On-Line Learning of Linear Dynamical Systems: Exponential Forgetting in Kalman Filters

Mark Kozdoba*, Jakub Marecek+, Tigran Tchrakian+, Shie Mannor*

The Technion (*) and IBM Research – Ireland (+)

Abstract

- We give an on-line time series prediction algorithm which considers only a few most recent observations.
- We compare, via regret bounds, the results of our algorithm to the best, in hindsight, Kalman filter for a given signal.
- Technically, we show that the dependence of a prediction of Kalman filter on the past is decaying exponentially, whenever the process noise is non-degenerate.
- Thus, Kalman filter may be approximated by regression on a few recent observations.
- Improper, off-model learning of a linear dynamical system (LDS).

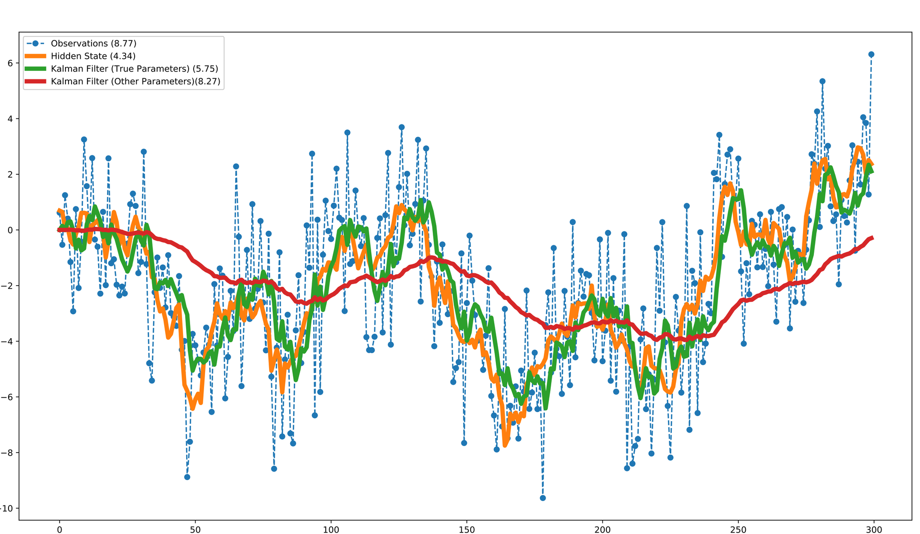

Figure 1: Prediction Errors: Last Seen Value: 8.77, Kalman Slow: 8.27, Kalman True (optimal): 5.75. Known Hidden State: 4.34

A Linear Dynamical System

A linear system $L = (G, F, v, W)$ is:

$$\phi_t = G\phi_{t-1} + \omega_t$$
$$Y_t = F'\phi_t + \nu_t,$$

where

- Y_t are scalar observations, $\phi_t \in \mathbb{R}^{n\times 1}$ is the hidden state, and ω_t, ν_t are iid noises with covariances W, v.

Kalman Filter

- An estimate of the current hidden state, given the observations for $t \geq 1$:
$$m_t = \mathbb{E}\left(\phi_t | Y_0, \ldots, Y_t\right),$$
and let C_t be the covariance matrix of ϕ_t given $Y_0, \ldots, Y_t$.
- Forecast of the next observation, given the current data:
$$f_{t+1} = \mathbb{E}\left(Y_{t+1} | Y_t, \ldots, Y_0\right) = F'Gm_t.$$
- m_t is usually computed recursively.

Kalman Filter Unrolled

$$f_{t+1} = F'GA_tY_t + F'\sum_{j=0}^{s-1}\left[\left(\prod_{i=0}^{j} Z_{t-i}\right)GA_{t-j-1}Y_{t-j-1}\right] + F'\left(\prod_{i=0}^{s} Z_{t-i}\right)a_{t-s}.$$

where A_t, Z_t, a_t are computed recursively using the LDS parameters (G, F, v, W) and Y_t.

Theorem (LDS Approximation)

Let $L = L(F, G, v, W)$ be an observable LDS with $\mathbf{W > 0}$.
For any $\varepsilon > 0$, and any $B_0 > 0$, there is $T_0 > 0$, $s > 0$ and $\theta \in \mathbb{R}^s$, such that for every sequence Y_t with $|Y_t| \leq B_0$, and for every $t \geq T_0$,

$$\left|f_{t+1} - \sum_{i=0}^{s-1}\theta_i Y_{t-i}\right| \leq \varepsilon.$$

- Similar result holds for Lipschitz sequences, $|Y_{t+1} - Y_t| \leq B_1$.

Necessity of Noise ($W > 0$)

With $n = 1$, assume that Y_t are generated by an LDS with $G = F = 1$, $W = 0$ and some $v > 0$. Assume that the true process starts from a deterministic state $m_0 > 0$.

- This is equivalent to estimating the mean (m_0) of a random variable from samples.
- In this case, based on fixed number of observations we can not compete with an estimator based on all observations.
- The decay is not exponential.
- Similar considerations apply more generally.

An Algorithm

Let us consider an on-line gradient descent:

1: **Input:** Regression length s, domain bound D. Observations $\{Y_t\}_0^\infty$, given sequentially.
2: Set the learning rate $\eta_t = t^{-\frac{1}{2}}$.
3: Initialize θ_s arbitrarily in $\mathcal{D}$.
4: **for** $t = s$ **to** ∞ **do**
5: Predict $\hat{y}_t = \sum_{i=0}^{s-1}\theta_{t,i}Y_{t-i-1}$
6: Observe Y_t and compute the loss $\ell_t(\theta_t)$
7: Update $\theta_{t+1} \leftarrow \pi_{\mathcal{D}}\left(\theta - \eta_t\nabla\ell_t(\theta_t)\right)$
where the gradient $\nabla_\theta \ell_t(\theta)$ of the cost at θ at time t is given by
$$-2\left(Y_t - \sum_{i=0}^{s-1}\theta_i Y_{t-i-1}\right)\left(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-s}\right).$$
8: **end for**

Theorem (Regret Bound)

Let S be a finite family of observable LDSs with $W_S > 0$ for all S. Let B_0 be given. For any $\varepsilon > 0$, there are s, D, and C_S, such that the following holds:
For every sequence Y_t with $|Y_t| \leq B_0$, if θ_t is a sequence produced by the algorithm with parameters s and D, then for every $T > 0$, the regret

$$\sum_{t=0}^{T}\ell_t(\theta_t) - \min_{L\in S}\sum_{t=0}^{T}\ell(Y_t, f_t(L))$$

is bounded by

$$C_S + 2(D^2 + B_0^2)\sqrt{T} + \varepsilon T.$$

Experiments

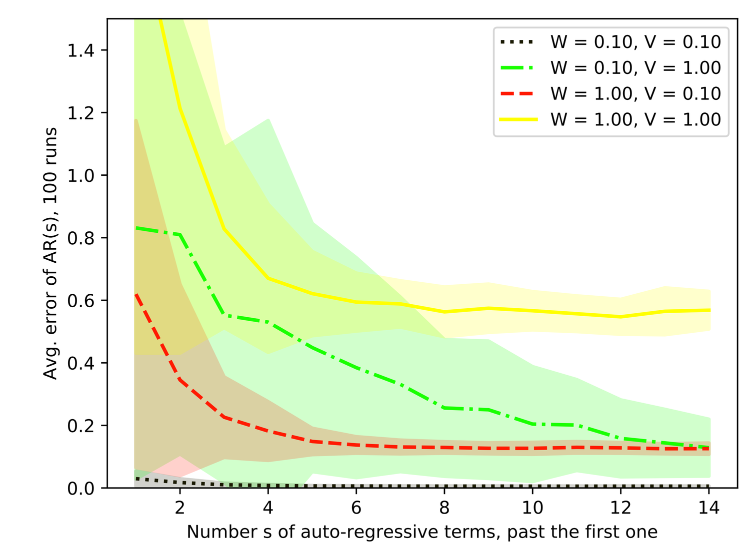

Figure 2: The error of AR($s+1$) as a function of $s+1$, in terms of the mean and standard deviation over $N = 100$ runs on Example of Hazan et al, for 4 choices of W, v of process, observation noise, respectively.

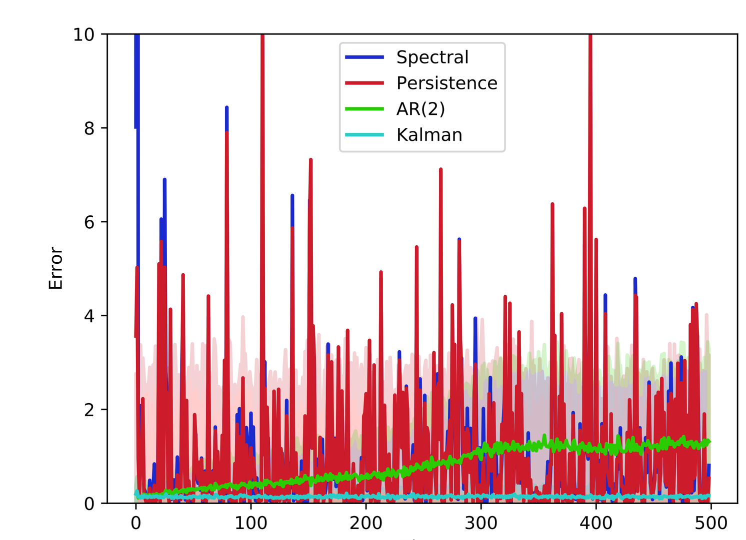

Figure 3: Illustrations on stock-market data used by Liu et al (AAAI 2016). Sample outputs and predictions with AR(2), compared against Kalman filter, last-value prediction, and spectral filtering.

References

[1] Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. COLT, 2013.

[2] Chenghao Liu, Steven C. H. Hoi, Peilin Zhao, and Jianling Sun. Online arima algorithms for time series prediction. AAAI, 2016.

[3] Elad Hazan, Karan Singh, and Cyril Zhang. Online learning of linear dynamical systems. NIPS, 2017.

Acknowledgements

This research received funding from the European Union Horizon 2020 Programme (Horizon2020/2014-2020) under grant agreement number 688380 (project VaVeL).

Contact Information

- markk@technion.ac.il
- jakub.marecek@ie.ibm.com
- tigran@ie.ibm.com
- shie@ee.technion.ac.il

TECHNION Israel Institute of Technology