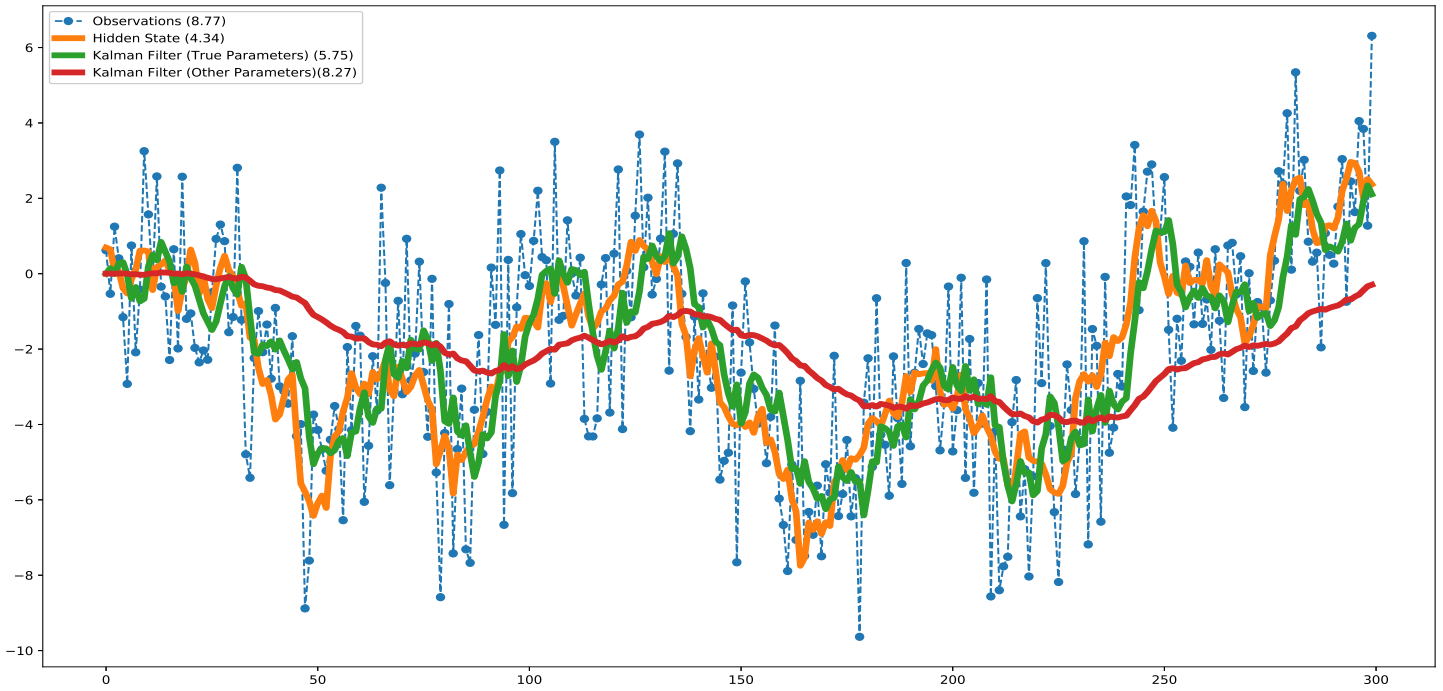
On-Line Learning of Linear Dynamical Systems: Exponential Forgetting in Kalman Filters

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

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

January 30th, 2019

Filter Demo



- Kalman filter is a key tool for time-series forecasting and analysis.
- Prediction Error: 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:

$$\phi_t = G\phi_{t-1} + \omega_t \tag{1}$$
$$Y_t = F'\phi_t + \nu_t, \tag{2}$$

where

- Y_t are scalar observations,
- $\phi_t \in \mathbb{R}^{n\times 1}$ is the hidden state,
- $G \in \mathbb{R}^{n\times n}$ is the state transition matrix,
- $F \in \mathbb{R}^{n\times 1}$ is the observation direction.
- ω_t is the process noise, $\mathcal{N}(0, W)$
- ν_t is the observation noise, $\mathcal{N}(0, v)$
- ϕ_0 is initial state, $\mathcal{N}(m_0, C_0)$.

Kalman Filter

- Kalman filter is a key tool for time-series forecasting and analysis.
- An estimate of the current hidden state, given the observations for $t \geq 1$:

$$m_t = \mathbb{E}\left(\phi_t | Y_0, \ldots, Y_t\right), \tag{3}$$

  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:

$$f_{t+1} = \mathbb{E}\left(Y_{t+1} | Y_t, \ldots, Y_0\right) = F' G m_t. \tag{4}$$

Kalman Filter Unrolled

$$f_{t+1} = \underbrace{F' G A_t Y_t + F' \sum_{j=0}^{s-1} \left[\left(\prod_{i=0}^{j} Z_{t-i} \right) G A_{t-j-1} Y_{t-j-1} \right]}_{AR(s+1)} + \underbrace{F' \left(\prod_{i=0}^{s} Z_{t-i} \right) a_{t-s}}_{\text{Remainder term}}. \quad (5)$$

The Results

Theorem (LDS Approximation)

Let $L = L(F, G, v, W)$ be an observable LDS with $W > 0$.

1. *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 $t \geq T_0$,*

$$\left| f_{t+1} - \sum_{i=0}^{s-1} \theta_i Y_{t-i} \right| \leq \varepsilon. \tag{6}$$

2. *For any $\varepsilon, \delta > 0$, and any $B_1 > 0$, there is $T_0 > 0$, $s > 0$ and $\theta \in \mathbb{R}^s$, such that for every sequence Y_t with $|Y_{t+1} - Y_t| \leq B_1$, and for every $t \geq T_0$,*

$$\left| f_{t+1} - \sum_{i=0}^{s-1} \theta_i Y_{t-i} \right| \leq 2 \max\left(\varepsilon, \delta |Y_t|\right). \tag{7}$$

The Results

Is the assumption necessary?

Example

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 $m_{0,ture} > 0$. Since we do not know $m_{0,true}$, we start the Kalman filter with $m_0 = 0$ and initial covariance $C_0 = 1$.

- This is equivalent to estimating the mean 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.

The Results

Key technical result:

Theorem

If the covariance matrix of the process noise is non-zero, then there is $\gamma = \gamma(W, v, F, G) < 1$ *such that for every* $x \in \mathbb{R}^n$,

$$\left[(I - A \otimes F)G'x, (I - A \otimes F)G'x\right] \leq \gamma \left[x, x\right], \tag{8}$$

where $[x, y] = \langle Rx, y \rangle$ *is the inner product induced by the limit* R *of* R_t *on* $\mathbb{R}^n$.

Kalman Filter Unrolled

$$f_{t+1} = \underbrace{F' G A_t Y_t + F' \sum_{j=0}^{s-1} \left[\left(\prod_{i=0}^{j} Z_{t-i} \right) G A_{t-j-1} Y_{t-j-1} \right]}_{AR(s+1)} + \underbrace{F' \left(\prod_{i=0}^{s} Z_{t-i} \right) a_{t-s}}_{\text{Remainder term}}. \quad (9)$$

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

$$-2\left(Y_t - \sum_{i=0}^{s-1} \theta_i Y_{t-i-1}\right)(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-s}). \qquad (10)$$

8: **end for**

The Regret Bound

Let us bound the regret of the algorithm:

Theorem

Let S be a finite family of LDSs, such that every $L = L(F, G, v, W) \in S$, is observable and has $W > 0$. 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 $T > 0$,

$$\sum_{t=0}^{T} \ell_t(\theta_t) - \min_{L \in S} \sum_{t=0}^{T} \ell(Y_t, f_t(L)) \leq C_S + 2(D^2 + B_0^2)\sqrt{T} + \varepsilon T. \qquad (11)$$

What did we Know?

- Subspace identification methods: State sequence of the Kalman filter of an LDS via an SVD of a certain matrix constructed from the inputs.
- Anava et al (COLT 2013): An approximation of a subset of ARMA by AR processes with instance-dependent approximation ratio.
- Liu et al (AAAI 2016): An approximation of a subset of ARIMA by AR processes with instance-dependent approximation ratio.
- Venkatraman et al (AAAI 2016): On-line method related to subspace identification works well.
- Hazan et al (NIPS 2017): Improper learning of LDS conditional on the history of inputs and the most recent observation, by approximating LDS with inputs by AR-like model.

Throughout, guarantees require that the observations are generated from an LDS that is *stationary*. In contrast, we approximate the optimal filter for an arbitrary sequence.

The Experiments

Example (Adapted from (Hazan et al., 2017))

Consider the system:

$$G = \text{diag}([0.999, 0.5]), \quad F' = [1, 1], \tag{12}$$

with process noise distributed as $\omega_t \sim \mathcal{N}(0, w \cdot Id_2)$ and observation noise $\nu_t \sim \mathcal{N}(0, v)$ for different choices of $v, w > 0$.

We will consider the ratio:

$$\frac{\sum_i^{10} \sum_t^T \ell(Y_{i,t}, f_t(L)))}{\sum_i^{10} \sum_t^T \ell(Y_{i,t}, \hat{y}_{i,t}(\theta_{AR(2)}))} \tag{13}$$

where $f_t(L)$ denotes the prediction of the Kalman filter with the ground truth system parameters.

The Experiments

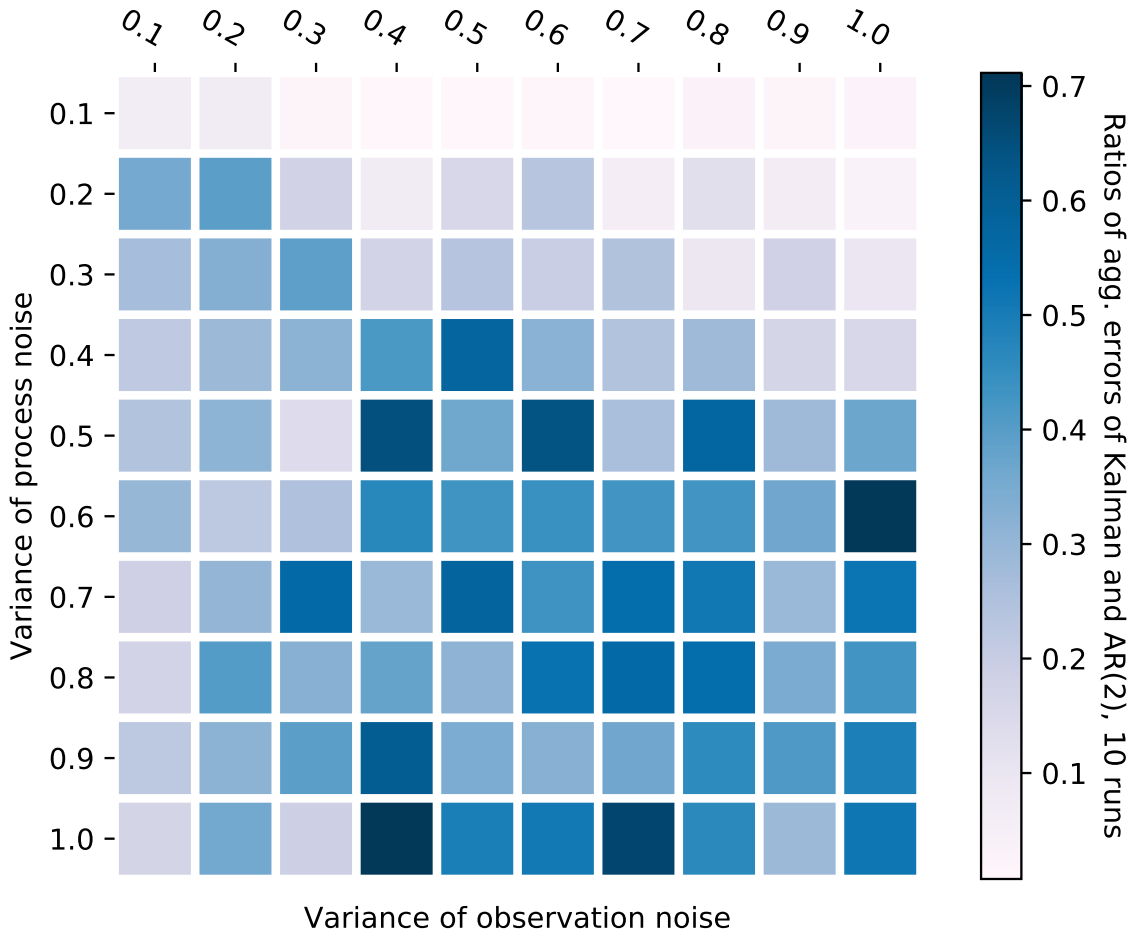


Figure: The ratio of the errors of Kalman filter and AR(2) on Example of Hazan et al, indicated by colours as a function of w, v of process and observation noise, on the vertical and horizontal axes, resp. Origin is the top-left corner.

The Experiments

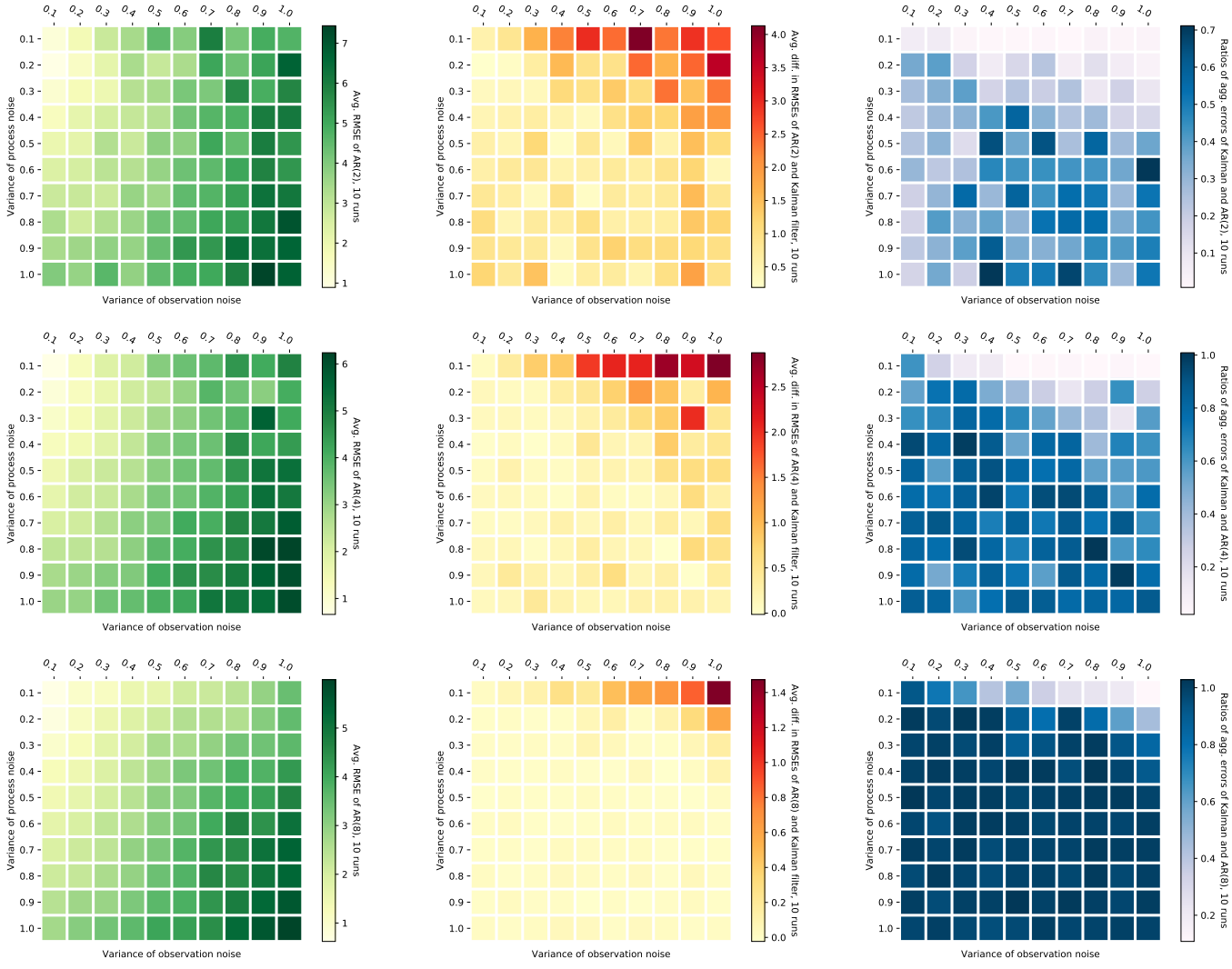


Figure: RMSE, differeces, and ratios for AR(2), AR(4), AR(8).

The Experiments

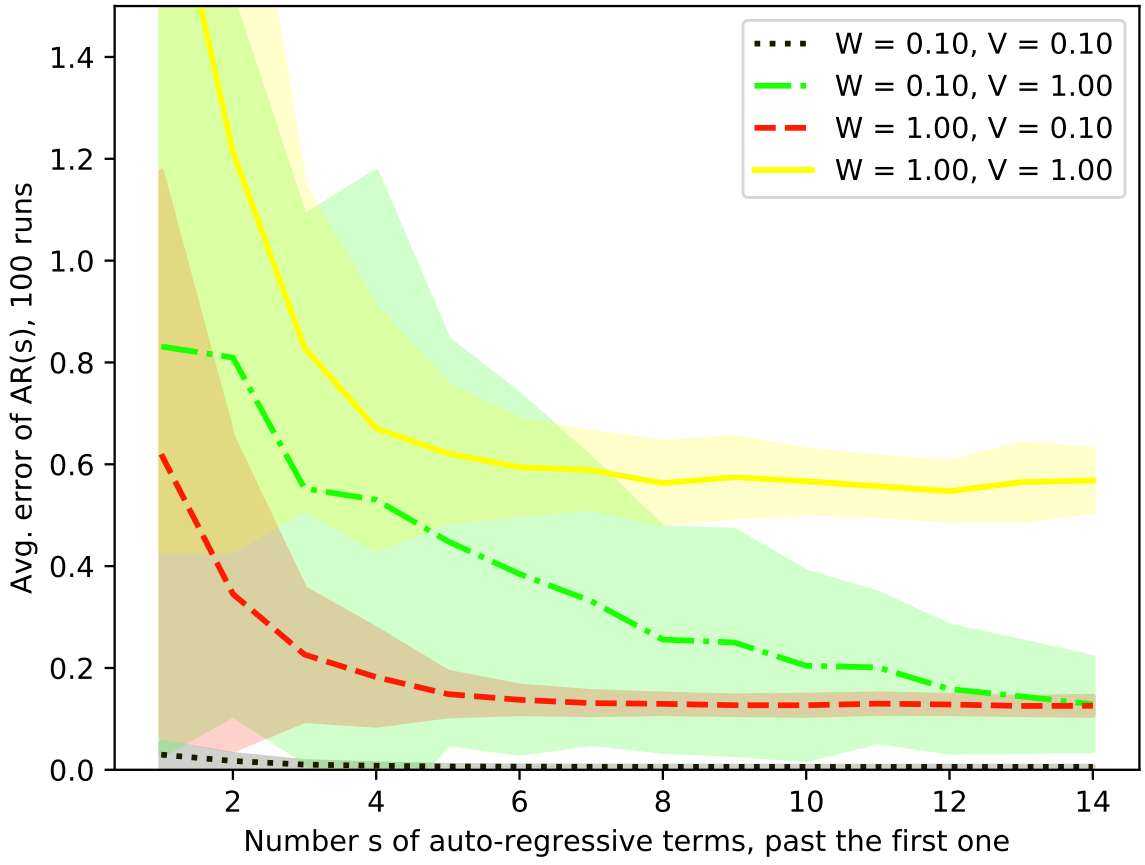


Figure: The error of AR($s+1$) as a function of $s+1$, in terms of the mean and standard deviation over $N = 100$ runs on Example of Hazan et al, for 4 choices of W, v of process, observation noise, respectively.

The Experiments

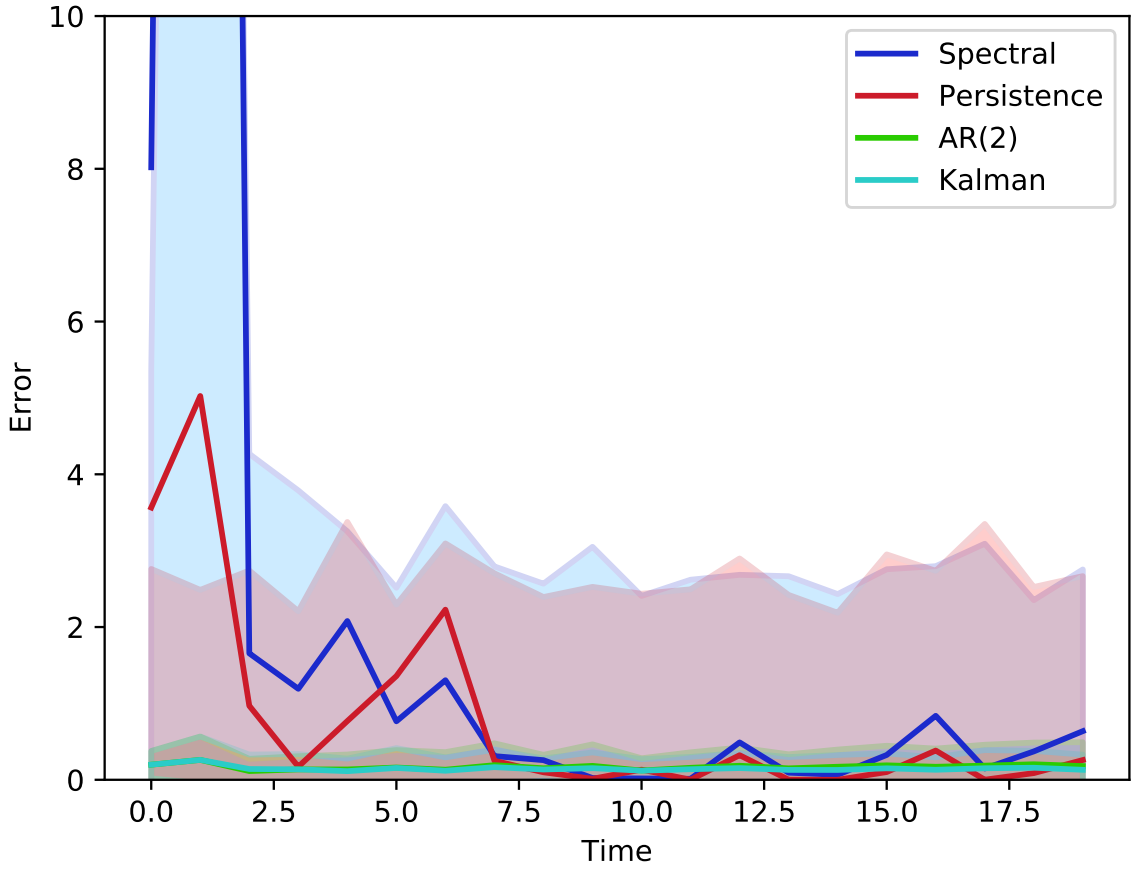


Figure: The error of AR(2) compared against Kalman filter, last-value prediction, and spectral filtering in terms of the mean and standard deviation over $N = 100$ runs on Example of Hazan et al.

The Experiments

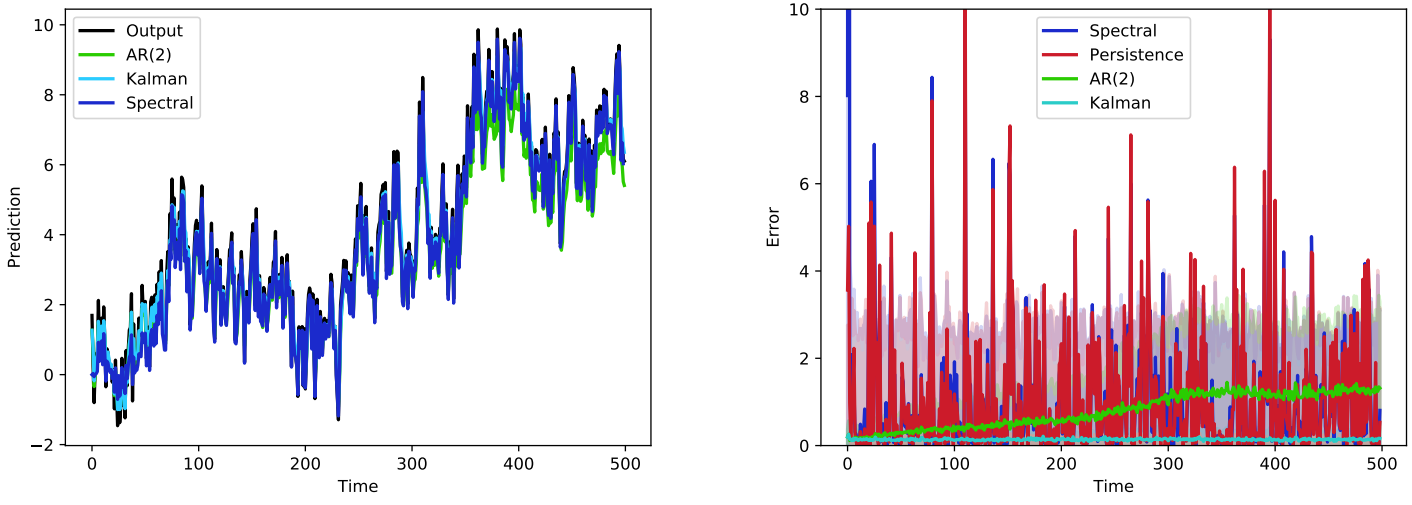


Figure: Illustrations on stock-market data used by Liu et al (AAAI 2016). Left: sample outputs and predictions with AR(2), compared against Kalman filter, last-value prediction, and spectral filtering. Right: Same as Figure 4, over longer time period.

The Conclusions

- A forecasting method applicable to arbitrary sequences.
- First ever regret bound competing against a class of methods including Kalman filters.
- Practical run-time and decent statistical performance.
- A considerable scope for future work.

- For all details, please see `https://arxiv.org/abs/1809.05870`
- For code, go to `https://github.com/jmarecek/OnlineLDS`
- Any questions or comments welcome!

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