Two-Layer Linear Auto-Regressive Models Estimate Latent States

Yahya Sattar [1] Sunmook Choi [1] Leo Maynard-Zhang [2] Yassir Jedra [3] Maryam Fazel [2,4] Sarah Dean [1]

[1]Cornell University [2]University of Washington [3]Imperial College London [4]Amazon

Core Message

Auto-regressive prediction models can recover Kalman filter state estimates.

- We can construct a **two-layer linear AR model** to approximate Kalman filtering.
- The loss landscape of training the model is **benign** under persistence of excitation.
- **The learned hidden representation coincides**, up to a similarity transform, with the optimal Kalman filter state estimate.

Motivation

Auto-regressive models are trained to predict future observations from past data. In partially observed linear dynamical systems, the natural target representation is the Kalman filter state estimate.

Question. Can a simple auto-regressive model trained end-to-end on input-output data recover the latent state estimates of a Kalman filter?

Problem Setup

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad y_t = Cx_t + v_t$$

- x_t is latent; u_t, y_t are observed.
- We train on one input-output trajectory $\{(u_t, y_t)\}_{t=0}^{T+H}$.
- The learner does not know $A, B, C, \Sigma_w, \Sigma_v$.

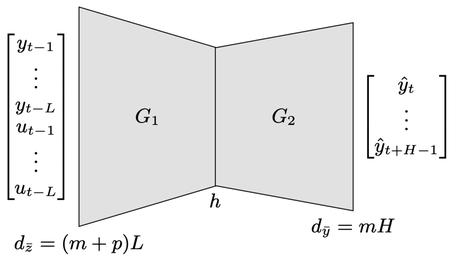
Identifiability. Latent states are identifiable only up to a similarity transform: x_t and Sx_t induce the same input-output behavior.

Assumptions

- During training, noise processes $\{w_t\}_{t\geq 0}$ and $\{v_t\}_{t\geq 0}$ and the excitations $\{u_t\}_{t\geq 0}$ are independent and zero-mean Gaussian with $\Sigma_w, \Sigma_v, \Sigma_u \succ 0$.
- The system is non-explosive: $\rho(A) \leq 1$.
- The system is $(A, \Sigma_w^{1/2})$ stabilizable and (A, C) observable.

Empirical Risk Minimization & Two-Layer AR Model

- Function class $\mathcal{F} = \{f(\bar{z}) = G_2 G_1 \bar{z} \colon (G_1, G_2) \in \mathcal{G}(h)\}$
  $\mathcal{G}(h) = \{(G_1, G_2) \colon G_1 \in \mathbb{R}^{h \times d_{\bar{z}}}, G_2 \in \mathbb{R}^{d_{\bar{y}} \times h}, \|G_1\|_F^2, \|G_2\|_F^2 \leq c\}$
- Squared Loss: $\mathcal{L}(G_1, G_2) = \frac{1}{2T} \sum_{t=1}^{T} \|y_{t:t+H-1} - G_2 G_1 \bar{z}\|_2^2$
- ERM: $(\hat{n}, \hat{G}_1, \hat{G}_2) \in \arg\min_{h \leq r, (G_1, G_2) \in \mathcal{G}(h)} \mathcal{L}(G_1, G_2)$.



Kalman Filter Predictor Form

$$\widehat{x}_{t+1} = \bar{A}\widehat{x}_t + Bu_t + Fy_t, \qquad y_t = C\widehat{x}_t + e_t$$

where $\bar{A} = A - FC$ and $e_t = C(x_t - \widehat{x}_t) + v_t$. Then,

$$y_{t:t+H-1} = \mathcal{O}\mathcal{C}\bar{z}_t + \mathcal{O}\bar{A}^L \widehat{x}_{t-L} + \xi_t$$

for $\xi_t = \mathcal{T}_u u_{t:t+H-2} + \mathcal{T}_e e_{t:t+H-1}$. Here,

$$\mathcal{O} = \begin{bmatrix} C \\ C\bar{A} \\ \vdots \\ C\bar{A}^{H-1} \end{bmatrix} \quad \text{and} \quad \mathcal{C} = \begin{bmatrix} F & \bar{A}F & \cdots & \bar{A}^{L-1}F & B & \bar{A}B & \cdots & \bar{A}^{L-1}B \end{bmatrix}$$

are extended observability and controllability matrices, respectively.

Optimization Landscape

two-layer factorization → nonconvex ERM → benign landscape

every local minimum is global

every saddle is strict

Takeaway: No spurious local minima from the two-layer factorization.

Finite-Sample Statistical Analysis

- **In-Sample Prediction Error**: On the **training data** $\{(u_t, y_t)\}_{t=0}^{T}$, w.p. $1-\delta$,

$$\frac{1}{T}\sum_{t=1}^{T} \|(\widehat{G}_2\widehat{G}_1 - \mathcal{O}\mathcal{C})\bar{z}_t\|_2^2 \lesssim \frac{H}{T}\Big(r(d_{\bar{y}} + d_{\bar{z}})\log(T) + \log(T/\delta)\Big)$$

- **Parameter Estimation Error**: For $T \gtrsim L\Big(d_{\bar{z}}\log(T/L) + \log(1/\delta)\Big)$ and $L \gtrsim \log(T)$, w.p. $1-\delta$,

$$\|\widehat{G}_2\widehat{G}_1 - \mathcal{O}\mathcal{C}\|_F^2 \lesssim \frac{H}{T}\Big(r(d_{\bar{y}} + d_{\bar{z}})\log(T) + \log(T/\delta)\Big)$$

- **Latent State Recovery**: Suppose $\mathcal{O}$ and $\mathcal{C}$ are full rank, and let $\sigma_n := \min\{\sigma_n(\mathcal{O}\mathcal{O}^\top), \sigma_n(\mathcal{C}\mathcal{C}^\top)\}$. For any **test** sequence of $\{(u_\tau, y_\tau)\}_{\tau=0}^{t}$, if

$$L \gtrsim \log(T) \quad \text{and} \quad \|\widehat{G}_2\widehat{G}_1 - \mathcal{O}\mathcal{C}\|_F \leq \sigma_n,$$

there exists a similarity transform S such that, w.p. $1-\delta$,

$$\|\widehat{x}_t - S\widehat{G}_1\bar{z}_t\|_2^2 \lesssim \frac{H\|\bar{z}_t\|_2^2}{\sigma_n T}\Big(r(d_{\bar{y}} + d_{\bar{z}})\log(T) + \log(T/\delta)\Big)$$

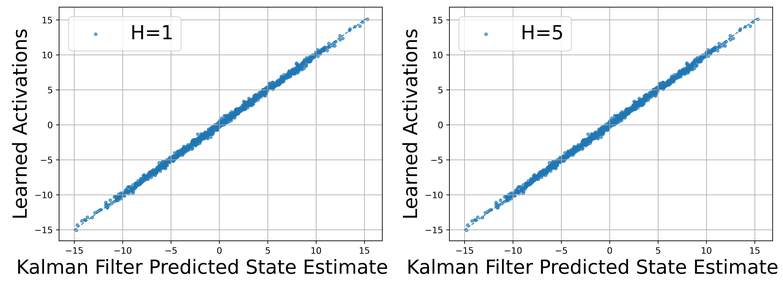
[arXiv link] [Yahya's Live Zoom Link]

Experiments

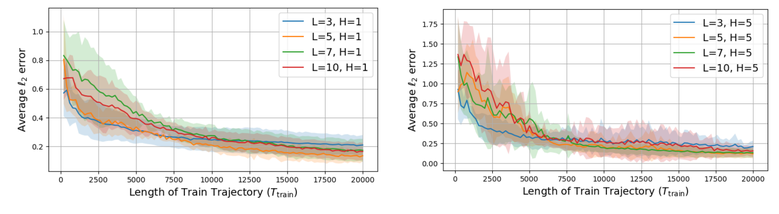
Protocol. Synthetic LDS and ControlGym systems; train a two-layer linear AR model on one trajectory; align hidden activations by

$$\widehat{S} = \arg\min_{S} \sum_{t} \|\widehat{x}_t - S\,\widehat{G}_1\bar{z}_t\|^2.$$

Latent recovery: Kalman estimate vs. mapped hidden activation



Sample complexity: recovery error decreases with T_{train}



Mean R^2 after similarity fit

System	State dim.	$H=1$	$H=5$
Random LDS	4	0.999	0.998
ControlGym `umv`	8	0.995	0.994
ControlGym `ac6`	10	0.979	0.980

Architecture Search for Random LDS

Hidden Dim.	Training Loss (mean ± std)
1	309.5980 ± 167.9855
2	0.6661 ± 0.0473
3	0.7204 ± 0.0565
4	**0.6179 ± 0.0408**
5	0.6921 ± 0.0586
6	0.6525 ± 0.0482
7	0.6212 ± 0.0434
8	0.6234 ± 0.0453
9	0.6372 ± 0.0452
10	0.6456 ± 0.0518

Empirical Pattern

- Hidden activations align with Kalman state estimates after a linear map.
- Recovery error decreases with T_{train}.
- Best hidden dimension matches the true state dimension in synthetic systems.

Key Takeaways

Takeaway. Prediction-trained two-layer linear AR models learn more than input-output statistics: their hidden activations recover Kalman filter state estimates up to similarity.

1. **Representation:** hidden activations recover latent state estimates.
2. **Optimization:** nonconvex two-layer linear ERM has a benign landscape.
3. **Theory:** finite-sample prediction, parameter estimation, and latent recovery guarantees.