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

Mark Kozdoba,[1] **Jakub Marecek,**[2] **Tigran Tchrakian,**[2] **Shie Mannor**[1]
[1]Technion, Israel Institute of Technology, [2]IBM Research, Ireland
markk@technion.ac.il, jakub.marecek@ie.ibm.com,tigran@ie.ibm.com, shie@ee.technion.ac.il

Abstract

The Kalman filter is a key tool for time-series forecasting and analysis. We show that the dependence of a prediction of Kalman filter on the past is decaying exponentially, whenever the process noise is non-degenerate. Therefore, Kalman filter may be approximated by regression on a few recent observations. Surprisingly, we also show that having some process noise is *essential* for the exponential decay. With no process noise, it may happen that the forecast depends on all of the past uniformly, which makes forecasting more difficult.

Based on this insight, we devise an on-line algorithm for improper learning of a linear dynamical system (LDS), which considers only a few most recent observations. We use our decay results to provide the first regret bounds w.r.t. to Kalman filters within learning an LDS. That is, we compare the results of our algorithm to the best, in hindsight, Kalman filter for a given signal. Also, the algorithm is practical: its per-update run-time is linear in the regression depth.

Introduction

Linear Dynamical Systems (LDS) are a key standard tool in modeling and forecasting time series, with an exceedingly large number of applications. In forecasting with an LDS, typically one learns the parameters of the LDS first, using a maximum likelihood principle, and then uses Kalman filter to generate predictions. The two features that seem to contribute the most to the success of LDS in practice are the ability of LDS to model a wide range of behaviors, and the recursive nature of Kalman filter, which allows for fast, real-time forecasts via a constant-time update of the previous estimate. On the other hand, a major difficulty with LDSs is that the process of learning system parameters, via expectation maximization (EM) or direct likelihood optimization, may be time consuming and prone to getting stuck in local maxima. We refer to (Anderson and Moore 1979; West and Harrison 1997; Hamilton 1994; Chui and Chen 2017) for book-length introductions.

Recently, there has been an interest in alternative, *improper* learning approaches, where one approximates the predictions of LDSs by a linear function of a few past observations. The advantage of such approaches is that it convexifies the problem, i.e., learning the linear function amounts to a convex problem, which avoids the issues brought by the non-convex nature of the likelihood function. The convexification allows for on-line algorithms, which are typically fast and simple. A crucial advance of these recent approaches is the guarantee that the predictions of the convexified, improper-learning algorithm are at least as good as the predictions of the proper one. One therefore avoids the long learning times and issues related to non-convexity associated with the classical algorithms, while maintaining the statistical performance.

Leading examples of this approach (Anava et al. 2013; Liu et al. 2016; Hazan, Singh, and Zhang 2017) utilise a framework of *regret bounds* (Cesa-Bianchi and Lugosi 2006) to provide guarantees on the performance of the convexifications. In this framework, one considers a sequence of observations Y_t, with or without additional assumptions. After observing $Y_0, \ldots, Y_t$, an algorithm for improper learning produces a forecast $\hat{Y}_{t+1}$ of the next observation. Then, roughly speaking, one shows that the sum of errors of the forecast thus produced is close to the sum of errors of the best model (in hindsight) from within a certain class. It is said that the algorithm competes against a certain class.

In this paper, we take several steps towards developing guarantees for an algorithm, which competes against Kalman filters. Specifically, we ask what conditions make it possible to model the predictions of Kalman filter as a regression of a few past observations? We show that for a natural, large, and well-known class of LDSs, the *observable* LDSs, the dependence of Kalman filter on the past decays exponentially if the process noise of the LDS is non-degenerate. Consequently, predictions of such LDS can be modeled as auto-regressions. In addition, we show that at least some non-degeneracy of the process noise is *necessary* for the exponential decay. We provide an example with no process noise, where the dependence on the past does not converge exponentially.

Next, based on the decay results, we give an on-line algorithm for time-series prediction and prove regret bounds for it. The algorithm makes predictions in the form $\hat{y}_{t+1} = \sum_{i=0}^{s-1} \theta_i(t) Y_{t-i}$, where Y_t are observations, and $\theta(t) \in \mathbb{R}^s$ is the vector of auto-regression (AR) coefficients, which is updated by the algorithm in an on-line manner.

For any LDS L, denote by $f_{L,t+1}$ the predicted value of Y_{t+1} by Kalman filter corresponding to L, given $Y_t, \ldots, Y_0$.

Denote by $E(T) = \sum_{t=1}^{T-1} |\hat{y}_{t+1} - Y_{t+1}|^2$ the total error made by the algorithm up to time T, and by $E(L,T) = \sum_{t=1}^{T-1} |f_{L,t+1} - Y_{t+1}|^2$ the total error made by Kalman filter corresponding to L. Let S be any finite family of observable linear dynamical systems with non-degenerate process noise. We show that for an appropriate regression depth s, for any bounded sequence $\{Y_t\}_0^T$ we have

$$\frac{1}{T}E(T) \leq \frac{1}{T}\min_{L\in S} E(L,T) + \frac{1}{T}C_S + \varepsilon, \qquad (1)$$

where C_S is a constant depending on the family S. In words, up to an arbitrarily small ε given in advance, the average prediction error of the algorithm is as good or better than the average prediction error of the *best* Kalman filter in S. We emphasize that while there is a dependence on S in the bounds, via the constant C_S, the algorithm itself depends on S only through the regression depth s. In particular, the algorithm does not depend on the cardinality of S, and the time complexity of each iteration is $O(s)$.

To summarize, our contributions are as follows: We show that the dependence of predictions of Kalman filters in a system with non-degenerate process noise is exponentially decaying and that therefore Kalman filters may be approximated by regressions on a few recent observations, cf. Theorem 2. We also show that the process noise is *essential* for the exponential decay. We given an on-line prediction algorithm and prove the first regret bounds against Kalman filters, cf. Theorem 6. Experimentally, we illustrate the performance on a single example in the main body of the text, and further examples in the supplementary material.

Literature

In this section, we review the relevant literature and place the current work in context.

We refer to (Hamilton 1994) for an exposition on LDSs, Kalman filter, and the classical approach to learning the LDS parameters via tha maximum likelihood optimization. See also (Roweis and Ghahramani 1999) for a survey of relations between LDSs and a large variety of other probabilistic models. A general exposition of on-line learning can be found in (Hazan 2016).

As discussed in the Introduction, we are concerned with improper learning, where we show that an alternative model can be shown to generate forecasts that are as good as Kalman filter, up to any given error. Perhaps the first example of an improper learning that is still used today is the moving average, or the exponential moving average (Gardner). In this approach, predictions for a process – of a possibly complex nature – are made using a simple auto-regressive (AR) or AR-like model. This is very successful in a multitude of engineering applications. Nevertheless, until recently, there were very few guarantees for the performance of such methods.

In (Anava et al. 2013), the first guarantees regarding prediction of a (non-AR) subset of auto-regressive moving-average (ARMA) processes by AR processes were given, together with an algorithm for finding the appropriate AR. In (Liu et al. 2016), these results were extended to a subset of autoregressive integrated moving average (ARIMA) processes, while at the same time the assumptions on the underlying ARMA model were relaxed.

In this paper, we show that AR models may also be used to forecast as well as Kalman filters. One major difference between our results and the previous work is that we obtain approximation results on *arbitrary* bounded sequences. Indeed, regret results of (Anava et al. 2013) and (Liu et al. 2016) only hold under the assumption that the data sequence was generated by a particular fixed ARMA or ARIMA process. Moreover, the constants in the regret bounds of (Anava et al. 2013) and (Liu et al. 2016) depend on the generating model, and the guaranteed convergence may be arbitrarily slow, even when the sequence to forecast is generated by appropriate model.

In contrast, we show that up to an arbitrarily small error given in advance, AR(s) will perform as well as Kalman filter on any bounded sequence. We also obtain approximation results in the more general case of bounded difference sequences.

Another related work is (Hazan, Singh, and Zhang 2017), which addresses a different aspect of LDS approximation by ARs. In the case of LDSs with *inputs*, building on known eigenvalue-decay estimates of Hankel matrices, it is shown that the influence of all past inputs may be effectively approximated by an AR-type model. However, the arguments and the algorithms in (Hazan, Singh, and Zhang 2017) were not designed to address model noise. In particular, the algorithm of (Hazan, Singh, and Zhang 2017) makes predictions based on the whole history of inputs and on only one most recent observation, Y_t, and hence clearly can not compete with Kalman filters in situations with no inputs. We demonstrate this in the Experiments section.

Previously, subspace identification methods (van Overschee and de Moor 1996) achieved significant advances in the learning of LDS. For a sequence of observations generated from an LDS, this family of methods allows to recover the state sequence of the Kalman filter of the LDS via a singular value decomposition (SVD) of a certain matrix constructed from the inputs. In a naive implementation, this requires the SVD to be performed at each time step, on a matrix constructed from the *full* history of observations. (Venkatraman et al. 2016) proposed an on-line method related to subspace identification using the notion of instrumental variables. While the experimental part of the paper deals with LDS forecasting, the provided theoretical guarantees apply only in cases of independent observations, which is not the case for LDSs. More broadly, guarantees we are aware of require that the observations are generated from an LDS that is *stationary*, in contrast to finding the optimal filter for a given arbitrary sequence. This therefore excludes most tracking applications.

Preliminaries

As usual in the literature (West and Harrison 1997), we define a linear system $L = (G, F, v, W)$ as:

$$\phi_t = G\phi_{t-1} + \omega_t \qquad (2)$$

$$Y_t = F'\phi_t + \nu_t, \qquad (3)$$

where Y_t are scalar observations, and $\phi_t \in \mathbb{R}^{n\times 1}$ is the hidden state. $G \in \mathbb{R}^{n\times n}$ is the state transition matrix which defines the system dynamics, and $F \in \mathbb{R}^{n\times 1}$ is the observation direction. The process-noise terms ω_t and observation-noise terms ν_t are mean zero normal independent variables. For all $t \geq 1$ the covariance of ω_t is W and the variance of ν_t is v. The initial state ϕ_0 is a normal random variable with mean m_0 and covariance C_0.

For $t \geq 1$ denote

$$m_t = \mathbb{E}\left(\phi_t | Y_0, \ldots, Y_t\right), \tag{4}$$

and let C_t be the covariance matrix of ϕ_t given $Y_0, \ldots, Y_t$. Note that m_t is the estimate of the current hidden state, given the observations. Further, the central quantity of this paper is

$$f_{t+1} = \mathbb{E}\left(Y_{t+1} | Y_t, \ldots, Y_0\right) = F' G m_t. \tag{5}$$

This is the forecast of the next observation, given the current data. The quantities m_t and f_{t+1} are known as Kalman Filter. In particular, in this paper we refer to the sequence f_t as the Kalman filter associated with the LDS $L = (G, F, v, W)$.

The Kalman filter satisfies the following recursive update equations: Set

$$\begin{aligned} a_t &= G m_{t-1} \\ R_t &= G C_{t-1} G' + W \\ Q_t &= F' R_t F + v \\ A_t &= R_t F / Q_t \end{aligned}$$

Note that in this notation we have

$$f_t = F' a_t.$$

Then the update equations of Kalman filter are:

$$m_t = a_t + A_t(Y_t - f_t) = A_t Y_t + (I - F \otimes A_t) a_t \tag{6}$$

$$C_t = R_t - A_t Q_t A_t' \tag{7}$$

where $x \otimes y$ is an $\mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times 1}$ operator which acts by $z \mapsto \langle z, x\rangle\, y = y x' z$. The matrix of $x \otimes y$ is given by the outer product yx', where $x, y \in \mathbb{R}^{n\times 1}$.

An important property of Kalman Filter is that while m_t depends on $Y_0, \ldots, Y_t$, the covariance matrix C_t does not. Indeed, note that R_t, Q_t, A_t, C_t are all deterministic sequences which do not depend on the observations.

We explicitly write the recurrence relation for R_t:

$$R_{t+1} = G\left(R_t - \frac{R_t F \otimes R_t F}{\langle F, R_t F\rangle + v}\right) G' + W \tag{8}$$

Also write for convenience

$$a_{t+1} = G m_t = G A_t Y_t + G(I - F \otimes A_t) a_t. \tag{9}$$

A more explicit form of the prediction of Y_{t+1} given $Y_t, \ldots, Y_0$, may be obtained by unrolling (6) and using (9):

$$\mathbb{E}\left(Y_{t+1} | Y_t, \ldots, Y_0\right) = f_{t+1} = F' a_{t+1} \tag{10}$$

$$= F' G A_t Y_t + F' G (I - F \otimes A_t) a_t \tag{11}$$

$$\begin{aligned} = {} & F' G A_t Y_t + F' G (I - F \otimes A_t) G A_{t-1} Y_{t-1} \\ & + F' G (I - F \otimes A_t) G (I - F \otimes A_{t-1}) a_{t-1} \end{aligned} \tag{12}$$

In general, set $Z_t = G(I - F \otimes A_t)$ and $Z = G(I - F \otimes A)$. Chose and fix some $s \geq 1$. Then for any $t \geq s+1$, the expectation (10) has the form displayed in Figure 1.

Next, a linear system $L = (G, F, v, W)$ is said to be *observable*, (West and Harrison 1997), if

$$\operatorname{span}\left\{F, G'F, \ldots, G'^{n-1} F\right\} = \mathbb{R}^n. \tag{14}$$

Roughly speaking, the pair (G, F) is observable if the state can be recovered from a sufficient number of observations, in a noiseless situation. Note that if there were parts of the state that do not influence the observations, these parts would be irrelevant for forecast purposes. Thus we are only interested in observable LDSs.

When L is observable, it is known (Harrison 1997) that the sequences C_t, R_t, Q_t, A_t converge. See also (Anderson and Moore 1979; West and Harrison 1997). We denote the limits by C, R, Q and A respectively. Moreover, the limits satisfy the recursions as equalities. In particular we have

$$R = G\left(R - \frac{RF \otimes RF}{\langle F, RF\rangle + v}\right) G' + W. \tag{15}$$

Finally, an operator $P : \mathbb{R}^n \to \mathbb{R}^n$ is *non-negative*, denoted $L \geq 0$, if $\langle Px, x\rangle \geq 0$ for all $x \neq 0$, and is *positive*, denoted $P > 0$, if $\langle Px, x\rangle > 0$ for all $x \neq 0$. Note that W, C_t, R_t, C, R are either covariance matrices or limits of such matrices, and thus are symmetric and non-negative.

Exponential Decay and AR Approximation

In what follows, we denote by

$$[x, y] = \langle Rx, y\rangle, \quad \langle\langle x, y\rangle\rangle = \langle Wx, y\rangle \tag{16}$$

the inner products induced by R and W on $\mathbb{R}^n$, where R is the limit of R_t as described above. In particular, we set $U = G'$ and rewrite (15) as

$$[x, y] = [Ux, Uy] - \frac{[Ux, F]\,[Uy, F]}{[F, F] + v} + \langle\langle x, y\rangle\rangle. \tag{17}$$

Observe that since $R = GCG' + W$, we have $R \geq W$, and in particular if $W > 0$ then $R > 0$. In other words, if $W > 0$, then $[\cdot, \cdot]$ and $\langle\langle \cdot, \cdot\rangle\rangle$ induce proper norms on $\mathbb{R}^n$:

$$[x, x] \geq \langle\langle x, x\rangle\rangle > 0 \text{ for all } x \neq 0. \tag{18}$$

Next, consider the remainder term in the prediction equation (13), where we have replaced Z_{t-i} with their limit values Z:

$$\begin{aligned} F' & \left(G(I - F \otimes A)\right)^{s+1} a_{t-s} \\ & = \left\langle F, \left(G(I - F \otimes A)\right)^{s+1} a_{t-s}\right\rangle \\ & = \left\langle \left((I - A \otimes F)U\right)^{s+1} F, a_{t-s}\right\rangle. \end{aligned}$$

Let us now state and prove the key result of this paper: if $W > 0$, then $\left((I - A \otimes F)U\right)^s F$ converges to zero exponentially fast with s. The key to the proof will be to consider contractivity properties with respect to the norm induced by $[\cdot, \cdot]$, rather than with respect to the the default inner product.

$$f_{t+1} = \underbrace{F'GA_tY_t + F'\sum_{j=0}^{s-1}\left[\left(\prod_{i=0}^{j} Z_{t-i}\right) GA_{t-j-1}Y_{t-j-1}\right]}_{AR(s+1)} + \underbrace{F'\left(\prod_{i=0}^{s} Z_{t-i}\right) a_{t-s}}_{\text{Remainder term}}. \tag{13}$$

Figure 1: The unrolling of the forecast f_{t+1}. The remainder term goes to zero exponentially fast with s, by Lemma 3.

Theorem 1. *If $W > 0$, then there is $\gamma = \gamma(W, v, F, G) < 1$ such that for every $x \in \mathbb{R}^n$,*

$$[(I - A \otimes F)Ux, (I - A \otimes F)Ux] \leq \gamma [x, x]. \tag{19}$$

Proof. Set

$$y = ((I - A \otimes F)U) x. \tag{20}$$

Then

$$y = (I - A \otimes F)Ux = Ux - \langle A, Ux\rangle F \tag{21}$$

$$= Ux - \frac{[Ux, F]}{[F, F] + v}F. \tag{22}$$

Therefore we have

$$[y, y] = [Ux, Ux] - 2\frac{[Ux, F]^2}{[F, F] + v} + \frac{[Ux, F]^2 [F, F]}{([F, F] + v)^2}. \tag{23}$$

In addition, by (17),

$$[Ux, Ux] = [x, x] + \frac{[Ux, F]^2}{[F, F] + v} - \langle\langle x, x\rangle\rangle. \tag{24}$$

Combining (23) and (24), we obtain

$$\begin{aligned}[y, y] &= [x, x] - \langle\langle x, x\rangle\rangle - \frac{[Ux, F]^2}{[F, F] + v}\left(1 - \frac{[F, F]}{[F, F] + v}\right) \\ &= [x, x] - \langle\langle x, x\rangle\rangle - \frac{[Ux, F]^2}{[F, F] + v}\frac{v}{[F, F] + v}.\end{aligned} \tag{25}$$

Equation (25) immediately implies that $[x, x]$ is non-increasing. Recall that by (18), W is dominated by R. However, since both R and W define proper norms, by the equivalence of finite dimensional norms, the inverse inequality is also true: There exists $0 < \kappa \leq 1$ such that

$$\langle\langle x, x\rangle\rangle \geq \kappa [x, x] \text{ for all } x \neq 0. \tag{26}$$

Therefore the decrease in (25) must be exponential:

$$[y, y] \leq [x, x] - \langle\langle x, x\rangle\rangle \leq (1 - \kappa) [x, x]. \tag{27}$$

□

It is of interest to stress the fact that Theorem 1 does not assume any contractivity properties of G. In particular, the very common assumption of the spectral radius of G being bounded by 1 is not required.

Let us state and prove our main approximation result:

Theorem 2 (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 $t \geq T_0$,*

$$\left| f_{t+1} - \sum_{i=0}^{s-1} \theta_i Y_{t-i} \right| \leq \varepsilon. \tag{28}$$

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 $t \geq T_0$,*

$$\left| f_{t+1} - \sum_{i=0}^{s-1} \theta_i Y_{t-i} \right| \leq 2\max(\varepsilon, \delta |Y_t|). \tag{29}$$

We first prove the bound on the remainder term in the prediction equation (13).

Lemma 3 (Remainder-Term Bound)**.** *Let $L = L(F, G, v, W)$ be an observable LDS with $W > 0$.*

1. *If a sequence Y_t satisfies $|Y_t| \leq B_0$ for all $t \geq 0$, then there are constants $\rho'_L < 1$ and c_L such that for any $s > 0$ and $t > s$,*

$$\left| \left\langle F, \left(\prod_{i=0}^{s} Z_{t-i}\right) a_{t-s} \right\rangle \right| \leq (\rho'_L)^s c_L. \tag{30}$$

2. *If a sequence Y_t satisfies $|Y_{t+1} - Y_t| \leq B_1$ for all $t \geq 0$, then there are constants ρ'_L and $c_{1,L}, c_{2,L}$ such that for all $s > 0$ and $t > s$,*

$$\begin{aligned}&|\langle F, (\textstyle\prod_{i=0}^{s} Z_{t-i})\, a_{t-s}\rangle| \leq \\ &(\rho'_L)^s c_{1,L} (|Y_t| + sB_1 + c_{2,L}).\end{aligned} \tag{31}$$

Proof. Recall that a_t satisfies the recursion (9),

$$a_{t+1} = G(I - F \otimes A_t)a_t + Y_t A_t = Z_t a_t + Y_t G A_t. \tag{32}$$

Denote by $[x] = [x, x]^{\frac{1}{2}}$ and and by $|x| = \langle x, x\rangle^{\frac{1}{2}}$ the norms induced by $[\cdot, \cdot]$ and $\langle \cdot, \cdot \rangle$ respectively. Set $P = Z'$ and $P_t = Z'_t$. By Theorem 1, there is $\rho = \gamma^{\frac{1}{2}} < 1$ such that P is a ρ-contraction with respect to $[\cdot]$. Fix some ρ' such that $\rho < \rho' < 1$. Since $P_t \to P$, there is some T_1 such that for all $t \geq T_1$, P_t is a ρ'-contraction. In addition, let T_2 be such that $[GA - GA_t] \leq 1$ for all $t \geq T_2$. Set $T_0 = \max(T_1, T_2) + 1$. Fix $s > 0$ and set $t' = t - s - 1$. For $t' > T_0$, using (32) write a_{t-s} as

$$\begin{aligned}a_{t'+1} &= Y_{t'} G A_{t'} \\ &+ \sum_{i=0}^{t'-T_0} \left(Y_{t'-i-1} \left(\prod_{j=0}^{i} Z_{t'-j} \right) G A_{t'-i-1} \right) \\ &+ \left(\prod_{j=0}^{t'-T_0} Z_{t'-j} \right) a_{T_0 - 1}.\end{aligned} \tag{33}$$

Observe that if an operator O' is a γ-contraction with respect to $[\cdot]$, then for any $x, y \in \mathbb{R}^n$,

$$\begin{aligned}\langle y, Ox\rangle &= \langle O'y, x\rangle \\ &\le |O'y|\,|x| \le \gamma\mu[y]\,|x| \le \gamma\mu^2[y][x],\end{aligned} \tag{34}$$

where μ is the equivalence constant between $[\cdot]$ and $|\cdot|$.

For every $x \in \mathbb{R}^n$ by (33) we have

$$\begin{aligned}\langle x, a_{t-s}\rangle &= \\ &= Y_{t'}\,\langle x, GA_{t'}\rangle + \\ &\sum_{i=0}^{t'-T_0} Y_{t'-i-1}\left\langle \left(\prod_{j=i}^{0} P_{t'-j}\right) x, GA_{t'-i-1}\right\rangle \\ &+ \left\langle \left(\prod_{j=t'-T_0}^{0} P_{t'-j}\right) x, a_{T_0-1}\right\rangle.\end{aligned} \tag{35}$$

By the choice of T_0, as since the expansion in (35) is only up to T_0, every $P_{t'-j}$ in (35) is a ρ'-contraction and all $GA_{t'-j}$ satisfy $[GA - GA_{t'-j}] \le 1$.

Combining this with (34) and using triangle inequality, we obtain

$$\begin{aligned}|\langle x, a_{t-s}\rangle| &\le \\ &= |Y_{t'}|\,\mu^2[x]\,([GA]+1) + \\ &+ \sum_{i=0}^{t'-T_0} |Y_{t'-i-1}|\,(\rho')^{i+1}\mu^2[x]\,([GA]+1) \\ &+ (\rho')^{t'-T_0}\mu^2[x][a_{T_0-1}].\end{aligned} \tag{36}$$

Finally, choose $x = \left(\prod_{i=s}^{0} P_{t-i}\right) F$. Note that $[x] \le (\rho')^{s+1}[F]$. Therefore,

$$\left|\left\langle F, \left(\prod_{i=0}^{s} Z_{t-i}\right) a_{t-s}\right\rangle\right| = \langle x, a_{t-s}\rangle \tag{37}$$

$$\le (\rho')^{s+1}\,|Y_{t'}|\,\mu^2[F]\,([GA]+1) + \tag{38}$$

$$+ \sum_{i=0}^{t'-T_0} (\rho')^{s+1}\,|Y_{t'-i-1}|\,(\rho')^{i+1}\mu^2[F]\,([GA]+1) \tag{39}$$

$$+ (\rho')^{s+1}(\rho')^{t'-T_0}\mu^2[F][a_{T_0-1}]. \tag{40}$$

Observe that the term $[a_{T_0-1}]$ in (40) is a constant, independent of t, and that the series in (39) are summable w.r.t t'. Therefore, in the bounded case $|Y_t| \le B_0$, the proof is complete.

In the Lipschitz case, for every $i > 0$, we have

$$|Y_{t'-i-1}| \le |Y_{t'}| + (i+1)B_1. \tag{41}$$

Substituting this into (37)-(40), and observing that the resulting series are still summable, we obtain

$$\left|\left\langle F, \left(\prod_{i=0}^{s} Z_{t-i}\right) a_{t-s}\right\rangle\right| \le (\rho')^s c_1\,(|Y_{t'}| + c_2)\,. \tag{42}$$

Thus using

$$|Y_{t'}| \le |Y_t| + sB_1, \tag{43}$$

completes the proof in the Lipschitz case. □

We now prove Theorem 2.

Proof. Recall that f_{t+1} is given by (13). Fix some $s > 0$ and set $\theta_0 = \langle F, GA\rangle$, and $\theta_{j+1} = \left\langle F, Z^{j+1}GA\right\rangle$ for $j = 0, \ldots, s-1$. Note that $\theta \in \mathbb{R}^{s+1}$ and s here corresponds to $s+1$ in the statement of the Theorem. Set also $r_t = \langle F, GA_t\rangle$ and for $j \ge 0$, $r_{t-j-1} = \left\langle F, \left(\prod_{i=0}^{j} Z_{t-i}\right) GA_{t-j}\right\rangle$. Clearly $r_t \to \theta_0$ with t and $r_{t-j-1} \to \theta_{j+1}$ for every fixed j. Next, using Lemma 3, the discrepancy between f_{t+1} and the θ predictor is given by

$$\begin{aligned}&\left|f_{t+1} - \sum_{j=0}^{s} Y_{t-j}\theta_j\right| \le \\ &|Y_t|\,|r_t - \theta_0| + \sum_{j=0}^{s-1} |Y_{t-j-1}|\,|r_{t-j-1} - \theta_{j+1}| + (\rho'_L)^s c_L\end{aligned} \tag{44}$$

in the bounded case. In this case, therefore, choosing regression depth s large enough so that $(\rho'_L)^s c_L \le \varepsilon/2$ and T_0 large enough so that for all $t \ge T_0$, $|r_{t-j-1} - \theta_{j+1}| \le \frac{\varepsilon}{2sB_0}$ for all $j \le s$, suffices to conclude the proof. The proof of the Lipschitz case follows similar lines and is given in the Supplementary Material due to space constraints. □

To conclude this section, we discuss the relation between exponential convergence and the non-degenerate noise assumption, $W > 0$. Note that the crucial part of Theorem 1, inequality (26), holds if and only if we can guarantee that $\langle\langle x, x\rangle\rangle > 0$ for every x for which $[x, x] > 0$. In particular, this holds when $W > 0$ – that is, the noise is full dimensional. We now demonstrate that at least some noise is *necessary* for the exponential decay to hold.

Consider first a one dimensional example.

Example 4. *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 $C_0 = 1$.*

In this case, clearly the observations Y_t are independent samples of a fixed distribution with mean $m_{0,true}$ and variance v. The Kalman filter in this situation is equivalent to a Bayesian mean estimator with prior distribution $N(0, C_0 = 1)$. From general considerations, it follows that $R_t \to R = 0$ with t. Indeed, if we start with $C_0 = 0$, then we have $R_t = 0$ for all t. Since the limit R does not depend on the initialization, (Harrison 1997), we have $R = 0$ for every initialization. As a side note, in this particular case it can be shown, either via the Bayesian interpretation or directly, that R_t decays as $1/t$ (that is, $tR_t \to const$, with t). Now, note that $Z_t = 1 - \frac{R_t}{R_t+v} = \frac{v}{R_t+v} \to 1$, and that for any fixed $j > 0$, $A_{t-j} \to 0$ as t grows. Next, for fixed $s > 0$, consider the prediction equation (13). On the one hand, we know that f_{t+1} converges to $m_{0,true} > 0$ in probability. This is clear for instance from the Bayesian estimator interpretation above. On the other hand, the coefficients of all Y_{t-j} in (13) converge to 0. It follows therefore, that the remainder

term in (13), $F'\left(\prod_{i=0}^{s} Z_{t-i}\right) a_{t-s}$, converges in probability to $m_{0,true}$ as $t \to \infty$. In particular, the remainder term does *not* converge to 0. This is in sharp contrast with the exponential convergence of this term to zero in the $W > 0$ case, as given by Lemma 3.

The above example can be generalized as follows:

Example 5. *In any dimension n, let (G, F) define an LDS such that G is a rotation, and such that G, F is observable. Again choose $W = 0$ and $v > 0$. As before, let the true process start from a state $m_{0,true} \neq 0$ and start the filter with $m_0 = 0$ and $C_0 = \mathrm{Id}$.*

Considerations similar to those of the previous example imply that $R_t \to 0$ but f_{t+1} does not. Consequently, the remainder term will not converge to zero.

An Algorithm and Regret Bounds

In this section, we introduce our prediction algorithm and prove the associated regret bounds. Our on-line algorithm maintains a state estimate, which is represented by the regression coefficients $\theta \in \mathbb{R}^s$, where s is the regression depth, a parameter of the algorithm. At time step t, the algorithm first produces a prediction of the observation Y_t, using the current state θ and previous observations, $Y_{t-1}, \ldots, Y_0$. Specifically, we will predict Y_t by

$$\hat{y}_t(\theta) = \sum_{i=0}^{s-1} \theta_i Y_{t-i-1}. \tag{45}$$

After the prediction is made, the true observation Y_t is revealed to the algorithm, and a loss associated with the prediction is computed. Here we consider the quadratic loss for simplicity: We define $\ell(x, y)$ as $(x - y)^2$. The loss function at time t will be given by

$$\ell_t(\theta) := \ell(Y_t, \hat{y}_t(\theta)). \tag{46}$$

In addition, the state is updated. We use the general scheme of on-line gradient decent algorithms, (Zinkevich 2003), where the update goes against the direction of the gradient of the current loss. In addition, it is useful to restrict the state to a bounded domain. We will use a Euclidean ball of radius D as the domain, where D is a parameter of the algorithm. We denote this domain by $\mathcal{D} = \{x \in \mathbb{R}^s \mid |x| \leq D\}$ and denote by $\pi_{\mathcal{D}}$ the Euclidean projection onto this domain. If the gradient step takes the state outside of the domain, the state is projected back onto $\mathcal{D}$. The pseudo-code is presented in Algorithm 1, where the gradient $\nabla_\theta \ell_t(\theta)$ of the cost at θ at time t is given by

$$-2\left(Y_t - \sum_{i=0}^{s-1} \theta_i Y_{t-i-1}\right)(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-s}). \tag{47}$$

Note a slight abuse of notation in Algorithm 1: the vector $\theta_t \in \mathbb{R}^s$ denotes the state at time t, while in (45) and elsewhere in the text, θ_i denotes the scalar coordinates of θ. Whether the vector or the coordinates are considered will always be clear from context.

For any LDS L, let $f_t(L)$, defined by (13), be the prediction of Y_t that Kalman filter associated with L makes, given $Y_{t-1}, \ldots, Y_0$. We start all filters with the initial state $m_0 = 0$, and initial covariance $C_0 = \mathrm{Id}_s$, the $s \times s$ identity matrix. Let S be any family of LDSs. Then for any sequence $\{Y_t\}_0^T$, the quantity

$$\sum_{t=0}^{T} \ell(\theta_t) - \min_{L \in S} \sum_{t=0}^{T} \ell(Y_t, f_t(L)), \tag{48}$$

where θ_t are the sequence of states produced by Algorithm 1, is called the *regret*. As discussed in the introduction, $\sum_{t=0}^{T} \ell(\theta_t)$ is the total error incurred by the algorithm, and $\min_{L \in S} \sum_{t=0}^{T} \ell(Y_t, f_t(L))$ is the loss of the best (in hindsight) Kalman filter in S. Therefore, small regret means that the algorithm performs on sequence $\{Y_t\}_0^T$ as well as the best Kalman filter in S, even if we are allowed to select that Kalman filter in hindsight, after the whole sequence is revealed.

In the Supplementary Material, we prove the following bound on the regret of Algorithm 1:

Theorem 6. *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 Algorithm 1 with parameters s and D, then for every $T > 0$,

$$\begin{aligned} &\textstyle\sum_{t=0}^{T} \ell_t(\theta_t) - \min_{L \in S} \sum_{t=0}^{T} \ell(Y_t, f_t(L)) \\ &\qquad \leq C_S + 2(D^2 + B_0^2)\sqrt{T} + \varepsilon T. \end{aligned} \tag{49}$$

Due to the limited space in the main body of the text, we describe only the main ideas of the proof here. Similarly to other proofs in this domain, it consists of two steps. In the first step we show that

$$\sum_{t=0}^{T} \ell_t(\theta_t) - \min_{\phi \in \mathcal{D}} \sum_{t=0}^{T} \ell(Y_t, \hat{y}_t(\phi)) \leq 2(D^2 + B_0^2)\sqrt{T}. \tag{50}$$

This means that Algorithm 1 performs as well as the best in hindsight *fixed* state vector ϕ. This follows from the general results in (Zinkevich 2003). In the second step, we use the approximation Theorem 2 to find for each $L \in S$ an appropriate $\theta_L \in \mathcal{D}$, such that the predictions $f_{t,L}$ are approximated by $\hat{y}_t(\theta_L)$. It follows from this step, that the best

Algorithm 1 On-line Gradient Descent

Input: Regression length s, domain bound D.
Observations $\{Y_t\}_0^\infty$, given sequentially.
Set the learning rate $\eta_t = t^{-\frac{1}{2}}$.
Initialize θ_s arbitrarily in $\mathcal{D}$.
for $t = s$ **to** ∞ **do**
  Predict $\hat{y}_t = \sum_{i=0}^{s-1} \theta_{t,i} Y_{t-i-1}$
  Observe Y_t and compute the loss $\ell_t(\theta_t)$ of (46)
  Update $\theta_{t+1} \leftarrow \pi_{\mathcal{D}}\left(\theta - \eta_t \nabla \ell_t(\theta_t)\right)$ using (47)
end for

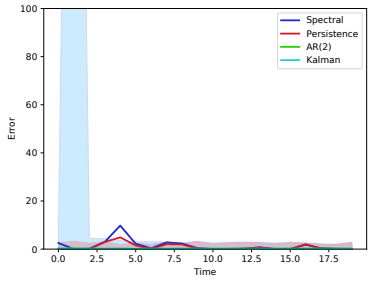

Figure 2: The error of AR(2) compared against Kalman filter, last-value prediction, and spectral filtering in terms of the mean and standard deviation over $N = 100$ runs on Example 7.

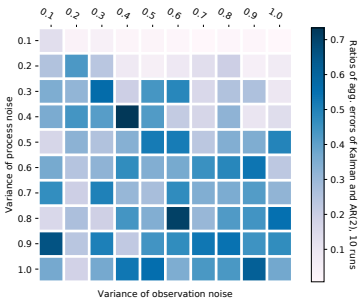

Figure 3: The ratio of the errors of Kalman filter and AR(2) on Example 7 indicated by colours as a function of w, v of process and observation noise, on the vertical and horizontal axes, resp. Origin is the top-left corner.

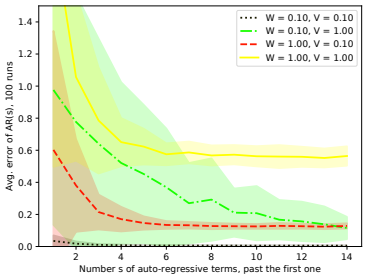

Figure 4: The error of AR($s + 1$) as a function of $s + 1$, in terms of the mean and standard deviation over $N = 100$ runs on Example 7, for 4 choices of W, v of process and observation noise, respectively.

Kalman filter performs approximately as well as the best θ_L. Specifically, we have

$$\min_{L\in S}\sum_{t=0}^{T}\ell(Y_t,\hat{y}_t(\theta_L)) \le \min_{L\in S}\sum_{t=0}^{T}\ell(Y_t,f_t(L)) + \varepsilon T \quad (51)$$

Because by construction $\theta_L \in \mathcal{D}$, clearly it holds that

$$\min_{\phi\in\mathcal{D}}\sum_{t=0}^{T}\ell(Y_t,\hat{y}_t(\phi)) \le \min_{L\in S}\sum_{t=0}^{T}\ell(Y_t,\hat{y}_t(\theta_L)),$$

and therefore combining (50) and (51) yields the statement of Theorem 6.

Experiments

To illustrate our results, we present experiments on a few well-known examples in the Supplementary Material. Out of those, we chose one to present here:

Example 7 (Adapted from (Hazan, Singh, and Zhang 2017))**.** *Consider the system:*

$$G = \mathrm{diag}([0.999, 0.5]), \quad F' = [1, 1], \qquad (52)$$

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* $v, w > 0$.

In Figure 2, we compare the prediction error for 4 methods: the standard baseline last-value prediction $\hat{y}_{t+1} := y_t$, also known as persistence prediction, the spectral filtering of (Hazan, Singh, and Zhang 2017), Kalman filter, and AR(2). Here AR(2) is the truncation of Kalman filter, given by (13) with regression depth $s = 1$ and no remainder term. Average error over 100 observation sequences generated by (52) with $v = w = 0.5$ is shown as solid line, and its standard deviation is shown a as shaded region. Note that from some time on, spectral filtering essentially performs persistence prediction, since the inputs are zero. Further, note that both Kalman filter and AR(2) considerably improve upon the performance of last- value prediction.

In Figure 3, we compare the performance of AR(2) and Kalman filter under varying magnitude of noises v, w. In particular, colour indicates the ratio of the errors of Kalman filter to the errors of AR(2), wherein the errors are the average prediction error over 10 trajectories of (52) for each cell of the heat-map, with each trajectory of length 50. (The formula is given in the Supplementary Material.) Consistent with our analysis, one can observe that increasing the variance of process noise improves the approximation of the Kalman filter by AR(2).

Finally, in Figure 4, we illustrate the decay of the remainder term by presenting the mean (line) and standard deviation (shaded area) of the error as a function of the regression depth s. There, 4 choices of the covariance matrix W of the process noise and the variance v of the observation noise are considered within Example 7 and the error is averaged over $N = 100$ runs of length $T = 200$. Of course, as expected, increasing s decreases the error, until the error approaches that of the Kalman filter. Observe again that for a given value of the observation noise, the convergence w.r.t s is slower for *smaller* process noise, consistently with our theoretical observations.

Conclusions

We have presented a forecasting method, which is applicable to arbitrary sequences and comes with a regret bound com-

peting against a class of methods, which includes Kalman filters.

We hope that our algorithms and Python code available from https://github.com/jmarecek/OnlineLDS will spur further research in forecasting and system identification.

Acknowledgments

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

References

Anava, O.; Hazan, E.; Mannor, S.; and Shamir, O. 2013. Online learning for time series prediction. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*.

Anderson, B., and Moore, J. 1979. *Optimal Filtering*. Prentice Hall.

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

Chui, C., and Chen, G. 2017. *Kalman Filtering: with Real-Time Applications*. Springer International Publishing.

Gardner, E. S. Exponential smoothing: The state of the art. *Journal of Forecasting* 4(1):1–28.

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

Harrison, P. J. 1997. Convergence and the constant dynamic linear model. *Journal of Forecasting* 16(5).

Hazan, E.; Singh, K.; and Zhang, C. 2017. Online learning of linear dynamical systems. In *Advances in Neural Information Processing Systems*, 6686–6696.

Hazan, E. 2016. Introduction to online convex optimization. *Found. Trends Optim.*

Liu, C.; Hoi, S. C. H.; Zhao, P.; and Sun, J. 2016. Online arima algorithms for time series prediction. AAAI'16.

Roweis, S., and Ghahramani, Z. 1999. A unifying review of linear gaussian models. *Neural Computation* 11(2):305–345.

van Overschee, P., and de Moor, L. 1996. *Subspace identification for linear systems: theory, implementation, applications*. Kluwer Academic Publishers.

Venkatraman, A.; Sun, W.; Hebert, M.; Bagnell, J.; and Boots, B. 2016. Online instrumental variable regression with applications to online linear system identification. In *AAAI Conference on Artificial Intelligence*, 2101–2107.

West, M., and Harrison, J. 1997. *Bayesian Forecasting and Dynamic Models (2nd ed.)*. Springer-Verlag.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. ICML.