matrix of polynomials of degree $n-1$.

This implies that each dimension of y_t can be generated by an $ARMAX(n, n, n-1)$ model, where the autoregressive parameters are the characteristic polynomial coefficients in reverse order and in negative values.

To prove the theorem, we introduce a lemma to analyze the autoregressive behavior of the hidden state projected to a generalized eigenvector direction.

Lemma A.2. *Consider a linear dynamical system with parameters $\Theta = (A, B, C, D)$, hidden states $h_t \in \mathbb{R}^n$, inputs $x_t \in \mathbb{R}^k$, and outputs $y_t \in \mathbb{R}^m$ as defined in (1). For any generalized eigenvector e_i of A^* with eigenvector λ and rank μ, the lag operator polynomial $(1-\lambda L)^\mu$ applied to time series $h_t^{(i)} := \langle h_t, e_i \rangle$ results in*

$$(1-\lambda L)^\mu h_t^{(i)} = \textit{linear transformation of } x_t, \cdots, x_{t-\mu+1}.$$

Proof. To expand the LHS, first observe that

$$\begin{aligned}(1-\lambda L)h_t^{(i)} &= (1-\lambda L)\langle h_t, e_i\rangle \\ &= \langle h_t^{(i)}, e_i\rangle - \lambda L \langle h_t^{(i)}, e_i\rangle \\ &= \langle Ah_{t-1} + Bx_t, e_i\rangle - \langle h_{t-1}, \lambda e_i\rangle \\ &= \langle h_{t-1}^{(i)}, (A^* - \lambda I)e_i\rangle + \langle Bx_t, e_i\rangle.\end{aligned}$$

We can apply $(1-\lambda L)$ again similarly to obtain

$$\begin{aligned}(1-\lambda L)^2 h_t^{(i)} = \langle h_{t-2}, (A^* - \lambda I)^2 e_i\rangle \\ + \langle Bx_{t-1}, (A^* - \lambda I)e_i\rangle + (1-\lambda L)\langle Bx_t, e_i\rangle,\end{aligned}$$

and in general we can show inductively that

$$\begin{aligned}(1-\lambda L)^k h_t^{(i)} - \langle h_{t-k}, (A^* - \lambda I)^k e_i\rangle = \\ \sum_{j=0}^{k-1} (1-\lambda L)^{k-1-j} L^j \langle Bx_t, (A^* - \lambda I)^j e_i\rangle,\end{aligned}$$

where the RHS is a linear transformation of $x_t, \cdots, x_{t-k+1}$.

Since $(\lambda I - A^*)^\mu e_i = 0$ by definition of generalized eigenvectors, $\langle h_{t-\mu}, (A^* - \lambda I)^\mu e_i\rangle = 0$, and hence $(1-\lambda L)^\mu h_t^{(i)}$ itself is a linear transformation of $x_t, \cdots, x_{t-\mu+1}$. □

Proof for Theorem A.1 Using Lemma A.2 and the canonical basis, we can prove Theorem A.1.

Proof. Let $\lambda_1, \cdots, \lambda_r$ be the eigenvalues of A with multiplicity $\mu_1, \cdots, \mu_r$. Since A is a real-valued matrix, its adjoint $A*$ has the same characteristic polynomial and eigenvalues as A. There exists a canonical basis $\{e_i\}_{i=1}^n$ for A^*, where $e_1, \cdots, e_{\mu_1}$ are generalized eigenvectors

with eigenvalue λ_1, $e_{\mu_1+1}, \cdots, e_{\mu_1+\mu_2}$ are generalized eigenvectors with eigenvalue λ_2, so on and so forth, and $e_{\mu_1+\cdots+\mu_{r-1}+1}, \cdots, e_{\mu_1+\cdots+\mu_r}$ are generalized eigenvectors with eigenvalue λ_r.

By Lemma (A.2), $(1-\lambda_1 L)^{\mu_1} h_t^{(i)}$ is a linear transformation of $x_t, \cdots, x_{t-\mu_1+1}$ for $i = 1, \cdots, \mu_1$; $(1-\lambda_2 L)^{\mu_2} h_t^{(i)}$ is a linear transformation of $x_t, \cdots, x_{t-\mu_2+1}$ for $i = \mu_1+1, \cdots, \mu_1+\mu_2$; so on and so forth; $(1-\lambda_r L)^{\mu_r} h_t^{(i)}$ is a linear transformation of $x_t, \cdots, x_{t-\mu_r+1}$ for $i = \mu_1+\cdots+\mu_{r-1}+1, \cdots, n$.

We then apply lag operator polynomial $\Pi_{j\neq i}(1-\lambda_j L)^{\mu_j}$ to both sides of each equation. The lag polynomial in the LHS becomes $(1-\lambda_1 L)^{\mu_1}\cdots(1-\lambda_r L)^{\mu_r} = \chi_A^\dagger(L)$. For the RHS, since $\Pi_{j\neq i}(1-\lambda_j L)^{\mu_j}$ is of degree $n-\mu_i$, it lags the RHS by at most $n-\mu_i$ additional steps, and the RHS becomes a linear transformation of $x_t, \cdots, x_{t-n+1}$.

Thus, for each i, $\chi_A^\dagger(L) h_t^{(i)}$ is a linear transformation of $x_t, \cdots, x_{t-n+1}$.

The outputs of the LDS are defined as $y_t = Ch_t + Dx_t + \xi_t = \sum_{i=1}^n h_t^{(i)} Ce_i + Dx_t + \xi_t$. By linearity, and since $\chi_A^\dagger(L)$ is of degree n, both $\sum_{i=1}^n h_t^{(i)} Ce_i$ and $\chi_A^\dagger(L) Dx_t$ are linear transformations of $x_t, \cdots, x_{t-n}$. We can write any such linear transformation as $\Gamma(L)x_t$ for some m-by-k matrix $\Gamma(L)$ of polynomials of degree $n-1$. Thus, as desired,

$$\chi_A^\dagger(L) y_t = \chi_A^\dagger(L)\xi_t + \Gamma(L) x_t.$$

Assuming that there are no common factors in $\chi_A^\dagger$ and Γ, $\chi_A^\dagger$ is then the lag operator polynomial that represents the autoregressive part of y_t. This assumption is the same as saying that y_t cannot be expressed as a lower-order ARMA process. The reciprocal polynomial has the same coefficients in reverse order as the original polynomial. According to the lag operator polynomial on the LHS, $1-\varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_n L^n = \chi_A^\dagger(L)$, and $L^n - \varphi_1 L^{n-1} - \cdots - \varphi_n = \chi_A(L)$, so the i-th order autoregressive parameter φ_i is the negative value of the $(n-i)$-th order coefficient in the characteristic polynomial χ_A.

□

A.3 The hidden input case as a corollary

The statement about LDS without external inputs in Theorem 4.1 comes as a corollary to Theorem A.1, with a short proof here.

Proof. Define $y'_t = Ch_t + Dx_t$ to be the output without noise, i.e. $y_t = y'_t + \xi_t$. By Theorem A.1, $\chi_A^\dagger(L) y'_t = \Gamma(L) x_t$. Since we assume the hidden inputs x_t are i.i.d. Gaussians, y'_t is then generated by an ARMA$(n, n-1)$ process with autoregressive polynomial $\chi_A^\dagger(L)$.

The output noise ξ_t itself can be seen as an ARMA$(0,0)$ process. By Lemma A.1, ARMA$(n, n-1)$ + ARMA$(0,0)$ = ARMA$(n+0, \max(n+0, n-1+0))$ = ARMA(n,n). Hence the outputs y_t are generated by an ARMA(n,n) process as claimed in Theorem 4.1. It is easy to see in the proof of Lemma A.1 that the autoregressive parameters do not change when adding a white noise [Granger and Morris, 1976]. □

B Proof for eigenvalue approximation theorems

Here we restate Theorem 4.2 and Theorem 4.3 together, and prove it in three steps for 1) the general case, 2) the simple eigenvalue case, and 3) the explicit condition number bounds for the simple eigenvalue case.

Theorem B.1. *Suppose y_t are the outputs from an n-dimensional latent linear dynamical system with parameters $\Theta = (A, B, C, D)$ and eigenvalues $\lambda_1, \cdots, \lambda_n$. Let $\hat{\Phi} = (\hat{\varphi}_1, \cdots, \hat{\varphi}_n)$ be the estimated autoregressive parameters with error $\|\hat{\Phi} - \Phi\| = \epsilon$, and let $r_1, \cdots, r_n$ be the roots of the polynomial $1 - \hat{\varphi}_1 z - \cdots - \hat{\varphi}_n z^n$.*

Assuming the LDS is observable, the roots converge to the true eigenvalues with convergence rate $\mathcal{O}(\epsilon^{1/n})$. If all eigenvalues of A are simple (i.e. multiplicity 1), then the convergence rate is $\mathcal{O}(\epsilon)$. If A is symmetric, Lyapunov stable (spectral radius at most 1), and only has simple eigenvalues, then

$$|r_i - \lambda_i| \leq \frac{\sqrt{n 2^{n-1}}}{\Pi_{k\neq j}|\lambda_j - \lambda_k|}\epsilon + \mathcal{O}(\epsilon^2).$$

B.1 General (1/n)-exponent bound

This is a known perturbation bound on polynomial root finding due to Ostrowski [Beauzamy, 1999].

Lemma B.1. *Let $\Phi(z) = z^n + \varphi_1 z^{n-1} + \cdots + \varphi_{n-1} z + \varphi_n$ and $\Psi(z) = z^n + \psi_1 z^{n-1} + \cdots + \psi_{n-1} z + \psi_n$ be two polynomials of degree n. If $\|\Phi - \Psi\|_2 < \epsilon$, then the roots (r_k) of Φ and roots $(\tilde{r}_k)$ of Ψ under suitable order satisfy*

$$|r_k - \tilde{r}_k| \leq 4Cp\epsilon^{1/n},$$

where $C = \max_{1,0\leq k\leq n}\{|\varphi_n|^{1/n}, |\psi_n|^{1/n}\}$.

The general $\mathcal{O}(\epsilon^{1/n})$ convergence rate in Theorem 4.2 follows directly from Lemma B.1 and Theorem 4.1.

B.2 Bound for simple eigenvalues

The $\frac{1}{n}$-exponent in the above bound might seem not very ideal, but without additional assumptions the

$\frac{1}{n}$-exponent is tight. As an example, the polynomial $x^2 - \epsilon$ has roots $x \pm \sqrt{\epsilon}$. This is a general phenomenon that a root with multiplicity m could split into m roots at rate $O(\epsilon^m)$, and is related to the *regular splitting property* [Hryniv and Lancaster, 1999, Lancaster et al., 2003] in matrix eigenvalue perturbation theory.

Under the additional assumption that all the eigenvalues are simple (no multiplicity), we can prove a better bound using the following idea with companion matrix: Small perturbation in autoregressive parameters results in small perturbation in companion matrix, and small perturbation in companion matrix results in small perturbation in eigenvalues.

Matrix eigenvalue perturbation theory The perturbation bound on eigenvalues is a well-studied problem [Greenbaum et al., 2019]. The *regular splitting property* states that, for an eigenvalue λ_0 with partial multiplicities $m_1, \cdots, m_k$, an $O(\epsilon)$ perturbation to the matrix could split the eigenvalue into $M = m_1 + \cdots + m_k$ distinct eigenvalues $\lambda_{ij}(\epsilon)$ for $i = 1, \cdots, k$ and $j = 1, \cdots, m_i$, and each eigenvalue $\lambda_{ij}(\epsilon)$ is moved from the original position by $O(\epsilon^{1/m_i})$.

For semi-simple eigenvalues, geometric multiplicity equals algebraic multiplicity. Since geometric multiplicity is the number of partial multiplicities while algebraic multiplicity is the sum of partial multiplicities, for semi-simple eigenvalues all partial multiplicities $m_i = 1$. Therefore, the regular splitting property corresponds to the asymptotic relation in equation 6. It is known that regular splitting holds for any semi-simple eigenvalue even for non-Hermitian matrices.

Lemma B.2 (Theorem 6 in [Lancaster et al., 2003])**.** *Let $L(\lambda, \epsilon)$ be an analytic matrix function with semi-simple eigenvalue λ_0 at $\epsilon = 0$ of multiplicity M. Then there are exactly M eigenvalues $\lambda_i(\epsilon)$ of $L(\lambda, \epsilon)$ for which $\lambda_i(\epsilon) \to \lambda_0$ as $\epsilon \to 0$, and for these eigenvalues*

$$\lambda_i(\epsilon) = \lambda_0 + \lambda_i' \epsilon + o(\epsilon). \tag{6}$$

Companion Matrix Matrix perturbation theory tell us how perturbations on matrices change eigenvalues, while we are interested in how perturbations on polynomial coefficients change roots. To apply matrix perturbation theory on polynomials, we introduce the *companion matrix*, also known as the *controllable canonical form* in control theory.

Definition B.1. *For a monic polynomial $\Phi(u) = z^n + \varphi_1 z^{n-1} + \cdots + \varphi_{n-1} z + \varphi_n$, the companion matrix of the polynomial is the square matrix*

$$C(\Phi) = \begin{bmatrix} 0 & 0 & \dots & 0 & -\varphi_n \\ 1 & 0 & \dots & 0 & -\varphi_{n-1} \\ 0 & 1 & \dots & 0 & -\varphi_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -\varphi_1 \end{bmatrix}.$$

The matrix $C(\Phi)$ is the companion in the sense that its characteristic polynomial is equal to Φ.

In relation to a pure autoregressive AR(p) model, the companion matrix corresponds to the transition matrix in the linear dynamical system when we encode the values form the past p lags as a p-dimensional state

$$h_t = \begin{bmatrix} y_{t-p+1} & \cdots & y_{t-1} & y_t \end{bmatrix}^T.$$

If $y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p}$, then $h_t =$

$$\begin{bmatrix} y_{t-p+1} \\ y_{t-p+2} \\ \cdots \\ y_{t-1} \\ y_t \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \varphi_p & \varphi_{p-1} & \varphi_{p-2} & \dots & \varphi_1 \end{bmatrix} \begin{bmatrix} y_{t-p} \\ y_{t-p+1} \\ \cdots \\ y_{t-2} \\ y_{t-1} \end{bmatrix}$$
$$= C(-\Phi)^T h_{t-1}. \tag{7}$$

Proof of Theorem 4.2 for simple eigenvalues

Proof. Let y_t be the outputs of a linear dynamical system S with only simple eigenvalues, and let $\Phi = (\varphi_1, \cdots, \varphi_n)$ be the ARMAX autoregressive parameters for y_t. Let $C(\Phi)$ be the companion matrix of the polynomial $z^n - \varphi_1 z^{n-1} - \varphi_2 z^{n-2} - \cdots - \varphi_n$. The companion matrix is the transition matrix of the LDS described in equation 7. Since this LDS the same autoregressive parameters and hidden state dimension as the original LDS, by Corollary 4.1 the companion matrix has the same characteristic polynomial as the original LDS, and thus also has simple (and hence also semi-simple) eigenvalues. The $O(\epsilon)$ convergence rate then follows from Lemma B.2 and Theorem 5.1, as the error on ARMAX parameter estimation can be seen as perturbation on the companion matrix. □

A note on the companion matrix One might hope that we could have a more generalized result using Lemma B.2 for all systems with semi-simple eigenvalues instead of restricting to matrices with simple eigenvalues. Unfortunately, even if the original linear dynamical system has only semi-simple eigenvalues, in general the companion matrix is not semi-simple unless the original linear dynamical system is simple. This is because the

companion matrix always has its minimal polynomial equal to its characteristic polynomial, and hence has geometric multiplicity 1 for all eigenvalues. This also points to the fact that even though the companion matrix has the form of the controllable canonical form, in general it is not necessarily similar to the transition matrix in the original LDS.

B.3 Explicit bound for condition number

In this subsection, we write out explicitly the condition number for simple eigenvalues in the asymptotic relation $\lambda(\epsilon) = \lambda_0 + \kappa\epsilon + o(\epsilon)$, to show how it varies according to the spectrum. Here we use the notation $\kappa(C, \lambda)$ to note the condition number for eigenvalue λ in companion matrix C.

Lemma B.3. *For a companion matrix C with simple eigenvalues $\lambda_1, \cdots, \lambda_n$, the eigenvalues $\lambda'_1, \cdots, \lambda'_n$ of the perturbed matrix by $C + \delta C$ satisfy*

$$|\lambda_j - \lambda'_j| \leq \kappa(C, \lambda_j)\|\delta C\|_2 + o(\|\delta C\|_2^2), \tag{8}$$

and the condition number $\kappa(C, \lambda_j)$ is bounded by

$$\frac{1}{\prod_{k\neq j}|\lambda_j - \lambda_k|} \leq \kappa(C, \lambda_j) \leq \frac{\sqrt{n}}{\prod_{k\neq j}|\lambda_j - \lambda_k|}\left(\max(1, |\lambda_j|)\right)^{n-1}\left(1 + \rho(C)^2\right)^{\frac{n-1}{2}}, \tag{9}$$

where $\rho(C)$ is the spectral radius, i.e. largest absolute value of its eigenvalues.

In particular, when $\rho(C) \leq 1$, i.e. when the matrix is Lyapunov stable,

$$|\lambda_j - \lambda'_j| \leq \frac{\sqrt{n}(\sqrt{2})^{n-1}}{\prod_{k\neq j}|\lambda_j - \lambda_k|}\|\delta C\|_2 + o(\|\delta C\|_2^2). \tag{10}$$

Proof. For each simple eigenvalue λ of the companion matrix C with column eigenvector v and row eigenvector w^*, the condition number of the eigenvalue is

$$\kappa(C, \lambda) = \frac{\|w\|_2\|v\|_2}{|w^* v|}. \tag{11}$$

This is derived from differentiating the eigenvalue equation $Cv = v\lambda$, and multiplying the differentiated equation by $w*$, which results in

$$w^*(\delta C)v + w^* C(\delta v) = \lambda w^*(\delta v) + w^* v(\delta\lambda).$$

$$\delta\lambda = \frac{w^*(\delta C)v}{w^* v}.$$

Therefore,

$$|\delta\lambda| \leq \frac{\|w\|_2\|v\|_2}{|w^* v|}\|\delta C\|_2 = \kappa(C, \lambda)\|\delta C\|_2. \tag{12}$$

The companion matrix can be diagonalized as $C = V^{-1}\mathrm{diag}(\lambda_1, \cdots, \lambda_n)V$, the rows of the Vandermonde matrix V are the row eigenvectors of C, while the columns of V^{-1} are the column eigenvectors of C. Since the the j-th row $V_{j,*}$ and the j-th column $V^{-1}_{*,j}$ have inner product 1 by definition of matrix inverse, the condition number is given by

$$\kappa(C, \lambda_j) = \|V_{j,*}\|_2\ \|V^{-1}_{*,j}\|_2. \tag{13}$$

Formula for inverse of Vandermonde matrix The Vandermonde matrix is defined as

$$V = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{p-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{p-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \lambda_p & \lambda_p^2 & \cdots & \lambda_p^{p-1} \end{bmatrix}. \tag{14}$$

The inverse of the Vandermonde matrix V is given by [El-Mikkawy, 2003] using elementary symmetric polynomial.

$$(V^{-1})_{i,j} = \frac{(-1)^{i+j} S_{p-i,j}}{\prod_{k<j}(\lambda_j - \lambda_k)\prod_{k>j}(\lambda_k - \lambda_j)}, \tag{15}$$

where $S_{p-i,j} = S_{p-i}(\lambda_1, \cdots, \lambda_{j-1}, \lambda_{j+1}, \cdots, \lambda_p)$.

Pulling out the common denominator, the j-th column vector of V^{-1} is

$$\frac{(-1)^j}{\prod_{k<j}(\lambda_j - \lambda_k)\prod_{k>j}(\lambda_k - \lambda_j)}\begin{bmatrix} (-1)S_{p-1} \\ (-1)^2 S_{p-2} \\ \vdots \\ (-1)^{p-1} S_1 \\ (-1)^p \end{bmatrix},$$

where the elementary symmetric polynomials are over variables $\lambda_1, \cdots, \lambda_{j-1}, \lambda_{j+1}, \cdots, \lambda_p$.

For example, if $p = 4$, then the 3rd column (up to scaling) would be

$$\frac{-1}{(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)(\lambda_4 - \lambda_3)}\begin{bmatrix} -\lambda_1\lambda_2\lambda_4 \\ \lambda_1\lambda_2 + \lambda_1\lambda_4 + \lambda_2\lambda_4 \\ -\lambda_1 - \lambda_2 - \lambda_4 \\ 1 \end{bmatrix}.$$

Bounding the condition number As discussed before, the condition number for eigenvalue λ_j is

$$\kappa(C, \lambda_j) = \|V_{j,*}\|_2\ \|V^{-1}_{*,j}\|_2.$$

where $V_{j,*}$ is the j-th row of the Vandermonde matrix V and $V^{-1}_{*,j}$ is the j-th column of V^{-1}.

By definition $V_{j,*} = \begin{bmatrix} 1 & \lambda_j & \lambda_j^2 & \cdots & \lambda_j^{p-1} \end{bmatrix}$, so

$$\|V_{j,*}\|_2 = \left(\sum_{i=0}^{p-1} \lambda_j^{2i} \right)^{1/2}.$$

Using the above explicit expression for V^{-1}, $\|V_{*,j}^{-1}\|_2 = \frac{1}{\prod_{k \neq j} |\lambda_j - \lambda_k|} \left(\sum_{i=0}^{p-1} S_i^2(\lambda_1, \cdots, \lambda_{j-1}, \lambda_{j+1}, \cdots, \lambda_p) \right)^{1/2}$.

Therefore,

$$\kappa(C, \lambda_j) = \frac{1}{\prod_{k \neq j} |\lambda_j - \lambda_k|} \left(\sum_{i=0}^{p-1} \lambda_j^{2i} \right)^{1/2} \left(\sum_{i=0}^{p-1} S_i^2(\lambda_1, \cdots, \lambda_{j-1}, \lambda_{j+1}, \cdots, \lambda_p) \right)^{1/2}. \tag{16}$$

Note that both parts under $(\cdots)^{1/2}$ are greater than or equal to 1, so we can bound it below by

$$\kappa(C, \lambda_j) \geq \frac{1}{\prod_{k \neq j} |\lambda_j - \lambda_k|}.$$

We could also bound the two parts above. The first part can be bounded by

$$\left(\sum_{i=0}^{p-1} \lambda_j^{2i} \right)^{1/2} \leq \sqrt{p} \max(1, |\lambda_j|)^{(p-1)}. \tag{17}$$

While for the second part, since

$$|S_i(\lambda_1, \cdots, \lambda_{j-1}, \lambda_{j+1}, \cdots, \lambda_p)| \leq \binom{p-1}{i} |\lambda|_{\max}^i,$$

we have that

$$\sum_{i=0}^{p-1} S_i^2(\lambda_1, \cdots, \lambda_{j-1}, \lambda_{j+1}, \cdots, \lambda_p) \leq \sum_{i=0}^{p-1} \binom{p-1}{i} |\lambda|_{\max}^{2i} = (1 + |\lambda|_{\max}^2)^{p-1}. \tag{18}$$

Combining equation 17 and 18 for the upper bound, and putting it together with the lower bound,

$$\frac{1}{\prod_{k \neq j} |\lambda_j - \lambda_k|} \leq \kappa(C, \lambda_j) \leq \frac{\sqrt{p}}{\prod_{k \neq j} |\lambda_j - \lambda_k|} (\max(1, |\lambda_j|))^{p-1} (1 + \rho(C)^2)^{\frac{p-1}{2}}, \tag{19}$$

as desired. □

Theorem 4.3 follows from Lemma B.3, because the estimation error on the autoregressive parameters can be seen as the perturbation on the companion matrix, and the companion matrix has the same eigenvalues as the original LDS.

C Iterated regression for ARMAX

Algorithm We generalize Algorithm 1 to accommodate for exogenous inputs. Since the exogenous inputs are explicitly observed, including exogenous inputs in the regression does not change the consistent property of the estimator.

Theorem A.1 shows that different output channels from the same LDS have the same autoregressive parameters in ARMAX models. Therefore, we could leverage multidimensional outputs by estimating the autoregressive parameters in each channel separately and average them.

Algorithm 2: Regularized iterated regression for AR parameter estimation in ARMAX

Input: A time series $\{y_t\}_{t=1}^T$ where $y_t \in \mathbb{R}^m$, exogenous input series $\{x_t\}_{t=1}^T$ where $x_t \in \mathbb{R}^k$, and guessed hidden state dimension n.

for $d = 1, \cdots, m$ **do**

- Let $y_t^{(d)}$ be the projection of y_t to the d-th dimension;
- Initialize error term estimates $\hat{\epsilon}_t = \vec{0} \in \mathbb{R}^m$ for $t = 1, \cdots, T$;
- **for** $i = 0, \cdots, n$ **do**
  - Perform ℓ_2-regularized least squares regression on y_t against lagged terms of y_t, x_t, and $\hat{\epsilon}_t$ to solve for coefficients $\hat{\varphi}_j \in \mathbb{R}$, $\hat{\theta}_j \in \mathbb{R}$, and $\hat{\gamma}_j \in \mathbb{R}^k$ in the linear equation $y_t^{(d)} = c + \sum_{j=1}^n \hat{\varphi}_j y_{t-j}^{(d)} + \sum_{j=1}^{n-1} \hat{\gamma}_j x_{t-j} + \sum_{j=1}^i \hat{\theta}_j \hat{\epsilon}_{t-j}$, with ℓ_2-regularization only on $\hat{\theta}_j$;
  - Update $\hat{\epsilon}_t$ to be the residuals from the most recent regression;
- **end**
- Record $\hat{\Phi}^{(d)} = (\hat{\varphi}_1, \cdots, \hat{\varphi}_n)$;

end

Return the average estimate $\hat{\Phi} = \frac{1}{d}(\hat{\Phi}^{(1)} + \cdots + \hat{\Phi}^{(m)})$.

Again as before the i-th iteration of the regression only uses error terms from the past i lags. In other words, the initial iteration is an ARMAX$(n, 0, n-1)$ regression, the first iteration is an ARMAX$(n, 1, n-1)$ regression, and so forth.

Time complexity The iterated regression in each dimension involves $n+1$ steps of least squares regression each on at most $n(k+2)$ variables. Therefore, the total time complexity of Algorithm 2 is $O(nm((nk)^2T+(nk)^3)) = O(mn^3k^2T+mn^4k^3)$, where T is the sequence length, n the hidden state dimension, m the output dimension, and k the input dimension.

D Additional simulation details

D.1 Synthetic data generation

First, we generate K cluster centers by generating LDSs with random matrices A, B, C of standard i.i.d. Gaussians. We assume that the output y_t only depends on the hidden state h_t but not the input x_t, i.e. the matrix D is zero. When generating the random LDSs, we require that the spectral radius $\rho(A) \leq 1$, i.e. all eigenvalues of A have absolute values at most 1, and regenerate a new random matrix if the spectral radius is above 1. Our method also applies to the case of arbitrary spectral radius, this requirement is for the purpose of preventing numeric overflow in generated sequence. We also require that the ℓ_2 distance $d(\Theta_1, \Theta_2) = \|\lambda(A_1) - \lambda(A_2)\|_2$ between cluster centers are at least 0.2 apart.

Then, we generate 100 LDSs by randomly assigning them to the clusters. To obtain a LDS with assigned cluster center $\Theta = (A_c, B_c, C_c)$, we generate A' by adding a i.i.d. Gaussians to each entry of A_c, while B' and C' are new random matrices of i.i.d. standard Gaussians. The standard deviation of the i.i.d. Gaussians for $A' - A_c$ is chosen such that the average distance to cluster centers is less than half of the inter-cluster distance between centers.

For each LDS, we generate a sequence by drawing hidden inputs $x_t \sim N(0, 1)$ and put noise $\xi_t \sim N(0, 0.01^2)$ on the outputs.

D.2 Empirical correlation between AR distance and LDS distance.

Theorem 4.2 shows that LDSs with similar AR parameters also have similar eigenvalues. The converse of Theorem 4.2 is also true: dynamical systems with small eigenvalue distance have small autoregressive parameter distance, which follows from perturbation bounds for characteristic polynomials [Ipsen and Rehman, 2008]. Figure 2 shows simulation results where the AR parameter distance and the LDS eigenvalue distance are highly correlated.

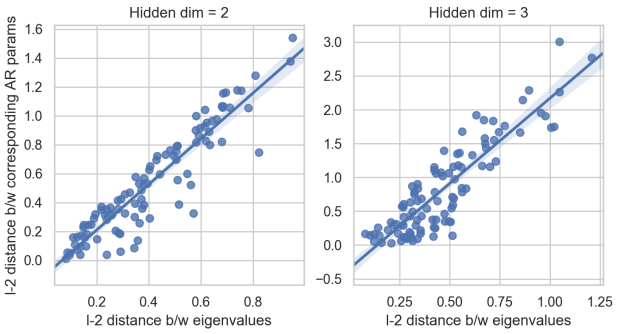

Figure 2: The eigenvalue ℓ_2 distance and the autoregressive parameter ℓ_2 distance for 100 random linear dynamical systems with eigenvalues drawn uniformly randomly from $[-1, 1]$. The two distance measures are highly correlated.