Long-Term Sequential Prediction Using Expert Advice

Eugeny Burnaev  
Alexander Korotin  
Vladimir V’yugin

Skolkovo Institute of Science and Technology  
Institute for Information Transmission Problems, Moscow, Russia

Abstract

For the prediction with expert advice setting, we consider methods to construct forecasting algorithms that suffer loss not much more than any of experts in the pool. In contrast to the standard approach, we investigate the case of long-term interval forecasting of time series, that is, each expert issues a sequence of forecasts for a time interval ahead and the master algorithm combines these forecasts into one aggregated sequence of forecasts. Two new approaches for aggregating experts long-term interval predictions are presented. One is based on Vovk’s aggregating algorithm and considers sliding experts, the other applies the approach of Mixing Past Posteriors method to the long-term prediction. The upper bounds for regret of these algorithms for adversarial case are obtained. We also present results of numerical experiments of time series long-term prediction.

Keywords: prediction with expert advice, long-term interval forecasting, sliding experts, aggregating algorithm, mixable game, Mixing Past Posteriors

1. Introduction

We consider the game-theoretic on-line learning model in which a master (aggregating) algorithm has to combine predictions from a set of $N$ experts (see e.g. Littlestone and Warmuth 1994, Freund and Schapire 1997, Vovk 1990, Vovk 1998, Cesa-Bianchi and Lugosi 2006 among others). In contrast to the standard approach, we investigate the case of interval long-term forecasting, that is, at each time step $t$ experts issue a sequence of forecasts for (integer) time interval $[t+1, t+d]$ ahead and, after that, the master algorithm combines these forecasts into one aggregated sequence of forecasts. The performance of the aggregating algorithm is measured by the regret.

The long-term forecasting of time series is often required in practice. In this case, the square loss function is widely used. This function is mixable in the sense of Vovk (1998). In classical one-step ahead forecasting games ($d=1$) with mixable loss functions, the Vovk’s aggregating algorithm from Vovk (1998) is the most appropriate, since it has a time-free upper bound for the regret with respect to the best expert.

The long-term forecasting is a case of the forecasting with a delayed feedback. As far as we know, the problem of the delayed feedback forecasting was first considered by Weinberger and Ordentlich (2002) and further developed by Langford et al. (2009), Mesterharm (2007), and Mesterharm (2009). Weinberger and Ordentlich (2002) investigated the adversarial setting with a fixed and known feedback delay $d$. Their main result is that the optimal solution is to run $d$ independent predictors on $d$ disjoint time grids $G_i = \{t \mid t \equiv i \pmod{d}\}$.
for \( i \in \overline{1,d} \). The most recent work is Adaptskiy et al. (2017) where a number of algorithms with tight upper bounds for predicting packs (vector valued outcomes) were presented.

The approach of Weinberger and Ordentlich (2002) serves as the starting point of our research. We take aggregating algorithm of Vovk (1998) as the base and modify it for the \( d \) steps ahead long-term interval forecasting. Despite the fact that the approach with learning on separate grids is theoretically optimal in the adversarial setting, in practice one can obtain better results if takes into account the relationship between learning processes on different time grids. We provide two ways to do this. Our first approach is called sliding experts; it takes into account all predictions available for a given time moment, and thus allows to use experts answers on different time grids. Our second approach is based on the Mixing Past Posteriors (MPP) method by Bousquet and Warmuth (2002) and mixes the experts weights on different grids to achieve better (practical) performance.

In Section 2 some preliminary notions are given. In Section 3 we present algorithms which take into account results of learning on different time grids. Subsections 3.1 and 3.2 are related to sliding experts. First, we consider the artificial case of experts predicting infinite future (Subsection 3.1) and use its results for more realistic case of experts whose predictions overlap (Subsection 3.2). In Subsection 3.3 we use modification of the method of MPP to connect separate time grids. The upper bounds for the regret of these algorithms are presented which gives us guaranties for the adversarial case. Results of numerical experiments of the long-term interval prediction aggregation are provided in Section B.

2. Preliminaries

We consider the Prediction with Expert Advice (PEA) framework in which a master (aggregating) algorithm has to combine predictions from a set of experts. Let a pool of \( N \) experts be given. Suppose that elements \( \omega_1, \omega_2, \ldots \) of a (scalar) time series are revealed online – step by step. Learning proceeds in trials \( t = 1, \ldots, T \). In each time moment \( t \) the experts \( i \in \{1, \ldots, N\} \) present their predictions and the aggregating algorithm presents its own forecast. When the corresponding outcome(s) will be revealed all players suffer their losses using a loss function. Let \( l_t = (l_{1t}, \ldots, l_{Nt}) \) be a vector of experts losses at time \( t \) and \( h_t \) be a loss of the aggregating algorithm.

The cumulative loss suffered by any expert \( i \) for the first \( T \) steps is defined as \( L^i_T = \sum_{t=1}^{T} l^i_t \) and the cumulative loss of the aggregating algorithm is \( H_T = \sum_{t=1}^{T} h_t \). The important characteristics of the algorithm is a regret over \( T \) rounds \( R^i_T = H_T - L^i_T \) with respect to an expert \( i \) and the regret with respect to the best expert \( R^*_T = \max_{1 \leq i \leq N} R^i_T \).

Following Kivinen and Warmuth (1999) we will use a more general notion of regret. The aggregating algorithm will compete not only with single experts, but with alternating sequences of experts. A comparison vector is a vector \( q = (q_1, \ldots, q_N) \) such that \( \|q\|_1 = 1 \) and \( q \geq 0 \) componentwisely. We compare the cumulative loss of the aggregating algorithm \( H_T \) and cumulative convex combination of losses of the experts \( \sum_{t=1}^{T} (q \cdot l_t) \) that is, \( R_T = H_T - \sum_{t=1}^{T} (q \cdot l_t) \) for a given \( q \). Note that \( R_T = R^*_T \) when \( q \) is unit vector, whose \( i \)-th coordinate is 1.
The goal of the aggregating algorithm is to minimize the regret. In order to achieve this goal, the aggregating algorithm evaluates performance of the experts in the form of a vector of weights $w_t = (w_{1,t}, \ldots, w_{N,t})$ assigned to all $N$ experts, where $\sum_{i=1}^{N} w_{i,t} = 1$ and $w_{i,t} \geq 0$ for all $i$. The weight $w_{i,t}$ of an expert $i$ is an estimate of the “quality” of the expert’s predictions at step $t$. In classical setting (see Freund and Schapire 1997, Vovk 1990 among others), the process of expert $i$ weights updating is based on the method of exponential weighting with a learning rate $\eta > 0$:

$$w_{i,t}^{m} = \frac{w_{i,t} e^{-\eta l_i^t}}{\sum_{j=1}^{N} w_{j,t} e^{-\eta l_i^t}},$$

where $w_1$ is some weight vector, for example, $w_1 = (\frac{1}{N}, \ldots, \frac{1}{N})$.

In case of the next outcome prediction we define $w_{t+1} = w_{t}^{m}$. In a more general case of the $d$th outcome ahead prediction, where $d \geq 1$, we set $w_{t+d} = w_{t}^{m}$.

Relative entropy is a basic tool for the regret analysis; it can be considered as a measure of progress. For any $n$ denote by $\Gamma_n$ the simplex of all probability distributions on a set of cardinality $n$. Let $D(p||q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$ be the relative entropy, where $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ are elements of $\Gamma_n$. We define $0 \ln 0 = 0$.

We use the notions of mixloss $m_t = -\frac{1}{\eta} \ln \sum_{i=1}^{N} w_{i,t} e^{-\eta l_i^t}$ and cumulative mixloss $M_T = \sum_{t=1}^{T} m_t$. We will show that the last quantity is less than the cumulative convex combination of the experts’ losses. The proof is based on a similar lemma from Bousquet and Warmuth (2002).

**Lemma 1** For any $t$, and for a comparison vector $q \in \Gamma_N$,

$$m_t = (q \cdot l_t) + \frac{1}{\eta} (D(q||w_t) - D(q||w_t^m)).$$

**Proof.** By method (1) of the weights update equality (2) is obtained as follows:

$$m_t - \sum_{i=1}^{N} q_i l_i^t = \sum_{i=1}^{N} q_i \left( \frac{1}{\eta} \ln e^{-\eta l_i^t} + m_t \right) = \frac{1}{\eta} \sum_{i=1}^{N} q_i \left( \ln e^{-\eta l_i^t} - \ln \sum_{j=1}^{N} w_{j,t} e^{-\eta l_i^t} \right) =$$

$$= \frac{1}{\eta} \sum_{i=1}^{N} q_i \ln \frac{\sum_{j=1}^{N} w_{j,t} e^{-\eta l_i^t}}{w_{i,t} e^{-\eta l_i^t}} = \frac{1}{\eta} \sum_{i=1}^{N} q_i \ln \frac{w_{i,t}^m}{w_{i,t}} = \frac{1}{\eta} (D(q||w_t) - D(q||w_t^m)).$$

Lemma is proved. $\triangle$

We apply Lemma 1 for the long-term prediction. Let a comparison vector $q = (q_1, \ldots, q_N)$ and a sequence $l_t = (l_1^t, \ldots, l_N^t)$ of losses of the experts be given online for $t = 1, 2, \ldots$. Let
\(d \geq 1\) and \(l_{i,s} = 0\) for \(1 \leq i \leq N\) and \(0 \leq s \leq d\). Summing (2) for \(1 \leq t \leq T\), we obtain

\[
M_T = \sum_{t=1}^{T} m_t = \sum_{t=1}^{T} (q \cdot l_t) + \sum_{t=1}^{T} \frac{1}{\eta} (D(q\|w^m_t) - D(q\|w^m_m)) = T \sum_{t=1}^{T} (q \cdot l_t) + \frac{1}{\eta} \sum_{t=d+1}^{T} (D(q\|w^m_{t-d}) - D(q\|w^m_m)) = \sum_{t=1}^{T} (q \cdot l_t) + \frac{1}{\eta} \sum_{t=1}^{d} D(q\|w^m_t) \leq \sum_{t=1}^{T} (q \cdot l_t) + \frac{1}{\eta} d \max_{1 \leq i \leq N, 1 \leq s \leq d} \ln \frac{1}{w_{i,s}^m}.
\]

In transition from (3) to (4) we have used equality \(w_t = w^m_{t-d}\) for \(d < t \leq T\). In transition from (4) to (5) positive and the corresponding negative terms telescope and only first \(d\) positive terms remain. Also, we have used the inequality \(D(q\|p) = \sum_{i=1}^{N} q_i \ln \frac{q_i}{p_i} = \sum_{i=1}^{N} q_i \ln q_i + \sum_{i=1}^{N} q_i \ln \frac{1}{p_i} \leq \max_{1 \leq i \leq N} \ln \frac{1}{p_i} \) for all probability vectors \(q\) and \(p\). In particular, the last term of this inequality is \(\frac{1}{\eta} d \ln N\) when \(w_{i,s}^m = 1/N\) for all \(i\) and \(1 \leq s \leq d\).

In classical PEA setting at each time moment \(t = 1, 2, \ldots\) the experts \(i = 1, \ldots, N\) present their forecasts \(\xi^t_i\) and then the aggregating algorithm computes its forecast \(\gamma_t\). After that an outcome \(\omega_t\) is revealed; the experts and the algorithm suffer their losses: \(l^t_i = \lambda(\omega_t, \xi^t_i)\) and \(h^t_t = \lambda(\omega_t, \gamma_t)\), where \(\lambda(\omega, \gamma)\) is some loss function.

We will use a specific case of PEA – learning with mixable loss functions and the Vovk’s aggregating algorithm (Vovk 1990, Vovk 1998). The main tool is a superprediction function by Vovk (1998)

\[
g(\omega) = -\frac{1}{\eta} \ln \int e^{-\eta \lambda(\omega, \xi(i))} dp(i),
\]

where \(p\) is a probability distribution on the set of all experts and \(\xi^t_i\) are the experts predictions. Although all considerations below are valid for infinite set of experts elsewhere below except of Section 3.1 we use finite sets of \(N\) experts.

A loss function \(\lambda(\omega, \gamma)\) is \(\eta\)-mixable, where \(\eta > 0\), if for any probability distribution \(p\) and any set of experts predictions \(\gamma\) exist such that \(\lambda(\omega, \gamma) \leq g(\omega)\) for all \(\omega\). We write \(\gamma = \Sigma(g)\), where \(\Sigma\) is called a substitution function. In what follows we suppose that a mixable loss function \(\lambda(\omega, \gamma)\) is given. For more details see Appendix A.

In case of one point long-term forecasting at any time moment \(t\), each expert \(i\) from the pool presents its forecast \(\xi^t_{i+t}\) for testing at a future time moment \(t + d\), where \(d \geq 1\). We use the weight update rule (1) and define the vector of the experts weights \(w_{t+d} = w^m_t\). After that, the aggregating algorithm presents its own forecast \(\gamma_{t+d}\) for the same time moment \(t + d\).

The corresponding outcome \(\omega_{t+d}\) will be revealed later at a time moment \(t + d\) and all players suffer losses \(\lambda(\omega_{t+d}, \xi^t_{i+t})\) and \(\lambda(\omega_{t+d}, \gamma_{t+d})\).

In other words, at any time moment \(t > d\) an expert \(i\) suffers the loss \(l^t_i = \lambda(\omega_t, \xi^t_i)\) and the aggregating algorithm suffers the loss \(h^t_t = \lambda(\omega_t, \gamma_t)\), where forecasts \(\xi^t_i\) and \(\gamma_t\) were
issued at the time moment $t - d$. Assume that at steps $1 \leq t \leq d$ no forecasts are given and experts and aggregating algorithm suffer zero loss: $l^t_i = h_t = 0$.

If the forecast $\gamma_t$ was defined at the time moment $t - d$ such that $\lambda(\omega, \gamma_t) \leq g(\omega)$ for all $\omega$ then by mixability $h_t \leq m_t$ for all $t$. Then an upper bound for the regret is given by (5).

In case of interval long-term forecasting, at any time moment $t$, each expert $i$ presents a vector $(\xi^t_{i+1}, \ldots, \xi^t_{i+d})$ of predictions for the time interval $[t+1, t+d]$. The aggregating algorithm mixes all such vectors into its own vector-valued forecast $\gamma_{t+d} = (\gamma_{t+1,1}, \ldots, \gamma_{t+d,d})$. When the outcomes $\omega_{t+1}, \ldots, \omega_{t+d}$ will be completely revealed at the time moment $t + d$ each expert $i$ suffers a loss $h^i_{t+d} = \sum_{s=1}^d \lambda(\omega_{t+s}, \xi^t_{i+s})$ and the aggregating algorithm suffers a loss $h_{t+d} = \sum_{s=1}^d \lambda(\omega_{t+s}, \gamma_{t+s,s})$.

Following Adamskiy et al. (2017), we apply the aggregation rule to each coordinate $d$ separately: since the loss function is $\eta$-mixable, for each $1 \leq s \leq d$ a prediction $\gamma_{t+d,s}$ exists such that $e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \geq \sum_{i=1}^N e^{-\eta \lambda(\omega, \xi^t_{i+s})} w_{i,t}$ for all $\omega$. Then by (31) from Appendix A

$$e^{-\frac{q}{d} h_{t+d}} \geq \sum_{i=1}^N w_{i,t} e^{-\frac{q}{d} \sum_{s=1}^d \lambda(\omega, \xi^t_{i+s})} = \sum_{i=1}^N w_{i,t} e^{-\frac{q}{d} h^i_{t+d}}.$$

Hence, the vector-valued game is $\frac{q}{d}$-mixable.

Let us apply Lemma 1 for the vector-valued game, where the weight update rule (1) is used with the learning rate $\frac{q}{d}$. By (7) $h_t \leq m_t$ for all $t$ and then the bound (5) becomes

$$H_T \leq \sum_{t=1}^T (q \cdot l_t) + \frac{1}{\eta} \frac{d^2}{1 \leq t \leq N, 1 \leq s \leq d} \max_{1 \leq t \leq N} \frac{1}{w_{i,s}^m} l^t_i,$$

where $q$ is a comparison vector and $l_t = (l^t_1, \ldots, l^t_N)$ is the vector of the experts losses: $l^t_i = \sum_{s=1}^d \lambda(\omega_t, \xi^t_{i,s})$.

In both cases of $d$ steps ahead prediction we set $w_t = w_{t-d}^m$. Thus we obtain an algorithm for the long-term forecasting based on the one from Vovk (1998). Our approach runs $d$ independent optimal (one-step ahead) predictors on $d$ non-overlapping subsequences $G_i = \{t : t = kd + i - 1, k = 1, 2, \ldots \}$ such that the $i$th predictor is used at time instants $t$ with $t \in G_i$. The sets $G_i$, $0 \leq i \leq d - 1$ form a grid on the time line.

### 3. Grid connection

In both methods of $d$ steps ahead prediction of Section 2 the game of prediction is divided in $d$ independent games played on $d$ non-intersecting time grids. Although Weinberger and Ordentlich (2002) showed that such algorithms for long-term forecasting are optimal in the adversarial case, in practice one can obtain better results if one takes into account the relationship between learning processes on different time grids. In this section we present two approaches to take into account this dependence without losing the guarantee for the adversarial case.

#### 3.1. Aggregating experts predicting infinite future

Let us consider a case where forecasting algorithms can present forecasts for unlimited number of future time moments. In this section we consider a simplified setting, where
at each time moment $t$ each expert $i \in \{1, \ldots, N\}$ presents an infinite sequence $p_i^t = (p_{i,1}^t, p_{i,2}^t, \ldots)$ of forecasts for time moments $t + 1, t + 2, \ldots$. The goal of the aggregating algorithm is for each time moment $t$ to present a vector of $d$ forecasts ahead based on all experts’ forecasts available up to this time moment.

At any time moment $t$ we observe the forecasts $p_i^t$ issued by the experts $1 \leq i \leq N$ at time moments $1 \leq s \leq t$. Using these predictions, the aggregating algorithm presents a vector $\gamma_t = (\gamma_{t+1}, \gamma_{t+2}, \ldots, \gamma_{t+d})$ of forecasts for $d$ time moments ahead: $t + 1, \ldots, t + d$.

At the time moment $t + d$ the outcomes $\omega_{t+1}, \omega_{t+2}, \omega_{t+d}$ will be revealed and the aggregating algorithm suffers the loss $\sum_{s=1}^{d} \lambda(\omega_{t+s}, \gamma_{t+d,s})$. Any expert $i$ issuing its forecast at a time moment $j \leq t$ suffers a loss $\sum_{s=1}^{d} \lambda(\omega_{t+s}, p_{j,t-j+s}^i)$ on the time interval $[t + 1, t + d]$, where $\lambda(\omega, \gamma)$ is a loss function.

We will take into account the relationship between learning processes on different grids by aggregating the forecasts of all experts and their shifted copies. We multiply any expert $i$ into an infinite sequence of the auxiliary experts $(i, j)$, where $1 \leq j < \infty$. At each time moment $j$, an expert $(i, j)$ presents a vector of forecasts $p_{t+d}^{i,j} = (p_{i,t+1}^{(i,j)}, \ldots, p_{i,t+d}^{(i,j)})$, where for $1 \leq j \leq t$, $p_{t+d}^{i,j} = p_{j,t-j+s}$ is a forecast of the expert $i$ for the time moment $t + s$ issued at the time moment $j$. For $j > t$, we artifically put $p_{t+d}^{i,j} = \gamma_{t+d}$.

We will define each component $\gamma_{t+d,s}$, $1 \leq s \leq d$, of the forecast of the generalized expert $(i, j)$ separately such that

$$e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \geq \sum_{i=1}^{N} \sum_{j=1}^{\infty} e^{-\eta \lambda(\omega, p_{i,j}^{(i,j)})} w_{(i,j),t}$$

for all $\omega$. Here $w_{(i,j),t}$ is the weight of the expert $(i, j)$ at a time moment $t$. Define $w_{(i,j),s} = \frac{1}{N} \nu(j)$ for $1 \leq s \leq d$, where $\nu(j) > 0$ for all $j$ and $\sum_{j=1}^{\infty} \nu(j) = 1$. In what follows $\nu(j) = \frac{1}{j(j+1)}$. The weights updating rule is given in Algorithm 1 below.

Rewrite inequality (8) in a more detailed form. For any $1 \leq s \leq d$,

$$e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \geq \sum_{i=1}^{N} \sum_{j=1}^{t} e^{-\eta \lambda(\omega, p_{i,j}^{(i,j)})} w_{(i,j),t} + \sum_{i=1}^{N} \sum_{j=t+1}^{\infty} e^{-\eta \lambda(\omega, \gamma_{t+d,s})} w_{(i,j),t} =$$

$$\sum_{i=1}^{N} \sum_{j=1}^{t} e^{-\eta \lambda(\omega, p_{i,j}^{(i,j)})} w_{(i,j),t} + e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \left(1 - \sum_{i=1}^{N} \sum_{j=1}^{t} w_{(i,j),t}\right).$$

Subtracting the identical terms from both side of this inequality, we obtain the inequality

$$e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \geq \sum_{i=1}^{N} \sum_{j=1}^{t} e^{-\eta \lambda(\omega, p_{i,j}^{(i,j)})} \tilde{w}_{(i,j),t},$$

(9)

where $\tilde{w}_{(i,j),t} = \frac{w_{(i,j),t}}{\sum_{i=1}^{N} \sum_{j=1}^{t} w_{(i,j),t}}$. Using mixability of our loss function, for each $1 \leq s \leq d$ define $\gamma_{t+d,s} = \Sigma(g_{t+d,s})$, where

$$g_{t+d,s}(\omega) = -\frac{1}{\eta} \ln \sum_{i=1}^{N} \sum_{j=1}^{t} e^{-\eta \lambda(\omega, p_{i,j}^{(i,j)})} \tilde{w}_{(i,j),t},$$

(10)
Algorithm 1

Define $w_{(i,j),s} = w_{(i,j),s}^m = \frac{1}{N} \nu(j)$ for $i = 1, \ldots, N$, $j = 1, 2, \ldots$ and $0 \leq s \leq d$.  

FOR $t = 1, \ldots, T$

IF $t \leq d$ THEN put $l_{t,j}^i = h_t = 0$ for all $i$ and $j$. ELSE

1. Observe the outcomes $\omega_{t-d+1}, \ldots, \omega_t$ and the algorithm predictions $\gamma_t = (\gamma_{t,1}, \ldots, \gamma_{t,d})$ issued at the time moment $t - d$.

2. Compute the algorithm loss $h_t = \sum_{s=1}^d \lambda(\omega_{t-d+s}, \gamma_{t,s})$.

3. Compute the losses $l_{t,s}^{(i,j)} = \sum_{s=1}^d l_{t,s}^{(i,j)}$ of all auxiliary experts $(i,j)$, where $1 \leq i \leq N$, $1 \leq j \leq t - d$ and $l_{t,s}^{(i,j)} = \lambda(\omega_{t-d+s}, p_{j,t-d-j+s}^{(i,j)})$ are real losses of experts $(i,j)$ for $1 \leq s \leq d$ and we artificially put $l_{t,s}^{(i,j)} = \lambda(\omega_{t-d+s}, \gamma_{t,s})$ for $j > t - d$, where the expert $(i,j)$ yet did not present predictions.

ENDIF

4. Update weights $w_{(i,j),t} = w_{(i,j),t}^m e^{\frac{-\eta l_{t,s}^{(i,j)}}{\eta}}$ for $1 \leq i \leq N$, $1 \leq j < \infty$.

5. Receive the experts $1 \leq i \leq N$ predictions $p_{j}^{(i)}$ issued at the time moments $j \leq t$ and define vectors of the experts $(i,j)$ forecasts $p_{t+d}^{(i,j)} = (p_{t+d,1}^{(i,j)}, \ldots, p_{t+d,d}^{(i,j)})$, where for $1 \leq j \leq t$, $p_{t+d,s}^{(i,j)} = p_{j,t-d-j+s}^{(i,j)}$ is a forecast of the expert $i$ for the time moment $t + s$ issued at the time moment $j$.

6. Compute the algorithm long-term prediction $\gamma_{t+d} = (\gamma_{t+d,1}, \ldots, \gamma_{t+d,d})$, where $\gamma_{t+d,s} = \Sigma(g_{t,s})$ for $1 \leq s \leq d$ and $g_{t,s}$ is defined by (10).

7. Prepare the weights for the next step: $w_{t+1} = w_{t-d+1}^m$.

ENDFOR

Let $m_t = -\frac{1}{\eta} \ln \sum_{i=1}^N \sum_{j=1}^\infty w_{(i,j),t} e^{-\frac{\eta l_{t,s}^{(i,j)}}{\eta}}$ and $M_T = \sum_{t=1}^T m_t$ be the mixloss and cumulative mixloss of the auxiliary experts. Using equivalence of inequalities (8) and (9), we obtain $h_t \leq m_t$ and $H_T \leq M_T$.

By Lemma 1 we obtain for any $1 \leq i \leq N$ and $1 \leq j < \infty$,

$$M_T \leq \sum_{t=1}^T l_{t,s}^{(i,j)} + \frac{1}{\eta} d^2 \max_{1 \leq i \leq N, 1 \leq j \leq T} \ln \frac{1}{w_{(i,j),0}}.$$  

$$H_T \leq \sum_{t=1}^T l_{t,s}^{(i,j)} + \frac{1}{\eta} d^2 \ln(NT(T + 1)).$$  \hspace{1cm} (11)

Evidently, $\ln(NT(T + 1)) \leq \ln N + 2 \ln T + \frac{1}{T}$. Since $l_{t,s}^{(i,j)} = h_t$ for $j > t - d$, we subtract the identical terms from both sides of the inequality (11) and obtain the following theorem.

**Theorem 2** Let an $\eta$-mixable loss function be given. For any $T$ and for any $1 \leq i \leq N$ and $1 \leq j \leq T$,

$$\sum_{t=j+d}^T h_t \leq \sum_{t=j+d}^T L_{[t-d+1,d]}^{i,j} + \frac{1}{\eta} d^2 \left( \ln N + 2 \ln T + \frac{1}{T} \right),$$  \hspace{1cm} (12)

1. It is easy to verify that Lemma 1 is valid for an infinite number of experts.
where $L_{[t-d+1,t]}^{i,j} = \sum_{s=1}^{d} \lambda(\omega_{t-d+s}, p_{j,t-d-j+s}^{i})$ is the loss suffered by the expert $i$ (starting at time $j$) on time interval $[t-d+1, t]$.

Theorem 2 compares the total loss suffered by the aggregating algorithm on time moments $j + d \leq t \leq T$ and the total loss suffered by an expert $i$ (which issued its forecasts at a time moment $j \leq t - d$) on the same time interval.

### 3.2. Aggregating sliding experts

In this section consider a more realistic scenario. At each time moment, each expert presents a vector valued forecast of dimension $d$. At any time moment we mix all overlapping forecasts issued by the experts in the past which are available up to the time moment $t$.

We represent any expert $i$ by $d$ experts $(i,j)$, where $0 \leq j \leq d - 1$. At time $t$, each expert $(i,j)$ presents a truncated vector of $d-j$ forecasts $p_{t+d-j}^{i(j)} = (p_{t+d-j+1}^{i}, \ldots, p_{t+d-j+d}^{i})$, where for $j + 1 \leq s \leq d$, $p_{t+d-j+s}^{i}$ is a forecast of the expert $i$ for time moment $t - j + s$ issued at time moment $t - j$. This means that we consider the last $d-j$ forecasts of the expert $i$ starting $j$ time moments back of $t$.

In particular, for $j = 0$ we have a vector $p_{t+d}^{i(0)} = (p_{t+d,1}^{i}, \ldots, p_{t+d,d}^{i})$ of $d$ forecasts of the expert $i$ issued at time $t$ for $d$ steps ahead.

As in Section 3.1 we will use a fixed point method by Chernov and Vovk (2009) to compute the aggregating algorithm forecast using incomplete experts forecasts.

Assume that a vector $\gamma_{t+d}$ of forecasts of the aggregating algorithm for step $t + d$ is known. To compute the correct weights for the experts $(i,j)$, we virtually extend these truncated vectors of their forecasts by the forecasts of the aggregating algorithm. Define the extended vector valued forecast of any expert $(i,j)$ as $p_{t+d}^{i(j)} = (p_{t+d,1}^{i(j)}, \ldots, p_{t+d,d}^{i(j)})$, where the first $d-j$ components are $p_{t+d-j+1}^{i(j)}, \ldots, p_{t+d-j+d}^{i(j)}$ and the rest $j$ components are $\gamma_{t+d,d-j+1}, \ldots, \gamma_{t+d,d}$.

The corresponding loss of the expert $(i,j)$ at step $t + d$ is $\sum_{s=1}^{d} \lambda(\omega_{t+d}, p_{t+d,s}^{i(j)})$.

By the aggregating algorithm rule we define a prediction $\gamma_{t+d} = (\gamma_{t+d,1}, \ldots, \gamma_{t+d,d})$ with each component $\gamma_{t+d,s}$ satisfying

$$ e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \geq \sum_{i=1}^{N} \sum_{j=1}^{d} e^{-\eta \lambda(\omega, p_{t+d,s}^{i,j})} w_{(i,j),t} $$

for all $\omega$. Rewrite this inequality in more detail

$$ e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \geq \sum_{i=1}^{N} \sum_{j=0}^{d-s} e^{-\eta \lambda(\omega, p_{t+d,s}^{i,j})} w_{(i,j),t} + \sum_{i=1}^{N} \sum_{j=d-s+1}^{d} e^{-\eta \lambda(\omega, \gamma_{t+d,s})} w_{(i,j),t} =$$

$$ = \sum_{i=1}^{N} \sum_{j=0}^{d-s} e^{-\eta \lambda(\omega, p_{t+d,s}^{i,j})} w_{(i,j),t} + e^{-\eta \lambda(\omega, \gamma_{t+d,s})} \left( 1 - \sum_{i=1}^{N} \sum_{j=0}^{d-s} w_{(i,j),t} \right). $$

2. For $j = 0$ these components are absent.
Subtracting identical terms from both sides of this inequality, we obtain the inequality

\[ e^{-\eta l_t(\omega; \gamma_{t+d},s)} \geq \sum_{i=1}^{N} \sum_{j=0}^{d-s} e^{-\eta l_t(\omega; p_{t+d}\gamma_{t+d},s)} \tilde{w}_{(i,j),t,s}, \tag{14} \]

where

\[ \tilde{w}_{(i,j),t,s} = \frac{w_{(i,j),t}}{\sum_{i=1}^{N} \sum_{j=0}^{d-s} w_{(i,j),t}} \tag{15} \]

for \( 1 \leq s \leq d \). By mixability of the loss function the forecast \( \gamma_{t+d,s} \) can be defined efficiently using the superprediction function: for each \( 1 \leq s \leq d \) define \( \gamma_{t+d,s} = \Sigma(g_{t+d,s}) \), where

\[ g_{t+d,s}(\omega) = -\frac{1}{\eta} \ln \sum_{i=1}^{N} \sum_{j=0}^{d-s} e^{-\eta l_t(\omega; p_{t+d-j}\gamma_{t+d+j},s)} \tilde{w}_{(i,j),t}. \tag{16} \]

Note that if \( \gamma_{t+d,s} \) satisfies (14) then it also satisfies inequality (13) for \( 1 \leq s \leq d \).

**Algorithm 2**

1. Put \( w_{(i,j),s} = w_{(i,j),s}^m = \frac{1}{dN} \) for \( i = 1, \ldots, N, j = 0, \ldots, d - 1 \), and \( 0 \leq s \leq d \).
2. **FOR** \( t = 1, \ldots, T \)
   - **IF** \( t \leq d \) **THEN** put \( l_{t}^{(i,j)} = h_t = 0 \) for all \( i \) and \( j \). **ELSE**
     1. Observe the outcomes \( \omega_{t-d+1}, \ldots, \omega_t \) and the algorithm predictions \( \gamma_t = (\gamma_{t,1}, \ldots, \gamma_{t,d}) \) issued at the time moment \( t - d \).
     2. Compute the algorithm loss \( h_t = \sum_{s=1}^{d} \lambda(\omega_{t-d+s}, \gamma_{t,s}) \).
     3. Compute the losses \( \tilde{l}_{t}^{(i,j)} = \sum_{s=1}^{d} \tilde{l}_{t}^{(i,j),s} \) of all experts \( (i, j) \), where \( 1 \leq i \leq N, 0 \leq j \leq d - 1 \) and \( \tilde{l}_{t}^{(i,j),s} = \lambda(\omega_{t-d+s}, p_{t-d+j}\gamma_{t+d+j},s) \) are real losses of experts \( (i, j) \) for \( 1 \leq s \leq d - j \); here we artificially put \( \tilde{l}_{t}^{(i,j),s} = \lambda(\omega_{t-d+s}, \gamma_{t,s}) \) for time moments \( d - j + 1 \leq s \leq d \), where the expert \( (i, j) \) did not present predictions.
   - **ENDIF**
3. Define \( w_{(i,j),t}^m = \frac{w_{(i,j),t}^m e^{-\eta \tilde{l}_{t}^{(i,j)}}}{\sum_{i'=1}^{N} \sum_{j'=0}^{d-s} w_{(i',j'),t} e^{-\eta \tilde{l}_{t}^{(i',j')}}} \) for \( 1 \leq i \leq N, 0 \leq j \leq d - 1 \).
4. For each \( 1 \leq i \leq N \), receive a vector of the expert \( i \) predictions \( p_{t+d}^{(i)} = (p_{t+d,1}^{(i)}, \ldots, p_{t+d,d}^{(i)}) \) issued at time \( t \) and recall predictions issued by the expert \( i \) in the past and overlapping with the time interval \( [t + 1, t + d] \). For any \( 0 \leq j \leq d - 1 \), collect these predictions in a vector \( p_{t-d+j,d+j,s}^{(i)} \), \( 1 \leq s \leq d - j \). Using these predictions, compute a forecast of the aggregating algorithm \( \gamma_{t+d} = (\gamma_{t+d,1}, \ldots, \gamma_{t+d,d}) \) where \( \gamma_{t+d,s} = \Sigma(g_{t+d,s}) \) and \( gt+d,s \) is defined by (16) for \( 1 \leq s \leq d \).
5. Define the weights for the next step by \( w_{t+1} = w_{t}^{m-d} \).
6. **ENDFOR**

Let \( m_t = -\frac{1}{\eta} \ln \sum_{i=1}^{N} \sum_{j=0}^{d-1} w_{(i,j),t} e^{-\eta \tilde{l}_{t}^{(i,j)}} \) be the mixloss and \( M_T = \sum_{t=1}^{T} m_t \) be the cumulative mixloss of the auxiliary experts.

Using equivalence of inequalities (14) and (13), where the number of experts is \( dN \), we obtain that \( h_t \leq m_t \) and \( H_T \leq M_T \). Then, using the general scheme of Lemma 1, we obtain

\[ H_T \leq \min_{1 \leq i \leq N, 0 \leq j \leq d - 1} \sum_{t=1}^{T} \tilde{l}_{t}^{(i,j)} + \frac{1}{\eta} d^2 \ln(dN). \tag{17} \]
Since for any \( t, \ i^{(i,j)}_t = \sum_{s=1}^{d-j} \lambda(\omega_{t-d+s}, p_{t-j+d,j+s}^i) + \sum_{s=d-j+1}^{d} \lambda(\omega_{t-d+s}, \gamma_{t,s}) \), we can subtract the identical terms from both sides of the inequality (17).\(^3\)

Denote by \( L_{i}^{i,j} = \sum_{s=1}^{d-j} \lambda(\omega_{t-d+s}, p_{t-j+d,j+s}^i) \) the loss of an expert \( i \) starting \( j \) time moments back of \( t-d \) suffered on the time interval \([t-d+1, t-j]\), and by \( H_{i}^{1} = \sum_{s=1}^{d-j} \lambda(\omega_{t-d+s}, \gamma_{t,s}) \) the loss of the aggregating algorithm suffered on the same time interval. For any \( t \) and \( 0 \leq j \leq d-1 \), both these sums are taken over all time moments where the vector of the expert \( i \) predictions is valid and has nonempty intersection with the time interval \([t-d+1, t]\).

Therefore, we obtain the following theorem.

**Theorem 3** Let an \( \eta \)-mixable loss function be given. For any \( 1 \leq i \leq N, \ 0 \leq j \leq d-1, \) and for all \( T \),

\[
\sum_{t=1}^{T} H_{i}^{1} \leq \sum_{t=1}^{T} L_{i}^{i,j} + \frac{1}{\eta} d^2 \ln(dN).
\]  

(18)

### 3.3. Grid connection using the method of Mixing Past Posteriors

A simplest way to take into account the relationships between the learning processes on different parts of a time series is to use the method of Mixing Past Posteriors (MPP) by Bousquet and Warmuth (2002). In this section we consider one-point \( d \)-steps ahead forecasting.

We will use simple properties of relative entropy. For two vectors \( p \) and \( q \), we write \( p > q \), if the relationship \( \geq \) holds componentwise. We use \( 0 \) to denote the all-zero vector. The proofs for MPP rely on the following simple inequalities for the relative entropy. For all \( p, q, w \in \Gamma_n \) such that \( q, w > 0 \) it holds that \( D(p||q) \leq D(p||w) + \ln \left( \sum_{i=1}^{n} p_i \frac{w_i}{q_i} \right) \). If \( q \geq \mu w \) for some \( \mu > 0 \) then

\[
D(p||q) \leq D(p||w) + \ln \frac{1}{\mu}.
\]  

(19)

In particular, for \( p = w \) we have \( D(w||q) \leq \ln \frac{1}{\mu} \) for \( q \geq \mu w \). Indeed, from concavity of the logarithm \( D(p||q) - D(p||w) = \sum_{i=1}^{n} p_i \ln \frac{w_i}{q_i} \leq \ln \left( \sum_{i=1}^{n} p_i \frac{w_i}{q_i} \right) \). If \( q \geq \mu w \) then \( \sum_{i=1}^{n} p_i \frac{w_i}{q_i} \leq \sum_{i=1}^{n} p_i \frac{w_i}{\mu w_i} = \frac{1}{\mu} \). The following inequality will be used below. Let \( p \in \Gamma_n \) and \( q = \sum_{i=0}^{t} \beta_i w_i \), where \( w_i \in \Gamma_n, w_i > 0 \) for \( 0 \leq i \leq t, \beta = (\beta_0, \ldots, \beta_t) \in \Gamma_{t+1} \) and \( \beta > 0 \). Then

\[
D(p||q) \leq D(p||w_s) + \ln \frac{1}{\beta_s}
\]  

(20)

for any \( 0 \leq s \leq t \). In particular, for \( p = w_s, \ D \left( w_s || \sum_{i=0}^{t} \beta_i w_i \right) \leq \ln \frac{1}{\beta_s} \) for any \( s \).

By the method MPP an extra update rule is added to the loss update (1). By a mixing scheme at step \( t \) we mean a vector of weights \( \beta^{t+1} = (\beta_0^{t+1}, \ldots, \beta_t^{t+1}) \), where \( \beta_s \geq 0 \) for \( 0 \leq s \leq t \) and \( \sum_{s=0}^{t} \beta_s^{t+1} = 1 \).

\(^3\) For \( j = 0 \) the second sum is absent.
At any time moment $t$ we choose a mixing scheme $\beta^{t+1}$ and define $v^m_t = \sum_{s=0}^{t} \beta^{s+1}_t w^m_s$, where $v^m_t = (v_{1,t}, \ldots, v_{N,t})$ and $w^m_s = (w_{1,s}, \ldots, w_{N,s})$ is defined by (1). For the next step we define the weights as $w_{t+1} = v_{t-d+1}$. We also put initial weights $w_{i,s}^m = v_{i,s} = \frac{1}{N}$ for all $1 \leq i \leq N$ and $0 \leq s \leq d$ and start the process of weights updates (1) at the time moment $d + 1$. Note that $v^m_t = w^m_t$ for the mixing scheme $\beta^{t+1}$, where $\beta^{t+1}_t = 1$ and $\beta^{t+1}_s = 0$ for $0 \leq s \leq t - 1$.

We will use the following nontrivial mixing scheme for grid connection: let $\sigma \in (0, 1]$ and $1 > \alpha_1 > \alpha_2 > \cdots > 0$ be parameters; in what follows $\alpha_t = \alpha$ or $\alpha_t = \frac{1}{t}$ for all $t$. Define $\beta^{t+1}_t = 1 - \alpha_t$, $\beta^{t+1}_s = \frac{1}{2} \alpha_t (1 - \sigma)$ for $1 \leq s \leq d$, $\beta^{t+1}_0 = \alpha_t \sigma$, and $\beta^{t+1}_0 = 0$ for $1 \leq s < t - d$.

In this case the corresponding intermediate weights are

$$v^m_t = \alpha_t \left[ \frac{\sigma}{N} e + \frac{1}{d} (1 - \sigma) \sum_{s=1}^{d} w^m_{t-s} \right] + (1 - \alpha_t) w^m_t$$

for all $t$, where $e = (1, \ldots, 1)$.

Let a sequence of comparison vectors $q_t = (q_{1,t}, \ldots, q_{N,t})$ be given online for $t = 1, 2, \ldots$. Assume that the comparison vector $q_t$ changes $K$ times for $1 \leq t \leq T$. Let $d < t_1 < t_2 < \ldots < t_K \leq T$ be the subsequence of indices in the sequence of comparators $q_1, \ldots, q_T$, where shifting occurs: $q_{t_k} \neq q_{t_k-1}$ and $q_{t} = q_{t-1}$ for all other steps, where $t > 1$. Define also $t_0 = 1$ and $t_{K+1} = T + 1$. We apply Lemma 1 for our mixing scheme. Summing (2) on the time interval where $q_t = q_{t_k}$ for $t_k \leq t \leq t_k + 1 - 1$, we obtain

$$\sum_{t=t_i}^{t_i+1} \frac{1}{\eta} (D(q_{t_k} || w_t) - D(q_{t_k} || w^m_t)) \leq \frac{1}{\eta} \ln \frac{1}{1 - \alpha_t}$$

for all $i$.

We have used Lemma 1 in line (21). In transition from (21) to (22), we use the inequalities $w_t \geq (1 - \alpha_t) w^m_{t-d}$ and (19), where $\mu = 1 - \alpha_t$. In transition from (22) to (23), the corresponding positive and negative terms telescope. In transition from (23) to (24), only first $d$ (positive) terms of the previous equation remain. In transition from (24) to (25) the inequality $w_t \geq \frac{\alpha_t \sigma}{N} e$ was used, where $e = (1, \ldots, 1)$, then

$$D(q_t || w_t) = \sum_{i=1}^{N} q_{t,i} \ln \frac{q_{t,i}}{w_{i,t}} \leq \sum_{i=1}^{N} q_{t,i} \ln q_{t,i} - \sum_{i=1}^{N} q_{t,i} \ln \frac{\alpha_t \sigma}{N} \leq \ln N + \frac{1}{\alpha_t} + \ln \frac{1}{\sigma}. \quad (26)$$
Summing these inequalities for $0 \leq k \leq K$ and bounding $\ln \frac{1}{\alpha_t} \leq \max_{t \in [1,T]} \ln \frac{1}{\alpha_t}$, we obtain the resulting inequality:

$$M_T - \sum_{t=1}^{T} (q_t \cdot l_t) \leq \frac{1}{\eta} \left( d(K + 1) \left( \ln N + \ln \frac{1}{\sigma} + \max_{t \in [1,T]} \ln \frac{1}{\alpha_t} \right) + \sum_{t=1}^{T} \ln \frac{1}{1-\alpha_t} \right).$$  \hspace{1cm} (27)

If we take $\sigma = 1$, then we separate the game into $d$ separate games on grids $G_i$.

A classical practical approach to take $\alpha_t = \alpha$ for some $\alpha \in (0,1)$ gives some regret dependence on time of the form $O(T)$. However, using non-constant $\alpha_t = \frac{1}{t}$ allows the left part of (27) to become equal to $O(\ln T)$.

**Theorem 4** Let a loss function be $\eta$-mixable and $\alpha_t = \frac{1}{t}$ for all $t$, $\sigma \in (0,1]$. Then, for any $T$, for any sequence of comparison vectors $q_t \in \Delta_N$ with no more than $K$ switches,

$$H_T \leq M_T \leq \sum_{t=1}^{T} (q_t \cdot l_t) + \frac{1}{\eta} \left( d(K + 1) \left( \ln N + \ln \frac{1}{\sigma} + \ln T \right) + \ln T \right)$$

that provides an $O(d \cdot K \cdot (\ln T + \ln N))$ bound for the regret $R_T$ with respect to any sequence of experts with no more that $K$ switches.

4. Acknowledgement

The first author was supported by the Ministry of Education and Science of Russian Federation, grant No. 14.606.21.0004, grant code: RFMEFI60617X0004. The second and the third authors were partially supported by the RFBR grants 16-01-00576 A and 16-29-09649 oFinm.

Vladimir V’yugin thanks Vladimir Vovk and Yuri Kalnishkan for useful discussions.

References

D. Adamskiy, T. Bellotti, R. Dzhamtyrova, Y. Kalnishkan. Aggregating Algorithm for Prediction of Packs. arXiv:1710.08114 [cs.LG], 2017.

Oren Anava, Elad Hazan, Shie Mannor. Online Learning for Adversaries with Memory: Price of Past Mistakes. NIPS 2015, 784–792, 2015.

O. Bousquet, M. Warmuth. Tracking a small set of experts by mixing past posteriors. Journal of Machine Learning Research. 3:363–396, 2002.

A. Chernov, V. Vovk. Prediction with expert evaluators’ advice. Lecture Notes in Artificial Intelligence. 8–22, 2009.

Y. Freund, R.E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 55:119–139, 1997.

N. Cesa-Bianchi, G. Lugosi. Prediction, Learning, and Games. Cambridge University Press. 2006.
Appendix A. Mixable loss functions

We study the learning with mixable loss functions $\lambda(\omega, \gamma)$, where $\omega$ is an element of some set of outcomes $\Omega$, and $\gamma$ is an element of some set of forecasts $\Gamma$. A set of experts $E$ is given. The experts $i \in E$ present the forecasts $\xi^i \in \Gamma$. Although all considerations are valid for an infinite set of experts elsewhere except of Section 3.1 we use finite sets of $N$ experts.

The Vovk’s aggregating algorithm (Vovk 1990, Vovk 1998) is the base algorithm in our study. In this case the main tool is a superprediction function

$$g(\omega) = -\frac{1}{\eta} \ln \int e^{-\eta \lambda(\omega, \xi^i)} dp(i),$$

where $p$ is a probability distribution on the set of all experts and $\xi^i$ are the experts predictions and $\omega \in \Omega$. 

13
A loss function $\lambda(\omega, \gamma)$ is $\eta$-mixable, where $\eta > 0$, if for any probability distribution $p$ on the set of experts and for any set of experts predictions $\xi^i$ there exists $\gamma$ such that $\lambda(\omega, \gamma) \leq g(\omega)$ for all $\omega$. We write $\gamma = \Sigma(g)$, where $\Sigma$ is called a substitution function.

A loss function $\lambda(\omega, \gamma)$ is $\eta$-exponential concave if for any $\omega$ the function $e^{-\eta \lambda(\omega, \gamma)}$ is concave w.r.t. $\gamma$. By definition any $\eta$-exponential concave function is $\eta$-mixable with the substitution function $\gamma = \sum_{i=1}^{N} \xi^i p(i)$, where $\xi^i$ are the experts predictions and $p(i)$ are their weights.

The square loss function $\lambda(\omega, \lambda) = (\omega - \gamma)^2$ is $\eta$-mixable for any $\eta$ such that $0 < \eta \leq \frac{1}{2L}$, where $\omega$ and $\lambda$ are real numbers from the interval $[-L, L]$, see (Vovk 1990, Vovk 1998). For the square loss function the corresponding prediction can be defined as

$$\gamma_t = \Sigma(g_t) = \frac{1}{4L} \left( g_t(-L) - g_t(L) \right) = \frac{1}{4\eta L} \ln \frac{\sum_{i=1}^{N} w_i e^{-\eta \lambda(L, \xi^i)}}{\sum_{i=1}^{N} w_i e^{-\eta \lambda(-L, \xi^i)}}.$$

The square loss function is also $\eta$-exponential concave for $0 < \eta \leq \frac{1}{8L^2}$.

Let $\lambda(\omega, \gamma)$, where $\eta > 0$ be an $\eta$-mixable loss function and $p = (p_1, \ldots, p_N)$ be a probability vector. Let any expert $i \in \{1, \ldots, N\}$ presents a sequence (vector) of forecasts $\xi^i = (\xi^i_1, \ldots, \xi^i_d)$. Following Adamskiy et al. (2017), to compute a vector-valued forecast we apply the aggregation rule to each coordinate separately: since the loss function is $\eta$-mixable, for each $1 \leq s \leq d$ a prediction $\gamma_s$ exists such that

$$e^{-\eta \lambda(\omega, \gamma_s)} \geq \sum_{i=1}^{N} e^{-\eta \lambda(\omega, \xi^i_s)} p_i$$

for all $\omega$. The prediction can be computed as $\gamma_s = \Sigma(g_s)$, where $\Sigma$ is a substitution function and $g(\omega) = -\frac{1}{\eta} \sum_{i=1}^{N} e^{-\eta \lambda(\omega, \xi^i_s)} p_i$.

Multiplying the inequalities (28) we obtain

$$e^{-\eta \sum_{s=1}^{d} \lambda(\omega, \gamma_s)} = \prod_{s=1}^{d} e^{-\eta \lambda(\omega, \gamma_s)} \geq \prod_{s=1}^{d} \sum_{i=1}^{N} e^{-\eta \lambda(\omega, \xi^i_s)} p_i.$$  \hspace{1cm} (29)

The generalized H"{o}lder inequality says that $\|f_1 f_2 \cdots f_d\|_r \leq \|f_1\|_{q_1} \|f_2\|_{q_2} \cdots \|f_d\|_{q_d}$, where $\frac{1}{q_1} + \cdots + \frac{1}{q_d} = \frac{1}{r}$, $q_s \in (0, +\infty)$ and $f_s \in L^{q_s}$ for $1 \leq s \leq d$. Let $q_s = 1$ for all $1 \leq s \leq d$, then $r = 1/d$. Consider a norm $\|f\|_1 = E_p(f)$. Let $f_s^i(i) = e^{-\eta \lambda(\omega, \xi^i_s)}$. Then $\|f_s\|_1 \leq e^{-\eta \lambda(\omega, \xi^i_s)}$.

The corresponding H"{o}lder inequality is

$$\left( \sum_{i=1}^{N} p_i \left( \prod_{s=1}^{d} e^{-\eta \lambda(\omega, \xi^i_s)} \right)^{\frac{1}{d}} \right)^{d} \leq \prod_{s=1}^{d} \sum_{i=1}^{N} p_i e^{-\eta \lambda(\omega, \xi^i_s)}.$$  \hspace{1cm} (30)

4. For a general (not necessary mixable) loss function the inequality (28) should be replaced by $e^{-\eta \lambda(\omega, \gamma_s)/c} \geq \sum_{i=1}^{N} e^{-\eta \lambda(\omega, \xi^i_s)} p_i$ for some constant $c > 0$. In this case the results of subsections 3.1 and 3.2 can be reformulated appropriately.
Rewrite (29) and (30) in the form

\[ e^{-\frac{\eta}{d} \sum_{s=1}^{d} \lambda(\omega, \gamma_s)} \geq \sum_{i=1}^{N} p_i e^{-\frac{\eta}{d} \sum_{s=1}^{d} \lambda(\omega, \xi^i_s)}. \]  

(31)

Here for any outcome \( \omega \), \( l^i = \sum_{s=1}^{d} \lambda(\omega, \xi^i_s) \) is the loss of the expert \( i \) presenting a sequence of predictions \( \xi^i = (\xi^i_1, \ldots, \xi^i_d) \) and \( h = \sum_{s=1}^{d} \lambda(\omega, \gamma_s) \) is the loss of the algorithm presenting predictions \( \gamma = (\gamma_1, \ldots, \gamma_d) \). We have shown that the forecasts \( \gamma = (\gamma_1, \ldots, \gamma_d) \) satisfying (28) also satisfy (31), i.e. the corresponding vector-valued game is \( \frac{\eta}{d} \)-mixable.

Appendix B. Experiments

In this section we run algorithm 2 on the artificial data and show when the sliding experts approach may be useful.

We create artificial time series for \( T = 200 \) as a realization of a random walk \( \omega_t = \sum_{\tau=1}^{t} \epsilon_\tau \), where \( \epsilon_\tau \sim U[-\frac{1}{6}, \frac{1}{6}] \) are independent random variables. The sampled time series is shown in Figure 1.

Figure 1: Sampled artificial time series \( \omega_t = \sum_{\tau=1}^{t} \epsilon_\tau \), where \( \epsilon_\tau \sim U[-\frac{1}{6}, \frac{1}{6}] \) are i.i.d.

We use three artificial experts. Each expert at each time step \( t \) predicts \( d = 10 \) next time series values. The predictions are realizations of i.i.d. random variables, distributed as \( U[\omega_{t+i}, \omega_{t+i} + \frac{1}{10}] \).

Expert \( i = 1 \) differs from the rest of experts: on time steps \( t \in G_0 = \{ t \mid t \equiv 0 \ (\text{mod} \ d) \} \) its predictions are exact, however, on other time steps it works in the same way as other experts. This means that the expert \( i = 1 \) on every \( d \)-th step provides very accurate (exact) prediction and its prediction is useful not only on the grid \( G_0 \) but on other grids too (as predictions of a sliding expert).

We apply Algorithm 2 for sliding experts (thus we obtain \( d \cdot N = 30 \) experts) and plot the cumulative losses of the base \( N = 3 \) experts and the loss of the aggregating algorithm in Figure 2. We also apply the version of Algorithm 2 for base experts without adding sliding copies. In this experiment trivial average prediction also serves as a comparator (note that average prediction performs approximately the same as the algorithm on the base experts).

The incredible performance of Algorithm 2 is explained by the availability of sliding copies of the expert \( i = 1 \), providing accurate predictions. More precisely, on the grid
Figure 2: The cumulative losses $L_i^t = \sum_{\tau=1}^{t} l_i^\tau$ of the base experts and the cumulative loss $H_i^t = \sum_{\tau=1}^{t} h_i^\tau$ of algorithm 2 (with and without sliding experts).

$G_j$ on every step $t \in G_j$ the sliding expert $(1,j)$ predicts exactly next $d - j$ time series elements. Thus, on every grid $G_j$ the weight of the expert $(1,j)$ increases. This leads to better prediction quality. The evolution of weights of some experts $(1,j)$ is shown in Figure 3. The weights of experts $(1,j)$ increase over $t$ on the corresponding grids.

Figure 3: Evolution of weights of sliding experts $(1,1)$ and $(1,5)$. 