Kalman Recursions Aggregated Online

Eric Adjakossa, Yannig Goude, Olivier Wintenberger

Kalman Recursions Aggregated Online

Eric Adjakossa[a,*], Yannig Goude[b], Olivier Wintenberger[a]

[a] *Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation (LPSM, UMR 8001), 4 place Jussieu, 75005 Paris, France*
[b] *EDF Lab, 7 Boulevard Gaspard Monge, 91120 Palaiseau*

Abstract

In this article, we aim at improving the prediction of expert aggregation by using the underlying properties of the models that provide expert predictions. We restrict ourselves to the case where expert predictions come from Kalman recursions, fitting state-space models. By using exponential weights, we construct different algorithms of Kalman recursions Aggregated Online (KAO) that compete with the best expert or the best convex combination of experts in a more or less adaptive way. We improve the existing results on expert aggregation literature when the experts are Kalman recursions by taking advantage of the second-order properties of the Kalman recursions. We apply our approach to Kalman recursions and extend it to the general adversarial expert setting by state-space modeling the errors of the experts. We apply these new algorithms to a real dataset of electricity consumption and show how it can improve forecast performances comparing to other exponentially weighted average procedures.

Keywords: online aggregation, Kalman filter, experts ensemble

1. Introduction

The aim of this paper is to aggregate Kalman recursions in an online setting in order to increase the accuracy of the prediction. We observe (y_t) sequentially through time $t \geq 1$ and predictions $\hat{y}_t^{(m)}$, $1 \leq m \leq M$, issued from Kalman recursions at time $t \geq 1$. Here $M \geq 0$ denotes the number of different Kalman recursions used as experts. The Kalman recursions are imbedded into a state-space model (see Section 2.1 for a formal definition). We introduce different Kalman recursions Aggregated Online (KAO) procedures that compute recursively weights $\rho_t^{(m)}$, $t \geq 1$, $1 \leq m \leq M$. We provide theoretical guarantees on the average prediction $\hat{y}_t = \sum_{m=1}^{M} \rho_t^{(m)} \hat{y}_t^{(m)}$.

We obtain bounds on the regret of KAO algorithms that are similar to the ones encountered in the literature. The book of reference on aggregation is undoubtedly the book from Cesa-Bianchi

*Corresponding author
Email addresses: ericadjakossah@gmail.com (Eric Adjakossa), yannig.goude@edf.fr (Yannig Goude), olivier.wintenberger@upmc.fr (Olivier Wintenberger)

and Lugosi (2006) and we refer to it for classical regret aggregation bounds. The novelty of our approach is to derive regret bounds directly on the cumulative quadratic predictive risk as defined by Wintenberger (2017). The predictive risk or risk of prediction of a predictor $\hat{y} \in \mathcal{F}_{t-1}$ is defined as

$$L_t(\hat{y}) = \mathbb{E}\left[(\hat{y} - y_t)^2 \mid \mathcal{F}_{t-1}\right], \qquad a.s., \qquad t \geq 1, \tag{1}$$

where $(\mathcal{F}_t)$ is the natural filtration of the past response variables $\sigma(y_s;\, 0 \leq s \leq t) = \mathcal{F}_t$, $t \geq 0$. The risk of prediction arises naturally when dealing with Kalman recursions. Indeed, Kalman recursions are online algorithms that provide the best linear predictions in gaussian state-space models. We refer to the classical monograph from Durbin and Koopman (2012) for details. The cumulative predictive risk of a recursive algorithm predicting $\hat{y}_t$ at each time $t \geq 1$ is the sum $\sum_{t=1}^{T} L_t(\hat{y}_t)$ up to the horizon $T \geq 1$. Our regret bounds on KAO algorithm predictions $(\hat{y}_t)$ are a.s. deterministic bound called respectively model selection regret or aggregation regret and defined as

$$R_T^S(m) \quad \geq \quad \sum_{t=1}^{T} L_t(\hat{y}_t) - L_t(\hat{y}_t^{(m)}) \qquad 1 \leq m \leq M, \tag{2}$$

$$R_T^A(\pi) \quad \geq \quad \sum_{t=1}^{T} L_t(\hat{y}_t) - L_t\left(\sum_{m=1}^{M} \pi^{(m)} \hat{y}_s^{(m)}\right), \tag{3}$$

for any vector of weights $\pi := (\pi^{(m)})_{1 \leq m \leq M}$ in the simplex. We suppress the dependence in m and π in R_t^S and R_t^A when the regret bounds are uniform in m and π, respectively.

Regret bounds on the cumulative predictive risk have attracted some interest since Audibert and Bubeck (2010) showed that the classical EWA algorithm from Vovk (1990) does not achieve a fast rate model selection regret even in the most favorable iid case. The fast rate model selection regret was first proved by the BOA algorithm in the iid setting with strongly convex loss in Wintenberger (2017) and then extended to any stochastic setting and exp-concave risk in Gaillard and Wintenberger (2016). For adaptative procedures, we improve their optimal regret

$$R_T^S = O(\log M + \log\log T + x), \qquad T \geq 1$$

with probability $1 - e^{-x}$, $x > 0$, to the a.s. bound

$$R_T^S = O(\log M + \log\log T) \qquad T \geq 1.$$

This optimal regret bound holds when the observations satisfy some unbounded state-space model defined in the next section 2.1. That the observations satisfy such a model is very unlikely in prac-

tice; It is the price to pay to guarantee a.s. optimal regret bounds on the cumulative predictive risk (1) for unbounded responses. Existing regret bounds such as the one of Gaillard and Wintenberger (2016) requires the boundedness of the response.

We present simulations and applications in cases where our assumptions are certainly not satisfied. Our new aggregation procedure improves the state of the art methods in aggregation such as MLPoly of Gaillard et al. (2014).

2. Preliminaries

2.1. State-space models

Assume that we observe (y_t, X_t) with $y_t \in \mathbb{R}$ the variable of interests and $X_t \in \mathbb{R}^d$ is the predictable design, i.e. $X_t \in \mathcal{F}_{t-1}$, $t \geq 1$. Notice that the design X_t can be either deterministic or random. We consider a collection of $M \geq 1$ experts $\hat{y}_t^{(m)} = X_t^\top \hat{\theta}_t^{(m)}$ issued from Kalman recursions as follows. We denote $\underset{t}{\mathbb{E}}[\cdot]$ and $\underset{t}{\mathrm{Var}}(\cdot)$ the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ and variance $\mathrm{Var}(\cdot \mid \mathcal{F}_t)$, respectively, for any $t \geq 0$.

For each $1 \leq m \leq M$, the sequence of experts $\hat{y}_t^{(m)} = X_t^\top \hat{\theta}_t^{(m)}$ is associated to a recursive hidden state model

$$\theta_t^{(m)} = K^{(m)} \theta_{t-1}^{(m)} + z_t^{(m)}\,, \qquad t \geq 1\,, \tag{4}$$

where $K^{(m)}$ is a $d \times d$ matrix, $z_t^{(m)} \sim \mathcal{N}(0, Q^{(m)})$ constitute an iid sequence and $\theta_0^{(m)} \in \mathbb{R}^d$ is deterministic. The sequence $(\theta_t^{(m)})$ is a Gaussian Markov chain and admits the representation

$$\theta_t^{(m)} = \left(K^{(m)}\right)^t \theta_0^m + \sum_{k=0}^{t-1} \left(K^{(m)}\right)^k z_{t-k}^{(m)}\,, \qquad t \geq 1\,, 1 \leq m \leq M\,. \tag{5}$$

It converges weakly to a stationary solution if and only if $\rho(K^{(m)}) < 1$, the spectral radius of $K^{(m)}$ is smaller than one. This assumption is not required in this work. Actually one of the most popular state models is the dynamic setting where one considers random walk under $K^{(m)} = I_d$, the identity matrix of $\mathbb{R}^d$. We notice that as the state model is hidden (latent, not observed), any assumption on the state recursion, such as the Gaussian assumption, is not restrictive for the observations (y_t, X_t).

Our main assumption is the following one.

(H) The vectors $(y_t, \theta_t^{(m)})$ constitute a Gaussian sequence,

$$\mathbb{E}[y_t \mid \theta_t^{(m)}] = X_t^\top \theta_t^{(m)}\,, \qquad t \geq 1\,, 1 \leq m \leq M\,,$$

Algorithm 1 Kalman recursion in the m-th state-space model

Parameters: The matrices $Q^{(m)}$ and $K^{(m)}$.
Initialization: The matrix $P_0^{(m)}$ and the vector $\hat{\theta}_0^{(m)}$.
Recursion: For each iteration $t = 1, \ldots, T$ do:

$$\begin{aligned}
\widehat{\theta}_{t+1}^{(m)} &= K^{(m)} \left(\widehat{\theta}_t^{(m)} + \frac{1}{X_t^\top P_t^{(m)} X_t + 1} P_t^{(m)} X_t (y_t - \hat{y}_t^{(m)}) \right) , \\
P_{t+1}^{(m)} &= K^{(m)} \left(P_t^{(m)} - \frac{1}{X_t^\top P_t^{(m)} X_t + 1} P_t^{(m)} X_t X_t^\top {P_t^{(m)}}^\top \right) {K^{(m)}}^\top + Q^{(m)} \\
\hat{y}_{t+1}^{(m)} &= X_{t+1}^\top \widehat{\theta}_{t+1}^{(m)} .
\end{aligned}$$

and the conditional variance $\sigma^{2(m)} := \mathrm{Var}(y_t \mid \theta_t^{(m)}) > 0$ is constant through time and known.

Condition (**H**) have different consequences upon the observations (y_t, X_t). The first obvious one is that (y_t) constitutes a Gaussian sequence. The second one is that y_t satisfies the linear model

$$y_t = X_t^\top \theta_t^{(m)} + \varepsilon_t^{(m)} , \qquad t \geq 1 , 1 \leq m \leq M .$$

The gaussian property of the couple $(y_t, \theta_t^{(m)})$ ensures that $\varepsilon_t^{(m)}$ is a gaussian random variable with mean zero and variance $\sigma^{2(m)} = \mathrm{Var}(y_t \mid \theta_t^{(m)})$ independent of $t \geq 1$. A direct implication from the expression (5) is the following mean-variance identity.

Proposition 2.1. Under Condition **(H)** the following mean-variance identity holds for all $1 \leq m \leq M$ and $t \geq 1$:

$$\begin{aligned}
\mathbb{E}[y_t] = \mathbb{E}[X_t^\top \theta_t^{(m)}] &= \mathbb{E}[X_t]^\top \left(K^{(m)} \right)^t \theta_0^m , \\
\mathrm{Var}(y_t) = \mathrm{Var}[X_t^\top \theta_t^{(m)}] &+ \sigma^{2(m)} \\
&= \mathbb{E}\left[X_t^\top \sum_{k=0}^{t-1} \left(K^{(m)} \right)^k Q^{(m)} \left(K^{(m)\top} \right)^k X_t \right] + \sigma^{2(m)} .
\end{aligned}$$

The static state-space model setting corresponds to the case where $Q^{(m)} = 0$ so that $\mathrm{Var}(y_t) = \sigma^{2(m)} = \sigma^2$, $1 \leq m \leq M$, $t \geq 1$.

2.2. The Kalman recursion

For the sake of completeness, we recall the Kalman recursion associated with the m-th state-space model in Algorithm 1. For details on the Kalman recursion we refer to the monograph from Durbin and Koopman (2012). We notice that the Kalman recursion does not require any inversion

of matrices. Each iteration has thus a $O(d^2)$ computational cost. Moreover it does not require the knowledge of the parameters $\sigma^{2(m)} > 0$. In addition, in many cases X_t is in fact a vector of size $d = \sum_{m=1}^{M} d_m$ that stacks M vectors $X_t^{(m)} \in \mathbb{R}^{d_m}$. In this case one considers d_m sparse vectors $\theta_t^{(m)}$ and identifies them with their non-null components $\theta_t^{(m)} \in \mathbb{R}^{d_m}$. Then the space equation is written as

$$y_t = {X_t^{(m)}}^\top \theta_t^{(m)} + \varepsilon_t^{(m)}\,, \qquad t \geq 1\,, 1 \leq m \leq M\,,$$

using similarly the notation $\varepsilon_t^{(m)} \in \mathbb{R}^{d_m}$. Doing so, each Kalman recursion holds in a state space of dimension $d^{(m)} < d$, lowering the computational cost of each iteration to $O(d_m^2)$.

In the static case when $K^{(m)} = I$ and $Q^{(m)}$ is the null matrix then, using the Shermann-Morrisson formula, we have the alternative recursion for $R_t^{(m)}$, the inverse of $P_t^{(m)}$,

$$R_{t+1}^{(m)} = R_t^{(m)} + X_t X_t^\top\,.$$

When $P_0^{(m)}$ is taken equals to $1/\lambda^{(m)} I_d$, for some $\lambda^{(m)} > 0$, then the estimator computed recursively using the Kalman recursion coincides with the Ridge estimator

$$\hat{\theta}_t^{(m)} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t} \left(y_s - X_s^\top \theta\right)^2 + \frac{\lambda^{(m)}}{2} \|\theta - \hat{\theta}_0^{(m)}\|_2^2 \right\}.$$

This equivalence has been first established by Diderrich (1985).

Notice that there is no assumption on the dependence among the Kalman recursions. Otherwise, it was possible to consider a Kalman recursion over the stack of the models in a dM dimensional state-space model. However, this approach is not practical when the computational cost $O((dM)^2)$ of the complete Kalman recursion is prohibitive because M is too large. The aim of this work is to show that this ideal procedure, that is uncertain in practice when the dependence among the recursions has to be estimated, can be easily overcome by a simple aggregation procedure over M Kalman recursions.

2.3. Examples

We provide some classical examples of state-space models satisfying **(H)**:

1. The static iid setting: this degenerate case coincides with the usual gaussian linear model for fixed or random (iid) design (X_t). We assume the relation $y_t = X_t^\top \theta_t^{(m)} + \varepsilon_t^{(m)}$, $t \geq 1$, associated with state equations $\theta_t^{(m)} = \theta_{t-1}^{(m)} = \cdots = \theta_0^{(m)}$ ($Q^{(m)} = 0$ for $1 \leq m \leq M$). Then the Kalman recursions are called static. Under **(H)** the mean-variance identity Proposition 2.1 implies $\sigma^{2(m)} = \sigma^2$.

2. The dynamical setting: This setting relies on the random walk state equations

$$\theta_t^{(m)} = \theta_{t-1}^{(m)} + z_t^{(m)}\,, \qquad t \geq 1\,, \tag{6}$$

from initial null state $\theta_0^{(m)} = 0$ for all $1 \leq m \leq M$. Then the mean identity of Proposition 2.1 is automatically satisfied as $\mathbb{E}[y_t] = 0$ for all $t \geq 1$. The variance identity requires that $\mathbb{E}[X_t^\top Q^{(m)} X_t] = \mathbb{E}[X_t^\top Q^{(m')} X_t]$ and $\sigma^{2(m)} = \sigma^{2(m')}$ for any $1 \leq m, m' \leq M$ and $t \geq 1$. The Kalman recursion can be used for tracking the signal (y_t) on different explanatory variables $X_t^{(m)}$ stacked in X_t.

3. The expert setting: We consider the case where we have M deterministic experts without any information about their generation process. This situation is very common in real-life applications as the forecast can come from different sources (physical models, different data sources, different machine learning models).
For each $1 \leq m \leq M$ expert we stack its prediction $f_{m,t} \in \mathbb{R}$ in $X_t^{(m)}$ together with the intercept and the past error $e_{m,t-1} = (y_{t-1} - f_{m,t-1})$, i.e.,

$$X_t^{(m)} = (1, f_{m,t}, e_{m,t-1}), \qquad t \geq 1\,.$$

and each state-space model is defined by the state equation:

$$\theta_t^{(m)} = K^{(m)}\theta_{t-1}^{(m)} + z_t^{(m)}\,, \qquad t \geq 1.$$

3. Kalman recursions Aggregated Online (KAO) algorithm

Consider the state-space models (coinciding with Equation (4) for the mth state equation, $1 \leq 1 \leq M$)

$$\left\{ \begin{array}{rcl} y_t & = & X_t^\top \theta_t^{(m)} + \varepsilon_t^{(m)} \\ \theta_t^{(m)} & = & K^{(m)}\theta_{t-1}^{(m)} + z_t^{(m)} \end{array} \right. , \quad t \geq 1, \quad 1 \leq m \leq M.$$

Recall that under **(H)** we have

$$\mathbb{E}[y_t \mid m] := \mathbb{E}[y_t \mid z_t^{(m)}, \ldots, z_1^{(m)}, \mathcal{F}_{t-1}] = X_t^\top \theta_t^{(m)}, \qquad 1 \leq m \leq M.$$

We aggregate Kalman recursions using a version of the exponentially weighted average forecaster defined as

$$\hat{y}_t = \sum_{m=1}^{M} \rho_t^{(m)} \hat{y}_t^{(m)} \tag{7}$$

with $\rho_t^{(m)} \geq 0$ for $1 \leq m \leq M$, $\sum_{m=1}^{M} \rho_t^{(m)} = 1$ and $\hat{y}_t^{(m)} = X_t^\top \theta_t^{(m)}$ is the mth Kalman forecaster of y_t.

3.1. Convex properties

The ability to find rapidly the solution of an optimization problem depends heavily on the convex properties of the objective function. In our case, the objective function is the conditional risk defined in (1) as $L_t(\hat{y}) = \mathop{\mathbb{E}}_{t-1}[(\hat{y} - y_t)^2]$. Due to the conditional expectation, it is a random convex function and its minimum $\mathop{\mathbb{E}}_{t-1}[y_t]$, called the best prediction, varies upon the time $t \geq 1$. Thus one cannot expect that our procedure converges in general and we rather study its regrets R_t^S and R_t^A defined in Equations (2) and (3) as the model selection and aggregation regret, respectively. The objective function is

$$\sum_{s=1}^{t} L_s\Big(\sum_{m=1}^{M} \pi^{(m)} \hat{y}_s^{(m)}\Big)$$

for any $(\pi^{(m)})_{1 \leq m \leq M}$ in the canonical basis or in the simplex, i.e. $\pi^{(m)} \geq 0$ such that $\sum_{m=1}^{M} \pi^{(m)} = 1$. The optimal rates of convergence in the model selection and the aggregation problems depend on the convex properties of the objective function and the observation of an approximation of the gradients. As the objective functions are convex, we will extensively use the gradient trick which consists to bounding the regret with the linearized risks $\mathscr{L}_s^{(m)} = L_s'(\hat{y}_s)(\hat{y}_s^{(m)} - \hat{y}_s)$ as

$$L_t(\hat{y}_t) - L_t\Big(\sum_{m=1}^{M} \pi^{(m)} \hat{y}_t^{(m)}\Big) \leq -\sum_{m=1}^{M} \pi^{(m)} \mathscr{L}_t^{(m)} \,. \tag{8}$$

Fast rates of convergence could be obtained easily if the objective function was strongly convex. Despite we use the square loss, it is not the case since the Hessian matrix

$$2\sum_{s=1}^{t} (\hat{y}_s^{(m)})_{1 \leq m \leq M} (\hat{y}_s^{(m)})_{1 \leq m \leq M}^\top$$

of the objective function is a sum of rank-one matrices which are very unlikely to converge in any non-stationary settings. This issue is bypassed in online convex optimization thanks to the notion of exp-concavity extensively studied by Hazan et al. (2016).

Definition 3.1. A loss function ℓ is η-exp-concave (with $\eta > 0$) on some convex set $\mathcal{Y}$ if the function $F(y) = \exp(-\eta \ell(y))$ is concave for all $y \in \mathcal{Y}$.

We need to find out for which values of η the conditional risks (1) are exp-concave. Moreover, we can use the exp-concave property of the risk L_t to refine the gradient trick.

Theorem 3.1. Assume that it exists $D > 0$ such that $|\hat{y}_t^{(m)} - \mu_t| \leq D$ a.s. $1 \leq m \leq M$, $t \geq 1$ with $\mu_t = \underset{t-1}{\mathbb{E}}[y_t]$. Then the conditional risk L_t is a.s. $(2D^2)^{-1}$-exp-concave for any $y = \sum_{m=1}^{M} \pi^{(m)} \hat{y}_t^{(m)}$, $t \geq 1$, with $(\pi^{(m)})_{1 \leq m \leq M}$ in the simplex. Moreover if the linearized risk satisfies $|\mathscr{L}_t^{(m)}| \leq G^{(m)}$ for any $t \geq 1$, $1 \leq m \leq M$, then we have

$$L_t(\hat{y}_t) - L_t(\hat{y}_t^{(m)}) \leq -\mathscr{L}_t^{(m)} - \eta^{(m)} {\mathscr{L}_t^{(m)}}^2,$$

with $\eta^{(m)} = \dfrac{1}{8(2G^{(m)} \vee D^2)}$, $1 \leq m \leq M$.

The refined linearized risk $\mathscr{L}_t^{(m)} + \eta^{(m)} {\mathscr{L}_t^{(m)}}^2$ is called the surrogate risk. It is itself exp-concave due to the quadratic term whereas the linearized risk cannot be exp-concave.

Proof. Let $0 < \eta \leq \dfrac{1}{(2D^2)}$. Consider the function $\varphi_\eta(y) = e^{-\eta L_t(y)}$ for $y = \sum_{m=1}^{M} \pi^{(m)} \hat{y}_t^{(m)}$ and $(\pi^{(m)})_{1 \leq m \leq M}$ in the simplex. The function φ is a.s. at least twice differentiable and we have

$$\varphi_\eta''(y) = -2\eta \varphi_\eta(y) \left(1 - 2\eta(y - \mu_t)^2\right).$$

using the derivation under the integral sign. For $\theta \in \Theta$, we get the concavity since $\varphi_\eta''(y) \leq 0$ as

$$2\eta(y - \mu_t)^2 \leq 2\eta(\sum_{m=1}^{M} \pi^{(m)}(\hat{y}_t^{(m)} - \mu_t))^2 \leq 2\eta D^2 = 1$$

and the first assertion follows.

We proceed as in the proof of Lemma 4.2 of Hazan et al. (2016) considering $\gamma^{(m)} = \dfrac{1}{2(2G^{(m)} \vee D^2)} \leq \eta$. One deduces from the concavity property of $\varphi_{\gamma^{(m)}}$ that $\varphi_{\gamma^{(m)}}(y) - \varphi_{\gamma^{(m)}}(z) \leq \varphi'_{\gamma^{(m)}}(z)(y - z)$ which, taking $y = \hat{y}_t^{(m)}$ and $z = \hat{y}_t$, provides

$$\exp(-\gamma^{(m)} L_t(\hat{y}_t^{(m)})) - \exp(-\gamma^{(m)} L_t(\hat{y}_t)) \leq -\gamma^{(m)} L_t'(\hat{y}_t) \exp(-\gamma^{(m)} L_t(\hat{y}_t))(\hat{y}_t^{(m)} - \hat{y}_t).$$

One deduces that

$$\gamma^{(m)}(L_t(\hat{y}_t) - L_t(\hat{y}_t^{(m)})) \leq \log(1 - \gamma^{(m)} L_t'(\hat{y}_t)(\hat{y}_t^{(m)} - \hat{y}_t)).$$

Using the relation $\log(1-z) \leq -z - \dfrac{1}{4}z^2$ that holds for any $|z| \leq 1/4$ applied on $|\gamma^{(m)} L_t'(\hat{y}_t)(\hat{y}_t^{(m)} -$

Algorithm 2 KAO for model selection

Parameters: The variances $\sigma^{2(m)}$, $1 \leq m \leq M$ and the learning rate η.
Initialization: The initial weights $\rho_1^{(m)} = \rho_0^{(m)}$, $1 \leq m \leq M$.
For each iteration $t = 1, \ldots, T$:
Inputs: The Kalman predictions $\hat{y}_{t+1}^{(m)}$ and the matrices $P_t^{(m)}$, $1 \leq m \leq M$.
Recursion: Do:

$$\begin{aligned} \rho_{t+1}^{(m)} &= \frac{\exp\left(-\eta\left(X_t^\top P_t^{(m)} X_t + \sigma^{2(m)}\right)\right)\rho_t^{(m)}}{\sum_{m=1}^M \exp\left(-\eta\left(X_t^\top P_t^{(m)} X_t + \sigma^{2(m)}\right)\right)\rho_t^{(m)}} \\ \hat{y}_{t+1} &= \sum_{m=1}^M \rho_{t+1}^{(m)} \hat{y}_{t+1}^{(m)} . \end{aligned}$$

$\hat{y}_t)| \leq 1/4$ one obtains

$$\gamma^{(m)}(L_t(\hat{y}_t) - L_t(\hat{y}_t^{(m)})) \leq \gamma^{(m)} L_t'(\hat{y}_t)(\hat{y}_t - \hat{y}_t^{(m)}) - \frac{1}{4}(\gamma^{(m)} L_t'(\hat{y}_t)(\hat{y}_t - \hat{y}_t^{(m)}))^2$$

and the second assertion follows. □

3.2. KAO for model selection

In this Section we assume the exp-concavity of the conditional risks and we adapt the classical analysis of the Exponentially Weighted Average (EWA) algorithm of Cesa-Bianchi and Lugosi (2006) to our setting. The aggregation procedure, called KAO, is described in Algorithm 2.

KAO achieves the optimal rate for model selection.

Theorem 3.2. Under assumption **(H)** and if it exists $D > 0$ such that $|\hat{y}_t^{(m)} - \mu_t| \leq D$ a.s. $1 \leq m \leq M$, $t \geq 1$ then KAO for model selection with $\eta = \dfrac{1}{(2D^2)}$ achieves the regret bound

$$R_t^S(m) \leq -2D^2 \log(\rho_0^{(m)}) \qquad 1 \leq t \leq T\,, 1 \leq m \leq M\,.$$

We note that the classical EWA algorithm satisfies a similar regret bound under the stronger assumption

$$|\hat{y}_t^{(m)} - y_t| \leq D, \qquad 1 \leq m \leq M,\, t \geq 1,\, a.s.$$

which never holds in our Gaussian setting. One usual way to bypass this well-known restriction of EWA is to use a doubling trick which deteriorates the regret bound, see Cesa-Bianchi and Lugosi (2006) for more details.

Proof. The proof is standard and follows the line of the proof of the EWA regret in Cesa-Bianchi and Lugosi (2006). The crucial step consists in identifying the conditional risk of any Kalman

prediction $\hat{y}_t^{(m)}$ under **(H)**. We have the following Lemma

Lemma 3.1. *Under* **(H)** *we have the identity* $\hat{y}_t^{(m)} = \underset{t-1}{\mathbb{E}}[\mathbb{E}[y_t \mid m]]$.

Proof. The Kalman recursion produces the best linear prediction which is equal to the conditional expectation in the gaussian case. Then we have $\hat{y}_t^{(m)} = \underset{t-1}{\mathbb{E}}[X_t\theta_t^{(m)}] = \underset{t-1}{\mathbb{E}}[\mathbb{E}[y_t|m]]$ by definition. □

We have explicitely

$$\begin{aligned}
L_s(\hat{y}_s^{(m)}) &= \underset{s-1}{\mathbb{E}}[(y_s - \underset{s-1}{\mathbb{E}}[\mathbb{E}[y_s \mid m]])^2] \\
&= \underset{s-1}{\mathbb{E}}[(y_s - \mathbb{E}[y_s \mid m])^2] + \underset{s-1}{\mathbb{E}}[(\mathbb{E}[y_s \mid m] - \underset{s-1}{\mathbb{E}}[\mathbb{E}[y_s \mid m]])^2] \\
&\quad +2\underset{s-1}{\mathbb{E}}[(y_s - \mathbb{E}[y_s \mid m])(\mathbb{E}[y_s \mid m] - \underset{s-1}{\mathbb{E}}[\mathbb{E}[y_s \mid m]])] \\
&= \underset{s-1}{\mathbb{E}}[(y_s - X_t\theta_t^{(m)})^2] + \underset{s-1}{\mathbb{E}}[(X_t(\theta_t^{(m)} - \hat{\theta}_t^{(m)})^2] \\
&= \sigma^{2(m)} + X_s^\top P_s^{(m)} X_s \,,
\end{aligned}$$

since the third term of the sum is zero and since $\underset{s-1}{\mathbb{E}}[(X_t(\theta_t^{(m)} - \hat{\theta}_t^{(m)})^2] = X_s^\top P_s^{(m)} X_s$ thanks to the Kalman recursion properties in the gaussian case. We also have, using the exp-concavity of L_t and Jensen inequality,

$$\begin{aligned}
e^{-\eta L_t(\hat{y}_t)} &= e^{-\eta L_t(\sum_{m=1}^M \rho_t^{(m)} \hat{y}_t^{(m)})} \\
&\geq \sum_{m=1}^M \rho_t^{(m)} e^{-\eta L_t(\hat{y}_t^{(m)})}, \\
&\geq \frac{\sum_{m=1}^M \rho_0^{(m)} e^{-\eta \sum_{s=1}^{t-1} L_s(\hat{y}_s)} e^{-\eta L_t(\hat{y}_t^{(m)})}}{\sum_{m=1}^M \rho_0^{(m)} e^{-\eta \sum_{s=1}^{t-1} L_s(\hat{y}_s)}} \\
&\geq \frac{\sum_{m=1}^M \rho_0^{(m)} e^{-\eta R_{t-1}^S(m)} e^{-\eta L_t(\hat{y}_t^{(m)})}}{\sum_{m=1}^M \rho_0^{(m)} e^{-\eta R_{t-1}^S(m)}},
\end{aligned}$$

multiplying by $e^{\eta \sum_{s=1}^{t-1} L_s(\hat{y}_s^{(m)})}$ above and below the fraction. We get the recursive relation

$$1 = \sum_{m=1}^M \rho_0^{(m)} \geq \sum_{m=1}^M \rho_0^{(m)} e^{-\eta R_{t-1}^S(m)} \geq \sum_{m=1}^M \rho_0^{(m)} e^{-\eta R_t^S(m)}$$

and the desired result follows. □

3.3. KAO for aggregation

In the case where the best expert is not worthy of confidence, it is generally much more interesting to compete with the best convex combination of the experts at hand. In this context, the aim is to provide a bound on the regret for aggregation $R_t^A(\pi)$ where $\pi := (\pi^{(m)})_{1 \leq m \leq M}$

Algorithm 3 KAO for aggregation

Parameters: The variances $\sigma^{2(m)}$, $1 \leq m \leq M$ and the learning rate η.
Initialization: The initial weights $\rho_0^{(m)}$, $1 \leq m \leq M$.
For each iteration $t = 0, \ldots, T$:
Inputs: The Kalman predictions $\hat{y}_{t+1}^{(m)}$ and the matrices $P_t^{(m)}$, $1 \leq m \leq M$.
Recursion: Do:

$$\begin{aligned} \mathscr{L}_t^{(m)} &= (9) \\ \rho_{t+1}^{(m)} &= \frac{\exp\left(-\eta \mathscr{L}_t^{(m)}\right) \rho_t^{(m)}}{\sum_{m'=1}^{M} \exp\left(-\eta \mathscr{L}_t^{(m)}\right) \rho_t^{(m')}} \\ \hat{y}_{t+1} &= \sum_{m=1}^{M} \rho_{t+1}^{(m)} \hat{y}_{t+1}^{(m)} \,. \end{aligned}$$

belongs to the simplex. As the conditional risk L_s is a convex function that is differentiable, one applies the gradient trick and we consider an explicit biased version of the linearized risk

$$\mathscr{L}_t^{(m)} = X_t^\top P_t^{(m)} X_t + \sigma^{2(m)} - (\hat{y}_t - \hat{y}_t^{(m)})^2 - \sum_{m'=1}^{M} \rho_t^{(m')} \left(X_t^\top P_t^{(m')} X_t + \sigma^{2(m')} - (\hat{y}_t - \hat{y}_t^{(m')})^2 \right). \quad (9)$$

By convention $\mathscr{L}_0^{(m)} = 0$. In our setting, the adaptation of the gradient-based EWA of Cesa-Bianchi and Lugosi (2006) yields Algorithm 3. The following theorem derives an upper bound for the regret R_t^A.

Theorem 3.3. Under Assumption **(H)**, suppose it exists $G > 0$ such that $|\mathscr{L}_t^{(m)}| \leq G$ a.s. for $1 \leq t \leq T$, $1 \leq m \leq M$. Then KAO for aggregation starting with $\rho_0^{(m)} = 1/M$ and $\eta = \frac{1}{G}\sqrt{\frac{2 \log M}{t}}$ satisfies the regret bound

$$R_t^A \leq G\sqrt{2t \log M}\,, \qquad 1 \leq t \leq T. \quad (10)$$

The regret bound matches the optimal bound for $M \geq \sqrt{t}$. Note that the boundedness assumption on $\mathscr{L}_t^{(m)}$ involves only the predictions and does not require to bound (y_t).

Proof. Since L_s is convex and differentiable, we apply the gradient trick

$$R_t^A(\pi) \leq -\sum_{s=1}^{t} \sum_{m=1}^{M} \pi^{(m)} \mathscr{L}_s^{(m)}.$$

Moreover, the expression of $L'_s(\hat{y}_s)\hat{y}_s^{(m)}$ can be developed as

$$\begin{aligned} L'_s(\hat{y}_s)\hat{y}_s^{(m)} &= 2\,\mathop{\mathbb{E}}_{s-1}[(\hat{y}_s - y_s)\hat{y}_s^{(m)}] \\ &= 2\hat{y}_s\hat{y}_s^{(m)} + \mathop{\mathbb{E}}_{s-1}[(y_s - \hat{y}_s^{(m)})^2] - {\hat{y}_s^{(m)}}^2 - \mathop{\mathbb{E}}_{s-1}[y_s^2] \\ &= \mathop{\mathbb{E}}_{s-1}[(y_s - \hat{y}_s^{(m)})^2] - (\hat{y}_s - \hat{y}_s^{(m)})^2 + \hat{y}_s^2 - \mathop{\mathbb{E}}_{s-1}[y_s^2] \\ &= X_s^\top P_s^{(m)} X_s + \sigma^{2(m)} - (\hat{y}_s - \hat{y}_s^{(m)})^2 + \hat{y}_s^2 - \mathop{\mathbb{E}}_{s-1}[y_s^2]. \end{aligned}$$

Since the two last summands of $L'_s(\hat{y}_s)\hat{y}_s^{(m)}$ do not depend on m we obtain the identity $\mathscr{L}_t^{(m)} = L'_s(\hat{y}_s)(\hat{y}_s^{(m)} - \hat{y}_s) = (9)$. As it exists $G > 0$ satisfying $|\mathscr{L}_t^{(m)}| \leq G$, by using the Hoeffding lemma (i.e. $\log \mathbb{E}[e^{\alpha X}] \leq \frac{\alpha^2}{2}G^2$, for any centered random variable $|X| \leq G$, with $\alpha \in \mathbb{R}$), and the identity

$$\rho_s^{(m)} = \frac{\exp(-\eta \sum_{r=1}^{s-1} \mathscr{L}_t^{(m)})\rho_0^{(m)}}{\sum_{m'=1}^{M} \exp(-\eta \sum_{r=1}^{s-1} \mathscr{L}_t^{(m')})\rho_0^{(m')}}$$

we get

$$\log\left(\sum_{m=1}^{M} \frac{\exp\left(-\eta \sum_{r=1}^{s} \mathscr{L}_r^{(m)}\right)\rho_0^{(m)}}{\sum_{m'=1}^{M} \exp\left(-\eta \sum_{r=1}^{s-1} \mathscr{L}_r^{(m')}\right)\rho_0^{(m')}}\right) \leq \frac{\eta^2}{2}G^2\,.$$

Then, by summing over s, a telescoping sum appears and leads to

$$\frac{1}{\eta}\log\left(\sum_{m=1}^{M} \exp\left(-\eta \sum_{s=1}^{t} \mathscr{L}_s^{(m)}\right)\rho_0^{(m)}\right) \leq \eta t \frac{G^2}{2}.$$

Moreover,

$$\exp\left(-\eta \sum_{s=1}^{t} \mathscr{L}_s^{(m)}\right)\rho_0^{(m)} \leq \sum_{m=1}^{M} \exp\left(-\eta \sum_{s=1}^{t} \mathscr{L}_s^{(m)}\right)\rho_0^{(m)},$$

which leads to

$$\sum_{m=1}^{M} \pi^{(m)}\left(\sum_{s=1}^{t} -\mathscr{L}_s^{(m)} + \frac{\log \rho_0^{(m)}}{\eta}\right) \leq \frac{1}{\eta}\log\left(\sum_{m=1}^{M} \exp\left(-\eta \sum_{s=1}^{t} \mathscr{L}_s^{(m)}\right)\rho_0^{(m)}\right).$$

Combining those bounds, we obtain

$$R_t^A(\pi) \leq \sum_{s=1}^{t}\sum_{m=1}^{M} -\pi^{(m)}\mathscr{L}_s^{(m)} \leq \sum_{m=1}^{M} \pi^{(m)}\left(\frac{-\log \rho_0^{(m)}}{\eta} + \eta t \frac{G^2}{2}\right).$$

We get the desired result noticing that $\rho_0^{(m)} = 1/M$ and the optimal choice of $\eta = \frac{1}{G}\sqrt{\frac{2\log M}{t}}$. □

We notice that a unique learning rate yields a uniform regret bound, independent of π. We also notice that we can use the gradient trick despite we only observed a biased version of the

Algorithm 4 KAO with multiple learning rates

Parameters: The variances $\sigma^{2(m)}$, the weights $\tilde{\rho}_0^{(m)}$ and the learning rates $\eta^{(m)}$, $1 \leq m \leq M$.
Initialization: The initial weights $\rho_0^{(m)} = \eta^{(m)}\tilde{\rho}_0^{(m)} / (\sum_{m'=1}^{M} \eta^{(m')}\tilde{\rho}_0^{(m')})$, $1 \leq m \leq M$.
For each iteration $t = 1, \ldots, T$:
Inputs: The Kalman predictions $\hat{y}_{t+1}^{(m)}$ and the matrices $P_t^{(m)}$, $1 \leq m \leq M$.
Recursion: Do:

$$
\begin{aligned}
\mathscr{L}_t^{(m)} &= (9) \\
\rho_{t+1}^{(m)} &= \frac{\exp\left(-\eta^{(m)}\mathscr{L}_t^{(m)}(1+\eta^{(m)}\mathscr{L}_t^{(m)})\right)\rho_t^{(m)}}{\sum_{m'=1}^{M}\exp\left(-\eta^{(m)}\mathscr{L}_t^{(m')}(1+\eta^{(m')}\mathscr{L}_t^{(m')})\right)\rho_t^{(m')}} \\
\hat{y}_{t+1} &= \sum_{m=1}^{M} \rho_{t+1}^{(m)}\hat{y}_{t+1}^{(m)} \,.
\end{aligned}
$$

linearized risk thanks to the exponential form of the weights that are not sensitive to the bias.

4. Online tuning of the learning rates

The theoretical guarantees on the regret for the model selection and the aggregation problems do not hold for the same algorithm. The gradient trick is a crucial step in the proof of the regret for the aggregation problem. However, the fast rate for the model selection does not hold for the gradient-based EWA since the linearized risk cannot be exp-concave. In order to bypass this issue, we adapt the approach of Wintenberger (2017) to our setting. The first step is to use a surrogate loss of the form $\mathscr{L}_t^{(m)}\left(1+\eta\mathscr{L}_t^{(m)}\right)$ where the quadratic part yields exp-concavity. The second step is to use a multiple learning rates version of KAO as described in Algorithm 4 in Section 4.1 where we show that multiple learning rates can be easily tuned online.

4.1. Multiple learning rates for KAO

In the context of expert aggregation, it is well known that using multiple learning rates help to increase the prediction accuracy, see Gaillard et al. (2014) and Wintenberger (2017). Here we aim to provide a multiple learning rates version for KAO in a similar way than the multiple learning rate version of the BOA procedure (see Wintenberger (2017)). The following theorem provides regret bounds both for model selection and aggregation on the same algorithm.

Theorem 4.1. Under assumption **(H)** suppose it exists $G > 0$ such that $|\mathscr{L}_t^{(m)}| \leq G$ a.s. for $1 \leq t \leq T$, $1 \leq m \leq M$. Then the aggregation regret of KAO with multiple learning rates

$\eta^{(m)} = \frac{1}{G}\left(\sqrt{-\frac{\log \tilde{\rho}_0^{(m)}}{t}} \wedge \frac{1}{2}\right)$ is bounded as

$$R_t^A(\pi) \leq 2G \sum_{m=1}^{M} \pi^{(m)} \left(\sqrt{-\log \tilde{\rho}_0^{(m)} t} - \log \tilde{\rho}_0^{(m)}\right), \qquad 1 \leq t \leq T\,. \tag{11}$$

If moreover there exists $D > 0$ such that $|\hat{y}_t^{(m)} - \mu_t| \leq D$ a.s. $1 \leq m \leq M$, $t \geq 1$ then the model selection regret of KAO with multiple learning rates $\eta^{(m)} = \frac{1}{8(2G \vee D^2)}$ for all $1 \leq m \leq M$ is bounded as

$$R_t^S(m) \leq -8(2G \vee D^2) \log \tilde{\rho}^{(m)}\,. \tag{12}$$

Proof. We start by applying the gradient trick as in (8) inferring that

$$R_t^A(\pi) \leq -\sum_{s=1}^{t} \sum_{m=1}^{M} \pi^{(m)} \mathscr{L}_s^{(m)}.$$

Moreover, as $x - x^2$ is 1-exp-concave for $x > 1/2$, Jensen's inequality implies that

$$\mathbb{E}\left[\exp\left(X - X^2\right)\right] \leq \exp\left(\mathbb{E}[X] - \mathbb{E}[X]^2\right) = 1 \tag{13}$$

for any centered random variable X such that $X \geq -1/2$ a.s. We notice that $\eta^{(m)} = \frac{1}{G}\sqrt{-\frac{\log \tilde{\rho}_0^m}{t}} \wedge \frac{1}{2}$ satisfies the relation

$$\eta^{(m)} \mathscr{L}_t^{(m)} \leq \frac{1}{2}\,.$$

Denoting

$$\tilde{\rho}_t^{(m)} = \frac{\exp\left[-\sum_{s=1}^{t-1} \eta^{(m)} \mathscr{L}_s^{(m)} \left(1 + \eta^{(m)} \mathscr{L}_s^{(m)}\right)\right] \tilde{\rho}_0^{(m)}}{\sum_{m'=1}^{M} \exp\left[-\sum_{s=1}^{t-1} \eta^{(m')} \mathscr{L}_s^{(m')} \left(1 + \eta^{(m')} \mathscr{L}_s^{(m')}\right)\right] \tilde{\rho}_0^{(m')}},$$

we have the identity

$$\tilde{\rho}_t^{(m)} = \frac{\rho_t^{(m)}}{\eta^{(m)}} \times \frac{1}{\sum_{m'=1}^{M} \rho_t^{(m')}/\eta^{(m')}},$$

which leads to

$$\sum_{m=1}^{M} \tilde{\rho}_t^{(m)} \eta^{(m)} \mathscr{L}_t^{(m)} = 0.$$

Thus the random variable $\left(\eta^{(m)} \mathscr{L}_t^{(m)}\right)_{1 \leq m \leq M}$ is therefore centered for the distribution $\left(\tilde{\rho}_t^{(m)}\right)_{1 \leq m \leq M}$, and we have

$$\sum_{m=1}^{M} \tilde{\rho}_t^{(m)} \exp\left[-\eta^{(m)} \mathscr{L}_t^{(m)} \left(1 + \eta^{(m)} \mathscr{L}_t^{(m)}\right)\right] \leq 1, \tag{14}$$

since $-\eta^{(m)} \mathscr{L}_t^{(m)} > -1/2$ a.s. for all $1 \leq m \leq M$ and $1 \leq t \leq T$, using (13). By putting the

expression of $\tilde{\rho}_t^{(m)}$ into Equation (14), we have

$$\sum_{m=1}^{M} \tilde{\rho}_0^{(m)} \exp\left[-\sum_{s=1}^{t} \eta^{(m)} \mathscr{L}_s^{(m)} \left(1 + \eta^{(m)} \mathscr{L}_s^{(m)}\right)\right] \leq \sum_{m=1}^{M} \tilde{\rho}_0^{(m)} \exp\left[-\sum_{s=1}^{t-1} \eta^{(m)} \mathscr{L}_s^{(m)} \left(1 + \eta^{(m)} \mathscr{L}_s^{(m)}\right)\right],$$

which implies for $1 \leq t \leq T$ that

$$\sum_{m=1}^{M} \tilde{\rho}_0^{(m)} \exp\left[-\sum_{s=1}^{t} \eta^{(m)} \mathscr{L}_s^{(m)} \left(1 + \eta^{(m)} \mathscr{L}_s^{(m)}\right)\right] \leq 1,$$

since by convention that $\mathscr{L}(\hat{y}_0^{(m)}) = 0$. Thus, for $1 \leq m \leq M$,

$$-\sum_{s=1}^{t} \eta^{(m)} \mathscr{L}_s^{(m)} \left(1 + \eta^{(m)} \mathscr{L}_s^{(m)}\right) \leq -\log \tilde{\rho}_0^{(m)}, \tag{15}$$

by applying the logarithm function using the previous inequality. We have

$$-\sum_{m=1}^{M} \tilde{\pi}^{(m)} \sum_{s=1}^{t} \eta^{(m)} \mathscr{L}_s^{(m)} \leq \sum_{m=1}^{M} \tilde{\pi}^{(m)} \left(\eta^{(m)^2} \sum_{s=1}^{t} \mathscr{L}_s^{(m)^2} - \log \tilde{\rho}_0^{(m)}\right)$$

for

$$\tilde{\pi}^{(m)} = \frac{\pi^{(m)}/\eta^{(m)}}{\sum_{m'=1}^{M} \pi^{(m')}/\eta^{(m')}}, \qquad 1 \leq m \leq M\,.$$

Multiplying with $\sum_{m'=1}^{M} \pi^{(m')}/\eta^{(m')}$ we obtain

$$\begin{aligned} -\sum_{m=1}^{M} \pi^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)} &\leq \sum_{m=1}^{M} \pi^{(m)} \left(\eta^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)^2} - \frac{\log \tilde{\rho}_0^{(m)}}{\eta^{(m)}}\right) \\ &\leq \sum_{m=1}^{M} \pi^{(m)} \left(\eta^{(m)} G^2 t - \frac{\log \tilde{\rho}_0^{(m)}}{\eta^{(m)}}\right) \end{aligned}$$

and the desired result on $R_t^A(\pi)$ follows from the specific choice of $\eta^{(m)}$.

The regret bound on $R_t^S(m)$ follows by an application of Theorem 3.1 on the last bound specified for π in the canonical basis

$$\sum_{s=1}^{t} -\mathscr{L}_s^{(m)} - \eta^{(m)} \mathscr{L}_s^{(m)^2} \leq \pi^{(m)} \left(-\frac{\log \tilde{\rho}_0^{(m)}}{\eta^{(m)}}\right)$$

for all $1 \leq m \leq M$ as $\eta^{(m)} = \dfrac{1}{8(2G \vee D^2)}$. □

Algorithm 5 KAO with adaptive multiple learning rates

Parameters: The variances $\sigma^{2(m)}$, $1 \leq m \leq M$.

Initialization: Any initial weights $\tilde{\rho}_0^{(m)} > 0$ such that $\sum_{m=1}^{M} \tilde{\rho}_0^{(m)} = 1$, $1 \leq m \leq M$.

For each iteration $t = 1, \ldots, T$:

Inputs: The Kalman predictions $\hat{y}_{t+1}^{(m)}$ and the matrices $P_t^{(m)}$, $1 \leq m \leq M$.

Recursion: Do:

$$
\begin{aligned}
\mathscr{L}_t^{(m)} &= (9) \\
\eta_t^{(m)} &= \sqrt{\frac{-\log \tilde{\rho}_0^{(m)}}{1 + \sum_{s=1}^{t} {\mathscr{L}_s^{(m)}}^2}} \\
\rho_{t+1}^{(m)} &= \frac{\eta_t^{(m)} \exp\left[-\eta_t^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)} \left(1 + \eta_{s-1}^{(m)} \mathscr{L}_s^{(m)}\right)\right] \tilde{\rho}_0^{(m)}}{\sum_{m'=1}^{M} \eta_t^{(m')} \exp\left[-\eta_t^{(m')} \sum_{s=1}^{t} \mathscr{L}_s^{(m')} \left(1 + \eta_{s-1}^{(m')} \mathscr{L}_s^{(m')}\right)\right] \tilde{\rho}_0^{(m')}} \\
\hat{y}_{t+1} &= \sum_{m=1}^{M} \rho_{t+1}^{(m)} \hat{y}_{t+1}^{(m)} \, .
\end{aligned}
$$

However the regrets bound do not apply on KAO with the same learning rates. The slow rate aggregation regret bound holds for a $O(1/\sqrt{t})$ learning rate whereas the fast rate model selection regret bound holds for a constant learning rate.

4.2. Adaptive multiple learning rates

Multiple learning rates are easily adaptable as in Algorithm 5. Moreover, a single algorithm with unique adaptive learning rates achieves optimal regret bounds for both model selection and aggregation problems as for the BOA algorithm developed by Wintenberger (2017) and refined by Gaillard and Wintenberger (2018).

Theorem 4.2. Under assumption **(H)** suppose there exist $G^{(m)} > 0$ and $D > 0$ such that $|\mathscr{L}_t^{(m)}| \leq G^{(m)}$ and $|\hat{y}_t^{(m)} - \mu_t| \leq D$ a.s. for $1 \leq t \leq T$, $1 \leq m \leq M$. Then the regret of KAO with adaptive multiple learning rates such that $\eta_{t-1}^{(m)} \mathscr{L}_t^{(m)} < 1/2$ for any $1 \leq t \leq T$ and $1 \leq m \leq M$ is bounded as

$$
\begin{aligned}
R_t^A(\pi) &\leq \sum_{m=1}^{M} \pi^{(m)} \left(G^{(m)}(3 + G^{(m)})\sqrt{t} + 1\right) \left(\sqrt{-\log \tilde{\rho}_0^{(m)}} + r_t^{(m)}\right) , \\
R_t^S(m) &\leq 8(2G^{(m)} \vee D^2)(3 + G^{(m)}) \left(\sqrt{-\log \tilde{\rho}_0^{(m)}} + r_t^{(m)}\right)^2 \\
&\qquad + \sqrt{-\log \tilde{\rho}_0^{(m)}} + r_t^{(m)} \, .
\end{aligned}
$$

where $r_t^{(m)} = \dfrac{\log\log\left(e^{1/4} + G^{(m)}\sqrt{t+1}\right)}{\sqrt{-\log\tilde{\rho}_0^{(m)}}}$.

Remark 4.1. The leading constant is proportional to ${G^{(m)}}^2$. It is not optimal and can be reduced to $G^{(m)}$ by refining the adaptive learning rates as in Cesa-Bianchi et al. (2007).

Proof. By adapting the inequality (14) as $(\eta_{t-1}^{(m)}\mathscr{L}_t^{(m)})_{1\le m\le M}$ is centered for $\left(\tilde{\rho}_t^{(m)}\right)_{1\le m\le M}$, where

$$\tilde{\rho}_t^{(m)} = \frac{\exp\left[-\eta_{t-1}^{(m)}\sum_{s=1}^{t-1}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right]\tilde{\rho}_0^{(m)}}{\sum_{m'=1}^{M}\exp\left[-\eta_{t-1}^{(m')}\sum_{s=1}^{t-1}\mathscr{L}_s^{(m')}\left(1+\eta_{s-1}^{(m')}\mathscr{L}_s^{(m')}\right)\right]\tilde{\rho}_0^{(m')}},$$

for any $t \ge 2$ we have

$$\sum_{m=1}^{M}\tilde{\rho}_0^{(m)}\exp\left[-\eta_{t-1}^{(m)}\sum_{s=1}^{t}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right] \le \sum_{m=1}^{M}\tilde{\rho}_0^{(m)}\exp\left[-\eta_{t-1}^{(m)}\sum_{s=1}^{t-1}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right]. \quad (16)$$

Since $x \le x^\alpha + \alpha^{-1}(\alpha-1)$ for $x \ge 0$ and $\alpha \ge 1$, by setting

$$\alpha = \frac{\eta_{t-2}^{(m)}}{\eta_{t-1}^{(m)}} \text{ and } x = \exp\left[-\eta_{t-1}^{(m)}\sum_{s=1}^{t-1}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right],$$

we have for any $t \ge 2$ the relation

$$\exp\left[-\eta_{t-1}^{(m)}\sum_{s=1}^{t-1}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right] \le \exp\left[-\eta_{t-2}^{(m)}\sum_{s=1}^{t-1}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right] + \frac{\eta_{t-2}^{(m)}-\eta_{t-1}^{(m)}}{\eta_{t-2}^{(m)}},$$

which leads to

$$\sum_{m=1}^{M}\tilde{\rho}_0^{(m)}\exp\left[-\eta_{t-1}^{(m)}\sum_{s=1}^{t}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right] \le \sum_{m=1}^{M}\tilde{\rho}_0^{(m)}\exp\left[-\eta_{t-2}^{(m)}\sum_{s=1}^{t-1}\mathscr{L}_s^{(m)}\left(1+\eta_{s-1}^{(m)}\mathscr{L}_s^{(m)}\right)\right] + \sum_{m=1}^{M}\tilde{\rho}_0^{(m)}\frac{\eta_{t-2}^{(m)}-\eta_{t-1}^{(m)}}{\eta_{t-2}^{(m)}}. \quad (17)$$

Using a recursion argument on $t \geq 2$ on Equation (17) yields

$$\sum_{m=1}^{M} \tilde{\rho}_0^{(m)} \exp\left[-\eta_{t-1}^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)} \left(1 + \eta_{s-1}^{(m)} \mathscr{L}_s^{(m)}\right)\right] \\ \leq \sum_{m=1}^{M} \tilde{\rho}_0^{(m)} \exp\left[-\eta_0^{(m)} \mathscr{L}_1^{(m)} \left(1 + \eta_0^{(m)} \mathscr{L}_1^{(m)}\right)\right] + \sum_{s=1}^{t-1} \sum_{m=1}^{M} \tilde{\rho}_0^{(m)} \frac{\eta_{s-1}^{(m)} - \eta_s^{(m)}}{\eta_{s-1}^{(m)}}. \quad (18)$$

Moreover, we have

$$\sum_{s=1}^{t-1} \frac{\eta_{s-1}^{(m)} - \eta_s^{(m)}}{\eta_{s-1}^{(m)}} \leq \sum_{s=1}^{t-1} \int_{\eta_s^{(m)}}^{\eta_{s-1}^{(m)}} \frac{dx}{x} \leq \int_{\eta_{t-1}^{(m)}}^{\eta_0^{(m)}} \frac{dx}{x} \leq \log\left(\frac{\eta_0^{(m)}}{\eta_{t-1}^{(m)}}\right)$$

the estimate of the ratio

$$\frac{\eta_0^{(m)}}{\eta_{t-1}^{(m)}} = \sqrt{1 + \sum_{s=1}^{t-1} {\mathscr{L}_s^{(m)}}^2} \leq G^{(m)} \sqrt{t+1}, \text{ for } G^{(m)} \geq 1.$$

and

$$-\eta_0^{(m)} \mathscr{L}_1^{(m)} \left(1 + \eta_0^{(m)} \mathscr{L}_1^{(m)}\right) \leq 1/4\,.$$

Equation (18) implies

$$-\eta_{t-1}^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)} \left(1 + \eta_{s-1}^{(m)} \mathscr{L}_s^{(m)}\right) \leq -\log \tilde{\rho}_0^{(m)} + r_t^{(m)},$$

and we obtain similarly than above that for any π we have

$$-\sum_{m=1}^{M} \pi^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)} \leq \sum_{m=1}^{M} \pi^{(m)} \left(\sum_{s=1}^{t} \eta_{s-1}^{(m)} {\mathscr{L}_s^{(m)}}^2 + \frac{-\log \tilde{\rho}_0^{(m)} + r_t^{(m)}}{\eta_{t-1}^{(m)}}\right).$$

In order to bound the second order term $\sum_{s=1}^{t} \eta_{s-1}^{(m)} {\mathscr{L}_s^{(m)}}^2$ we denote $V_t = 1 + \sum_{s=1}^{t} {\mathscr{L}_s^{(m)}}^2$ so that

$$\begin{aligned}
\eta_{s-1}^{(m)} {\mathscr{L}_s^{(m)}}^2 &= \sqrt{-\log \tilde{\rho}_0^{(m)}} \frac{V_s - V_{s-1}}{\sqrt{V_{s-1}}} \\
&= \sqrt{-\log \tilde{\rho}_0^{(m)}} \frac{\sqrt{V}_s + \sqrt{V}_{s-1}}{\sqrt{V_{s-1}}} (\sqrt{V}_s - \sqrt{V}_{s-1}) \\
&= \sqrt{-\log \tilde{\rho}_0^{(m)}} \left(\sqrt{V_s / V_{s-1}} + 1\right) (\sqrt{V}_s - \sqrt{V}_{s-1}) \\
&\leq \sqrt{-\log \tilde{\rho}_0^{(m)}} \left(\sqrt{1 + {G^{(m)}}^2} + 1\right) (\sqrt{V}_s - \sqrt{V}_{s-1})\,.
\end{aligned}$$

A telescoping sum argument yields

$$\sum_{s=1}^{t} \eta_{s-1}^{(m)} {\mathscr{L}_s^{(m)}}^2 \leq (2 + G^{(m)}) \sqrt{-\log \tilde{\rho}_0^{(m)}} \left(\sqrt{1 + \sum_{s=1}^{t} {\mathscr{L}_s^{(m)}}^2} - 1 \right)$$
$$\leq (2 + G^{(m)}) \sqrt{-\log \tilde{\rho}_0^{(m)} \sum_{s=1}^{t} {\mathscr{L}_s^{(m)}}^2}.$$

Finally we get

$$-\sum_{m=1}^{M} \pi^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)} \leq \sum_{m=1}^{M} \pi^{(m)} \left(\sqrt{\sum_{s=1}^{t} {\mathscr{L}_s^{(m)}}^2} \left((3 + G^{(m)}) \sqrt{-\log \tilde{\rho}_0^{(m)}} \right. \right.$$
$$\left. \left. + r_t^{(m)} \right) + \sqrt{-\log \tilde{\rho}_0^{(m)}} + r_t^{(m)} \right)$$

and the desired bound on R_t^A follows.

In order to obtain the regret bound on R_t^S we use the Young inequality $2\sqrt{ab} \leq \gamma a + b/\gamma$ with $\gamma = 4(2G^{(m)} \vee D^2)$ so that

$$-\sum_{m=1}^{M} \pi^{(m)} \sum_{s=1}^{t} \mathscr{L}_s^{(m)} \leq \sum_{m=1}^{M} \pi^{(m)} \left(\frac{1}{8(2G^{(m)} \vee D^2)} \sum_{s=1}^{t} {\mathscr{L}_s^{(m)}}^2 \right.$$
$$+ 8(2G^{(m)} \vee D^2)(3 + G^{(m)}) \left(\sqrt{-\log \tilde{\rho}_0^{(m)}} + r_t^{(m)} \right)^2$$
$$\left. + \sqrt{-\log \tilde{\rho}_0^{(m)}} + r_t^{(m)} \right)$$

and the desired result follows from an application of Theorem 3.1. □

5. Discussion and examples

KAO algorithms require the knowledge of the variances $\sigma^{2(m)} > 0$. A natural estimator of this quantity is the mean square residuals

$$\widehat{\sigma}_t^{2(m)} = \frac{1}{t} \sum_{s=1}^{t} \left(y_s - \hat{y}_s^{(m)} \right)^2.$$

It can be tuned online but without any guarantee on the regret of the corresponding algorithm. In our applications, we prefer to estimate $\widehat{\sigma}_t^{2(m)}$ on a burn-in period and use this fixed value in KAO.

5.1. Comparison with BOA

We start this Section with a short comparison with the BOA algorithm of Wintenberger (2017) that achieves similar regret bounds than the one obtained here. Theorem 4.2 in Wintenberger (2017) shows that BOA, which is an algorithm based on surrogate losses, has nice generalization properties that extend the regret bounds in the adversarial setting into similar regret bounds in the stochastic adversarial setting. The price to pay for the generalization is a factor 2 in the regret bounds. We show that this factor 2 is avoidable under assumption **(H)** with an algorithm such as KAO which uses the surrogate risks rather than the surrogate losses. Finally, notice that the use of the risk allows getting a.s. regret bounds in the well-specified stochastic unbounded setting rather than high-probability regret bounds only in bounded settings.

5.2. The static iid setting

In the iid setting, we consider an aggregation of static Kalman recursions with $P_0^{(m)} = 1/\lambda^{(m)} I$, $\lambda^{(m)} > 0$ which coincides with online ridge regression starting at $\hat{\theta}_0^{(m)}$. A natural estimator for σ^2 is the mean of the mean square residuals $M^{-1}\sum_{m=1}^{M} \hat{\sigma}_t^{(m)}$. This setting is very specific since $\mu_t = \underset{t-1}{\mathbb{E}}[y_t] = \mathbb{E}[y_t] = X_t^\top \theta_t^{(m)} = X_t^\top \theta^{(m)} = X_t^\top \theta^*$ for any $1 \le m \le M$ under **(H)** and some fixed $\theta^* \in \Theta$ corresponds to the well specified setting. Consider for a moment $D = \max_{t\ge 1} \max_{1\le m\le M} |X_t^\top(\theta^* - \hat{\theta}_t^{(m)})|$ as random. Moreover one can estimate $G^{(m)} = 4D^2$ such that, applying KAO with adaptive multiple learning rates we obtain the model selection regret bound

$$\sum_{s=1}^{t} L_s(\hat{y}_s) \le \min_{1\le m\le M} \Big(\sum_{s=1}^{t} L_s(\hat{y}_s^{(m)}) + O(-D^2 \log \tilde{\rho}_0^{(m)})) \Big) .$$

It is interesting to combine this bound with the regret bounds on the ridge regression when the design (X_t) is iid, bounded by X and such that $\mathbb{E}[X_t X_t^\top]$ has a positive lowest eigenvalue Λ_{min}. Applying Theorem 14 of de Vilmarest and Wintenberger (2020), the mth Kalman recursion achieves for any $\theta \in \mathbb{R}^d$ and any $t \ge 1$

$$\sum_{s=1}^{t} L_s(\hat{y}_s^{(m)}) \le \sum_{s=1}^{t} L_s(X_s^\top \theta) + O\Big({\lambda^{(m)}}^3 \|\theta - \hat{\theta}_0^{(m)}\|_2^6 + d \log\Big(\frac{t}{\lambda^{(m)}}\Big) + \log(\delta^{-1})^3 \Big) ,$$

with probability at least $1-\delta$. Moreover, the localization strategy of de Vilmarest and Wintenberger (2020) shows that under the same probability D can be considered as a constant. Then KAO achieves the regret bound in expectation, valid for any $\theta \in \mathbb{R}^d$ and any $1 \le m \le M$,

$$\sum_{s=1}^{t} L_s(\hat{y}_s) \le \sum_{s=1}^{t} L_s(X_s^\top \theta) + O\Big({\lambda^{(m)}}^3 \|\theta - \hat{\theta}_0^{(m)}\|_2^6 + d \log\Big(\frac{t}{\lambda^{(m)}}\Big) - D^2 \log \tilde{\rho}_0^{(m)} + \log(\delta^{-1})^3 \Big) ,$$

with probability $1 - M\delta$. Aggregation can be seen as an online alternative of cross-validation for tuning the starting point of the ridge regression algorithm and the regularization parameter.

As an illustration one should consider $\hat{\theta}_0^{(m)}$ may be taken equal to $\alpha(e_i)_{1\leq i\leq d}$ where $(e_i)_{1\leq i\leq d}$ is the canonical basis and α takes value on $[-d, d] \cap \mathbb{Z}$. Moreover $\lambda^{(m)}$ should be taken on an exponential d finite grid of $(0, \infty)$. The number of Kalman recursions is $M = O(d)$ and choosing uniform weights yields to a regret for any $\lambda > 0$ on the grid, any $1 \leq i \leq d$ and any $\alpha \in [-d, d] \cap \mathbb{Z}$ as

$$\sum_{s=1}^{t} L_s(\hat{y}_s) \leq \sum_{s=1}^{t} L_s(X_s^\top \theta) + O\Big(\lambda^3 \|\theta - \alpha e_i\|_2^6 + d \log\Big(\frac{t}{\lambda}\Big) + D^2 \log d + \log(\delta^{-1})^3\Big)\,,$$

with probability $1 - d\delta$. Other aggregation strategies on least-squares estimators are described in Leung and Barron (2006). Restrictions of our framework are the well-specification condition **(H)** and the presence of the large constant D^2 in the model selection bound. One clear advantage is an explicit online procedure whereas least square estimators require the inversion of inverse matrices at each batch step.

5.3. The dynamic setting

In the dynamic setting, we consider that (y_t) behaves as a centered random walk conditionally on the design. The Kalman recursions track the trajectory of the linear coefficients associated to the explanatory variables. Assume that the design is standardized such that $\mathbb{E}[{X_t^{(m)}}^2] = \mathbb{E}[{X_t^{(m')}}^2]$ for any $1 \leq m, m' \leq d$. Consider $M = d$ univariate Kalman recursions $d_m = 1$ with $K^{(m)} = Q^{(m)} = 1$. Then the random coefficients $\theta_t^{(m)}$ satisfies the relation (6) and constitutes a random walk. If there exist $D, X > 0$ satisfying $D = \max_{t\geq 1} |\hat{y}_t - \mu_t|$ and $|X_{t,m}| \leq X$ then one can bound, with high probability

$$\begin{aligned}\max_{1\leq t\leq T} \max_{1\leq m\leq d} |\mathscr{L}_t^{(m)}| &\leq 2DX \max_{1\leq t\leq T} \sum_{m=1}^{d} |\hat{\theta}_t^{(m)}| \\ &\leq CDX \left(\sum_{t=1}^{T} \left(\sum_{m=1}^{d} \mathbb{E}[(\hat{\theta}_t^{(m)})^2]^{1/2}\right)^2\right)^{1/2}\end{aligned}$$

for some high constant $C > 0$. Then we can apply the result of Guo (1994) asserting that $\mathbb{E}[(\hat{\theta}_t^{(m)} - \theta_t^{(m)})^2]^{1/2} \leq E$ for some $E > 0$. Together with the fact that $\mathrm{Var}(\theta_t^{(m)}) = t$ by definition we obtain

$$\max_{1\leq t\leq T} \max_{1\leq m\leq d} |\mathscr{L}_t^{(m)}| \leq CDXdT\,.$$

Then applying KAO with adaptive learning rate and doubling trick as in Remark 4.1 with $G = CDXdT$, we obtain with high probability the aggregation regret bound

$$\sum_{s=1}^{T} L_s(\hat{y}_s) \leq \sum_{s=1}^{T} L_s\left(\sum_{i=1}^{M} \pi^{(m)} \hat{y}_s^{(m)}\right) + O(DXT^{3/2} d \log d)\,.$$

This super-linear rate is due to the high fluctuations of the Kalman recursions when they track random walks $(\theta_t^{(m)})$. The Kalman recursions inherit the high variability of the random walks which is responsible for the high variability of the gradient and large $G = O(T)$. However, due to the unboundedness of the response, none of the existing regret bounds seem to apply in this setting.

5.4. The expert aggregation setting

The setting is similar to the previous one as $K^{(m)}$ is a diagonal matrix with non-null coefficients equals to 1. Thus one has to assume the boundedness of the gradients to get a $\sqrt{T}$ regret for the aggregation problem. It is the usual assumption in the setting of aggregation of experts and then the regret is essentially divided by a factor 2 compared with the regret bound obtained for BOA in Wintenberger (2017) under the boundedness of the response. It is worth mentioning again that the boundedness of the gradients of the conditional risk does not imply the boundedness of the response.

6. Simulation study

In this simulation study, we use some of the variables contained in the downloadable data set on the website of the RTE company (french TSO) that describes the hourly electricity consumption and production per type of production units in France from 2013 to 2017. We chose to simulate synthetic data from these true ones to be closer to a real application but controlling the true model at the same time. We generate synthetic data from a subset of these variables: the temperature, the gas production, the fuel production, the charcoal production, and the nebulosity. The square of the temperature and the cubic of the gas are jointly utilized as predictors in X_t to simulate, under a state-space model, the signal y_t that represents the electricity consumption. All the covariates are normalized to be in $[0, 1]$ by dividing each of them by their maximum value. The true model (that generates the true or the best expert) is a state-space model using the square of the temperature and the cubic of the gas as covariates in X_t and Gaussian noise. Regarding the parameters of this state-space model, $\sigma = 1.5$, Q is of values 1 on the diagonal and 0.9 otherwise, θ_0 is generated according to a gaussian law of mean 500 and covariance matrix identity, and K is the identity matrix. We also compute 27 other Kalman experts using other combinations of covariates that are different from those used for getting the true (or best) expert.

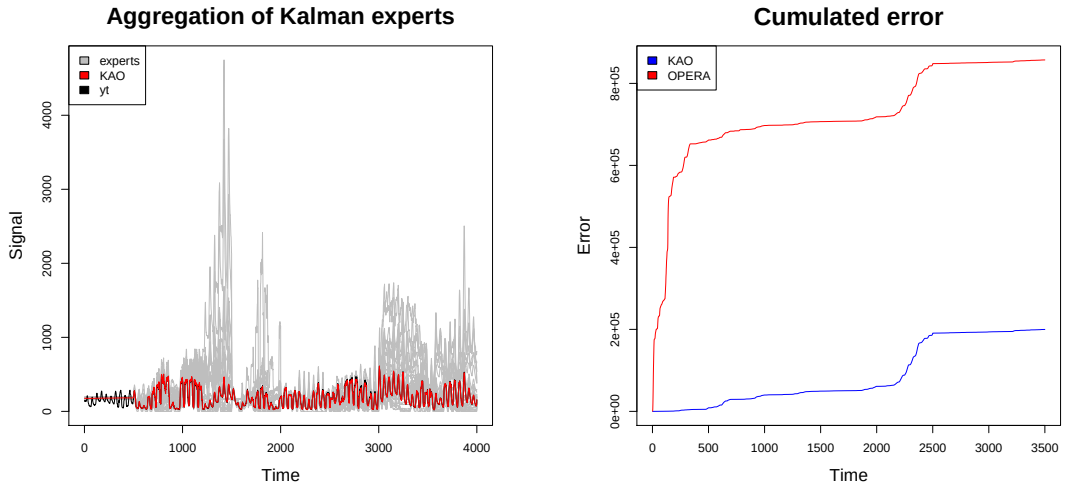

Figure 1: **One hour ahead prediction of y_t using KAO, and cumulated prediction errors for KAO and OPERA**. The left panel shows one hour ahead prediction of y_t using KAO and η within a grid of values. The value of η that minimizes the MSE is utilized to perform the prediction. These predictions are done in the case where the oracle is the best expert. The right panel shows the cumulated prediction errors for KAO and OPERA using the Kalman experts. the blue line represents the Kalman aggregation error, and the red one represents the error of the aggregation coming from the opera package. These predictions are done in the case where the oracle is the best expert.

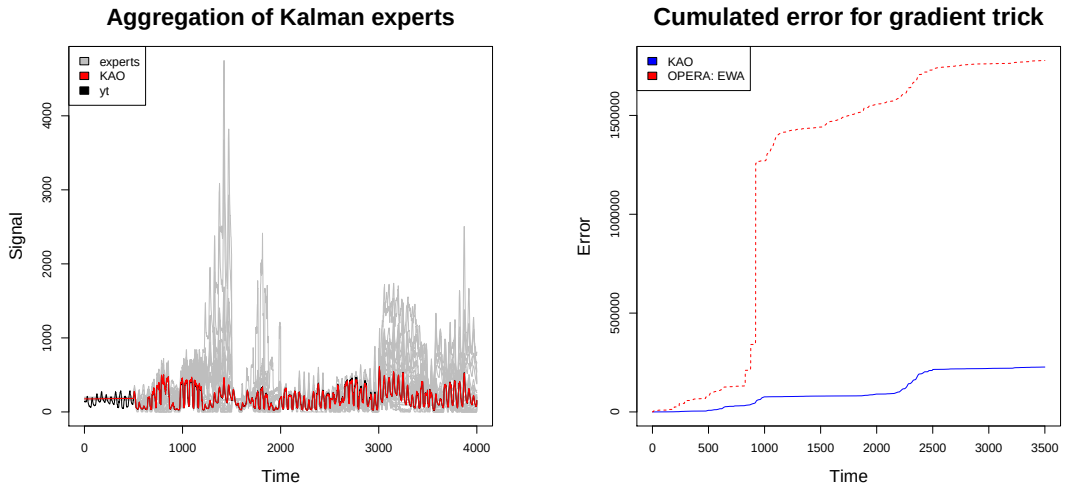

Figure 2: **One hour ahead prediction of y_t using KAO, and cumulated prediction errors for KAO and OPERA**. The left panel shows one hour ahead prediction of y_t using KAO and η within a grid of values. The value of η that minimizes the MSE is utilized to perform the prediction. These predictions are done in the case where the oracle is the best expert. The right panel shows the cumulated prediction errors for KAO and OPERA using the Kalman experts. the blue line represents the Kalman aggregation error, and the red one represents the error of the aggregation coming from the opera package. These predictions are done in the case where the oracle is the best convex combination of the Kalman experts.

Each Kalman expert is computed in the sequential way as follows. We begin by fitting the model using the first observations $(y_1, \ldots, y_{\text{window}})$ contained in a window. Then the fitted model is utilized to predict the observations contained in a window ahead (i.e., $y_{\text{window}+1}, \ldots, y_{2\text{window}}$).

At the pth step we use the observations $y_1, \ldots, y_{p\text{window}}$ to fit the model that is utilized to predict $y_{p\text{window}+1}, \ldots, y_{(p+1)\text{window}}$. We chose window $= 500$ as a good trade-off between a correct number of observations to estimate the state-space models and a good adaptation to changes. The prediction resulting from this procedure is called the Kalman expert and we, therefore, have 28 Kalman experts.

Simulations are done under the R software (R Core Team, 2019) and the predictive performance of the Kalman experts aggregation using KAO is compared with the aggregation performed using the R package *opera* (Gaillard and Goude, 2016) and the aggregation procedures therein. The aggregation obtained from the package opera is named OPERA when we are competing with the best expert and do not want to mention any specific aggregation procedure. We make one hour ahead prediction using KAO on the 28 Kalman experts. In the case where the oracle is the best Kalman expert, the resulting prediction is plotted by a red curve in Figure 1 at the left panel, where the signal y_t is plotted by a black curve, and the experts are plotted using the gray color. We can see that the red line tracks well the black one, meaning that the aggregation from KAO performs well its prediction. More precisely, the MSE of KAO is 66.507 which is approximately equal to the MSE of the best Kalman expert (66.503), and the MSE of OPERA is 253.06. The right panel of Figure 1 shows the cumulated error of KAO (in blue color) and OPERA (in red color). We can see that KAO performs better than OPERA. Though both KAO and OPERA (precisely, EWA or BOA procedure) are based on exponential weights, the difference seen in their respective cumulated errors can be explained by the fact that KAO takes into account the underlying models that provide the experts, and OPERA doesn't have this information.

In the case where the oracle is the best convex combination of the Kalman experts, the one hour ahead predictions of y_t, using KAO, are plotted in the left panel of Figure 2 in red color and the Kalman experts are plotted in gray color. We can also see that KAO tracks well the signal y_t that is plotted in black color. Here, the MSE of KAO is 65.02 against 223.37 for OPERA, using the procedure BOA (Wintenberger, 2017). The corresponding cumulated errors are plotted in the right panel for KAO (in blue color) and OPERA (in red color). The curves of the cumulated errors show that KAO has a better predictive performance than OPERA.

We simulate 100 collections of Kalman experts corresponding to 100 simulated datasets. Each collection of Kalman experts contains 28 different experts. We then perform the aggregation of each collection of Kalman experts using KAO and the procedures within OPERA for each type of oracle. The MSE of the aggregations are computed and plotted in Figure 3. The left panel (Figure 3(a)) shows the curves of the MSE of the aggregations performed in the case where the oracle is the best expert. KAO (dashed blue curve) presents the lowest MSE within all the aggregation procedures, followed by MLpoly and EWA. The right panel (Figure 3(b)) shows the aggregations' MSE in the case where the oracle is the best convex combination of the experts.

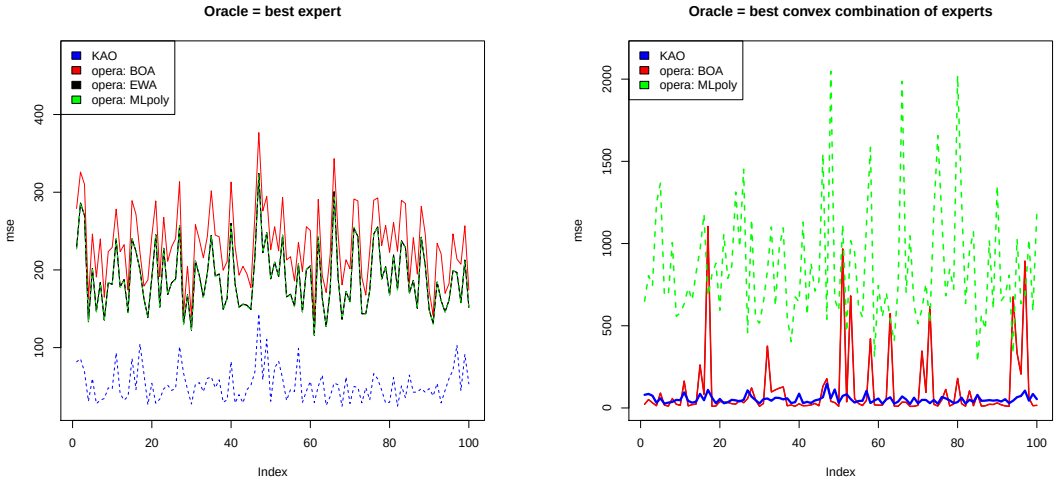

Figure 3: **MSE of** 100 **aggregations of Kalman experts using KAO and the procedures contained in opera.** The left panel shows the curves of computed mse, where each aggregation procedure competes with the best expert. KAO is plotted in blue color, BOA in red color, EWA (Exponentially Weighted Average) in black and MLpoly (Gaillard et al., 2014) in green color. The right panel shows the curves of the mse computed when the aggregation competes with the best convex combination of experts. KAO is in blue color, BOA in red color and MLpoly in green color.

We can see that KAO (blue curve) has not only the best MSE but also presents more stability than BOA (red curve) and MLpoly (dashed green curve). Here, reversely to the case where the aggregation competes with the best expert, BOA is better than MLpoly. This simulation study seems to point out that it may be worth of interest to take into account the underlying model that generates the experts when aggregating them.

Table 1: **Root Mean Square Error square of different aggregation procedures (relative to RMSE of the best convex combination). Kalman Experts.**

Procedure	rmse (with GT)	rmse (without GT)
Best expert	1.15	1.15
Uniform	1.11	1.11
MLpoly	1.06	1.16
BOA	1.07	1.11
KAO	1.05	1.07
Best convex	1	1

7. Application

In this section we apply the KAO algorithm to aggregate ten experts $f_{m,t}, 1 \leq m \leq M$ that are meant to predict the daily electricity consumption in France (see Ba et al. (2012), Gaillard and Goude (2014) for previous work on french load data) at times $t \in (1, 2, ..., T)$. These experts are

Table 2: **Root Mean Square Error square of different aggregation procedures (relative to RMSE of the best convex combination). AR experts.**

Procedure	rmse (with GT)	rmse (without GT)
Best expert	1.18	1.18
Uniform	1.11	1.11
MLpoly	1.07	1.19
BOA	1.07	1.09
Best convex	1	1

provided by different models that are black boxes. Thus, we consider the expert setting previously defined in 2.2. For each expert we stack their predictions $f_{m,t} \in \mathbb{R}$ in $X_t^{(m)}$ together with the intercept and the past error $e_{m,t-1} = (y_{t-1} - f_{m,t-1})$, i.e.,

$$X_t^{(m)} = (1, f_{m,t}, e_{m,t-1}), \qquad t \geq 1\,.$$

and each state-space model m is defined by the state equation:

$$\theta_t^{(m)} = \theta_{t-1}^{(m)} + z_t^{(m)}\,, \qquad t \geq 1.$$

the covariance matrices $Q^{(m)}, 1 \leq m \leq M$ and the variance of the noise $\sigma^{2(m)}$ are estimated using an EM algorithm on the first half of the data ($t \in (1, 2, ..., T/2)$) and we use the second half to evaluate KAO performances and compare it to other aggregation rules.

The predictive risk of $\hat{y}_t^{(m)} = X_t^{(m)}\theta_t^{(m)}$ is used for computing the loss and the pseudo-loss that are needed to perform KAO. The aggregation performance of KAO (on these experts) is compared with that of both MLpoly and BOA that are two aggregation procedures available in the opera package. The results are contained in Table 1 where GT means Gradient Trick. GT, therefore, refers to the case where the oracle of the aggregation procedure is the experts' best convex combination. For confidentiality reasons, errors are expressed relatively to the RMSE of the best convex combination.

The *uniform* procedure is the experts mean, and the procedure *best convex* is indeed the experts' best convex combination. All of these procedures are performed on the corrected experts. We clearly see that KAO performs slightly better (rmse = 1.05 with GT and rmse = 1.07 without GT) than both MLpoly (rmse = 1.06 with GT and rmse = 1.16 without GT) and BOA (rmse = 1.07 with GT and rmse = 1.11 without GT). In order to check if the Kalman correction is worth of interest, we make a direct autoregressive correction of the experts that are then aggregated using MLpoly and BOA (not KAO as we need an estimate of the risk for that). The results are contained in Table 2 and show that all the procedures are less accurate when the Kalman correction is not

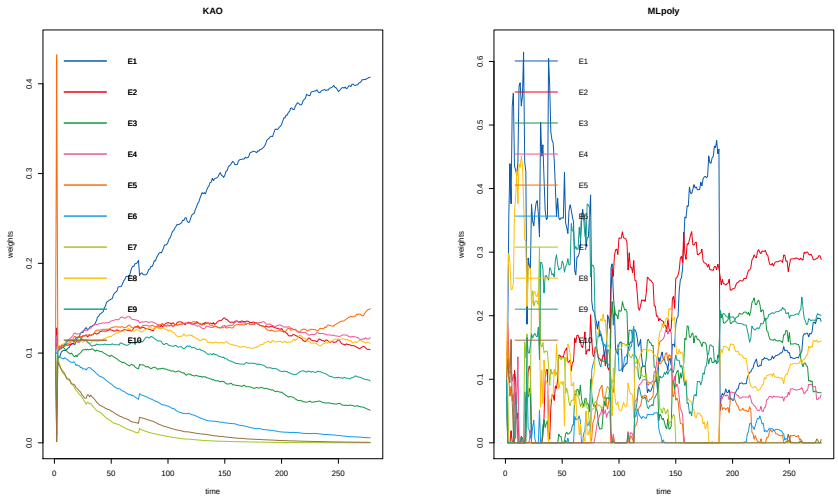

Figure 4: **Experts weight according to KAO and MLpoly, using the gradient trick**. The left panel shows the weights assigned to the corrected experts by KAO where the oracle is the best experts combination. The right panel shows the weights assigned by MLpoly, using the gradient trick. The experts are denoted by $E1, \ldots, E10$.

applied.

The weights that are assigned to the corrected experts (by KAO and MLpoly) are plotted in Figure 4. The weights coming from KAO are more smooth (see Figure 4(a)) than those provided by MLpoly (see Figure 4(b)). This smoothness of KAO weights can be explained by the fact that the procedure uses the underlying properties of the model that provide the experts. This information is used to anticipate the forthcoming performance of each expert.

8. Conclusion

In this paper, we show that the prediction obtained by aggregating the predictions coming from a finite set of experts can be improved by taking into account the properties of the underlying models that provide the experts' prediction. We place ourselves in the case where all the predictions provided by the experts come from fitting state-space models using Kalman recursions. By using exponential weights, two settings are considered: 1) the aggregation competes with the best expert (also considered as model selection), and 2) the aggregation competes with the best convex combination of the experts. We consider adaptive multiple learning rates in order to achieve the optimal rates in these two schemes for a unique procedure. The quality of the aggregation's prediction has been improved by taking advantage of the full knowledge of the Kalman experts, using their predictive risk in an unbounded well-specified setting. In the simulations studies, we notice a great recovery of stability of KAO (our aggregation procedure), where all other existing aggregation procedures may be sometime somewhat unstable, potentially due to the lack of boundedness of the responses. The aggregation procedure KAO is also applied to some existing

experts coming from unknown models, where we suggest correcting the errors of the experts using Kalman recursions. This strategy allows for approximating the theoretical weights needed for KAO and shows a quite important increase in the accuracy of the aggregation. In the case where the errors of the experts show no stationary behavior (for example, when there exist some cluster of variance), it should be interesting to adapt the fitting of the underlying state-space model in order to remain accurate.

References

Audibert, J.Y., Bubeck, S., 2010. Regret bounds and minimax policies under partial monitoring. Journal of Machine Learning Research 11, 2785–2836.

Ba, A., Sinn, M., Goude, Y., Pompey, P., 2012. Adaptive learning of smoothing functions: Application to electricity load forecasting, in: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems 25, pp. 2519–2527. URL: `http://books.nips.cc/papers/files/nips25/NIPS2012_1205.pdf`.

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

Cesa-Bianchi, N., Mansour, Y., Stoltz, G., 2007. Improved second-order bounds for prediction with expert advice. Machine Learning 66, 321–352.

Diderrich, G.T., 1985. The kalman filter from the perspective of goldberger?theil estimators. The American Statistician 39, 193–198.

Durbin, J., Koopman, S.J., 2012. Time series analysis by state space methods. Oxford university press.

Gaillard, P., Goude, Y., 2014. Forecasting electricity consumption by aggregating experts; how to design a good set of experts. to appear in Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension .

Gaillard, P., Goude, Y., 2016. opera: Online Prediction by Expert Aggregation. URL: `https://CRAN.R-project.org/package=opera`. r package version 1.0.

Gaillard, P., Stoltz, G., Van Erven, T., 2014. A second-order bound with excess losses, in: Conference on Learning Theory, pp. 176–196.

Gaillard, P., Wintenberger, O., 2016. Sparse accelerated exponential weights. arXiv preprint arXiv:1610.05022 .

Gaillard, P., Wintenberger, O., 2018. Efficient online algorithms for fast-rate regret bounds under sparsity, in: Advances in Neural Information Processing Systems, pp. 7026–7036.

Guo, L., 1994. Stability of recursive stochastic tracking algorithms. SIAM Journal on Control and Optimization 32, 1195–1225.

Hazan, E., et al., 2016. Introduction to online convex optimization. Foundations and Trends® in Optimization 2, 157–325.

Leung, G., Barron, A.R., 2006. Information theory and mixing least-squares regressions. IEEE Transactions on Information Theory 52, 3396–3410.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

de Vilmarest, J., Wintenberger, O., 2020. Stochastic online optimization using kalman recursion. arXiv preprint arXiv:2002.03636 .

Vovk, V.G., 1990. Aggregating strategies. Proc. of Computational Learning Theory, 1990 .

Wintenberger, O., 2017. Optimal learning with bernstein online aggregation. Machine Learning 106, 119–141.