Pattern Recognition 121 (2022) 108193

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Online aggregation of probability forecasts with confidence

Vladimir V'yugin[a,*], Vladimir Trunov[b]

[a] Institute for Information Transmission Problems, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19, Moscow GSP-4, 127994, Russia
[b] Institute for Information Transmission Problems, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19, Moscow GSP-4, 127994, Russia

ARTICLE INFO

Article history:
Received 15 September 2020
Revised 22 June 2021
Accepted 4 July 2021
Available online 29 July 2021

Keywords:
On-line learning
Prediction with expert advice
Aggregating algorithm
Probabilistic prediction
Continuous ranked probability score (CRPS)
Smooth confidence levels for experts

ABSTRACT

The paper presents numerical experiments and some theoretical developments in prediction with expert advice (PEA). One experiment deals with predicting electricity consumption depending on temperature and uses real data. As the pattern of dependence can change with season and time of the day, the domain naturally admits PEA formulation with experts having different "areas of expertise". We consider the case where several competing methods produce online predictions in the form of probability distribution functions. The dissimilarity between a probability forecast and an outcome is measured by a loss function (scoring rule). A popular example of scoring rule for continuous outcomes is Continuous Ranked Probability Score (CRPS). In this paper the problem of combining probabilistic forecasts is considered in the PEA framework. We show that CRPS is a mixable loss function and then the time-independent upper bound for the regret of the Vovk aggregating algorithm using CRPS as a loss function can be obtained. Also, we incorporate a "smooth" version of the method of specialized experts in this scheme which allows us to combine the probabilistic predictions of the specialized experts with overlapping domains of their competence.

1. Introduction

Probabilistic forecasts in the form of probability distributions over future events have become popular in several fields, including meteorology, hydrology, economics, demography. Probabilistic predictions are used in the theory of conformal predictions, where a predictive distribution that is valid under a nonparametric assumption can be assigned to any forecasting algorithm (see Vovk et al. [28]).

The dissimilarity between a probability forecast and an outcome is measured by a loss function (scoring rule). A popular example of scoring rule for continuous outcomes is Continuous Ranked Probability Score (CRPS).

$$\mathrm{CRPS}(F,y)=\int (F(u)-H(u-y))^2 du,$$

where $F(u)$ is a probability distribution function, y is an outcome – a real number, and $H(x)$ is the Heaviside function: $H(x)=0$ for $x<0$ and $H(x)=1$ for $x\geq 0$ (Epstein [11], Gneiting and Raftery [15]).

The paper presents theoretical developments in prediction with expert advice (PEA) and some numerical experiments. One experiment deals with predicting electricity consumption depending on temperature and uses real data. As the pattern of dependence can change with season and time of the day, the domain naturally admits PEA formulation with experts having a different "areas of expertise".

We consider the case where several competing methods produce online predictions in the form of probability distribution functions. These predictions can lead to large or small losses. Our task is to combine these forecasts into one optimal forecast, which will lead to a relatively small possible loss in the framework of the available past information.

We solve this problem in the PEA framework. We consider the game-theoretic on-line learning model in which a learner (aggregating) algorithm has to combine predictions from a set of N experts (see e.g. Littlestone and Warmuth [19], Freund and Schapire [12], Vovk [25], Kivinen and Warmuth [17], Vovk [26], Cesa-Bianchi and Lugosi [9] among others).

In contrast to the standard PEA approach, we consider the case where each expert presents probability distribution functions rather than a point prediction. The learner presents his forecast also in the form of probability distribution function computed using probabilistic predictions presented by the experts.

In online setting, at each time step t each expert issues a probability distribution as a forecast. The aggregating algorithm combines these forecasts into one aggregated forecast, which is a probability distribution function. The effectiveness of the aggregating algorithm on any time interval $[1,T]$ is measured by the regret which is the difference between the accumulated loss of the ag-

* Corresponding author.
E-mail addresses: vuygin@iitp.ru (V. V'yugin), trunov@iitp.ru (V. Trunov).

https://doi.org/10.1016/j.patcog.2021.108193
0031-3203/

gregating algorithm and the accumulated loss of the best expert suffered on first T steps.

There are many papers on probabilistic predictions and on CRPS scoring rule (some of them are Brier [4], Bröcker and Smith [5], Bröcker and Smith [6], Bröcker [7], Epstein [11], Raftery et al. [22]). In some cases, experts use for their predictions probability distributions functions (data models) which are defined explicitly in an analytic form. In this paper, we propose the rules for aggregation of such probability distributions functions. We present the exact formulas for direct calculation of the aggregated probability distribution function given probability distribution functions are presented by the experts.

We obtain a tight upper bound of the regret for a special case when the outcomes and the probability distributions are supported on a finite interval $[a, b]$ of real line. In Section 4 we prove that the CRPS function is mixable and then all machinery of the aggregating algorithm (AA) by Vovk [26] and of the exponentially weighted average forecaster (WA) (see Cesa-Bianchi and Lugosi [9]) can be applied. We present a method for computing online the aggregated probability distribution function given the probability distribution functions are presented by the experts and prove a time-independent bound for the regret of the proposed algorithm.

The application we will consider below in Section 5 (which is the sequential forecasting of probability distribution function of electricity consumption) will take place in a variant of the basic problem of prediction with expert advice called prediction with specialized (or sleeping) experts. At each round, only some of the experts output a prediction while the other ones are inactive. Each expert is expected to provide accurate forecasts mostly under given external conditions that can be known beforehand. For instance, in the case of the prediction of electricity consumption, experts can be specialized to a season, temperature forecast, and time of the day.

Each expert is trained on its specific domain. Moving from one domain to another, an expert which was tuned to the previous domain gradually loses his predictive ability. To take this into account, we define a smooth extension of the domain of any expert. Thus, each expert competes with other experts working at overlapping intervals. The second contribution of this paper is that we have incorporated a smooth generalization of the method of specialized experts (Sections 3 and 4.1) which allows us to combine the probabilistic predictions into the aggregating algorithm (AA) of the specialized experts with overlapping domains of theirs competence.

We demonstrate the effectiveness of the proposed methods in Section 5, where the results of numerical experiments with synthetic and real data are presented.

2. Preliminaries

In this section we present the main definitions and the auxiliary results of the theory of prediction with expert advice, namely, learning with mixable loss functions.

2.1. Online learning

Let Ω be a set of outcomes and Γ be a set of forecasts (decision space).[1] We consider the learning with a loss function $\lambda(f, y)$, where $f \in \Gamma$ and $y \in \Omega$. Let also, a set E of experts be given. For simplicity, we assume that $E = \{1, \dots, N\}$.

In PEA approach the learning process is represented as a game. The experts and the learner observe past real outcomes generated online by some adversarial mechanism (called nature) and present their forecasts. After that, a current outcome is revealed by the nature.

In more detail, at any round $t = 1, 2, \dots$, each expert $i \in E$ presents a forecast $f_{i,t} \in \Gamma$, then the learner presents its forecast $f_t \in \Gamma$, and after that, an outcome $y_t \in \Omega$ is revealed. Each expert i suffers the loss $\lambda(f_{i,t}, y_t)$, and the learner suffers the loss $\lambda(f_t, y_t)$. The game of prediction with expert advice is presented by Protocol 1 below.

Protocol 1

FOR $t = 1, \dots, T$

1. Receive the experts' predictions $f_{i,t}$, where $1 \le i \le N$.
2. Present the learner's forecast f_t.
3. Observe the true outcome y_t and compute the losses $\lambda(f_{i,t}, y_t)$ of the experts and the loss $\lambda(f_t, y_t)$ of the learner.

ENDFOR

Let $H_T = \sum_{t=1}^{T} \lambda(f_t, y_t)$ be the accumulated loss of the learner and $L_T^i = \sum_{t=1}^{T} \lambda(f_{i,t}, y_t)$ be the accumulated loss of an expert i. The difference $R_T^i = H_T - L_T^i$ is called regret with respect to an expert i, and $R_T = H_T - \min_i L_T^i$ is the regret with respect to the best expert. The goal of the learner is to minimize regret.

2.2. Aggregating algorithm (AA)

The Vovk Aggregating algorithm (Vovk [25] and Vovk [26]) is the base algorithm for computing the learner predictions. This algorithm starting from the initial weights $w_{i,1}$ (usually $w_{i,1} = \frac{1}{N}$ for all i) assign weights $w_{i,t}$ for the experts $i \in E$ using the weights update rule:

$$w_{i,t+1} = w_{i,t} e^{-\eta \lambda(f_{i,t}, y_t)} \text{ for } t = 1, 2, \dots, \tag{1}$$

where $\eta > 0$ is a learning rate. The normalized weights are defined

$$w_{i,t}^* = \frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}}. \tag{2}$$

The main tool of AA is a superprediction function

$$g_t(y) = -\frac{1}{\eta} \ln \sum_{i=1}^{N} e^{-\eta \lambda(f_{i,t}, y)} w_{i,t}^*. \tag{3}$$

We consider probability distributions $\mathbf{q} = (q_1, \dots, q_N)$ on the set E of the experts: $\sum_{i=1}^{N} q_i = 1$ and $q_i \ge 0$ for all i. By Vovk [26] a loss function is called η-mixable if for any probability distribution $\mathbf{q}$ on the set E of experts and for any predictions $\mathbf{f} = (f_1, \dots, f_N)$ of the experts there exists a forecast f such that

$$\lambda(f, y) \le g(y) \text{ for all } y, \tag{4}$$

where

$$g(y) = -\frac{1}{\eta} \ln \sum_{i=1}^{N} e^{-\eta \lambda(f_i, y)} q_i. \tag{5}$$

We fix some rule for calculating a forecast f and write

$$f = \text{Subst}(\mathbf{f}, \mathbf{q}). \tag{6}$$

The function Subst is called the substitution function.

As follows from (4) and (5), if a loss function $\lambda(f, y)$ is η-mixable, then the loss function $c\lambda(f, y)$ is $\frac{\eta}{c}$-mixable for any $c > 0$.

[1] In general, these sets can be of arbitrary nature. We will specify them when necessary.

The upper bound $H_T \le \sum_{t=1}^{T} g_t(y_t) \le L_T^i + \frac{\ln N}{\eta}$ for any expert i is obtained in A.1. Therefore, there is a strategy for the learner that guarantees the time-independent upper bound for the regret $R_T \le \frac{\ln N}{\eta}$ for all T regardless of which sequence of outcomes is observed.

2.3. Exponentially concave loss functions

Assume that all forecasts form a linear space. In this case, the mixability is a generalization of the notion of exponentially concavity. A loss function $\lambda(f,y)$ is called η-exponentially concave if for each y the function $\exp(-\eta\lambda(f,y))$ is concave in f (see Kivinen and Warmuth [17], Cesa-Bianchi and Lugosi [9]). For exponentially concave loss function the inequality (4) holds for all y by definition if the forecast of the learner is computed using the weighted average (WA) of the experts predictions:

$$f = \sum_{i=1}^{N} q_i f_i, \tag{7}$$

where $\mathbf{q} = (q_1, \dots, q_N)$ is a probability distribution on the set of experts, and $f_1, \dots, f_N$ are theirs forecasts.

For exponentially concave loss function and the game defined by Protocol 1, where the learner's forecast is computed by (7), we also have the time-independent bound (A.1) for the regret.

2.4. Square loss function

The important special case is $\Omega = \{0,1\}$ and $\Gamma = [0,1]$. The square loss function $\lambda(\gamma,\omega) = (\gamma-\omega)^2$ is η-mixable loss function for any $0 < \eta \le 2$, where $\gamma \in [0,1]$ and $\omega \in \{0,1\}$.[2] In this case, at any step t, the corresponding forecast f_t (in Protocol 1) can be defined as

$$f_t = \mathrm{Subst}(\mathbf{f}_t, \mathbf{w}_t^*) = \frac{1}{2} - \frac{1}{2\eta} \ln \frac{\sum_{i=1}^{N} w_{i,t}^* e^{-\eta\lambda(f_{i,t},0)}}{\sum_{i=1}^{N} w_{i,t}^* e^{-\eta\lambda(f_{i,t},1)}}, \tag{8}$$

where $\mathbf{f}_t = (f_{1,t}, \dots, f_{N,t})$ is the vector of the experts' forecasts and $\mathbf{w}_t^* = (w_{1,t}^*, \dots, w_{N,t}^*)$ is the vector of theirs normalized weights defined by (1) and (2). We refer the reader for details to Vovk [25], Vovk [26], and Vovk [27].

The square loss function $\lambda(f,\omega) = (f-\omega)^2$ is η-exponential concave for any $0 < \eta \le \frac{1}{2}$ (see Cesa-Bianchi and Lugosi [9]).

Note that the larger the learning rate, the faster the weights update rule (1) adapts to the changing predictive abilities of the experts.

3. AA For experts with confidence

In the experiments, which will be presented below in Section 5.2, the specialized experts will be used, where each expert is associated with specific type of domain (time interval).

We define a smooth extension of the domain of any expert. The scope of each expert will be determined by its confidence values. Inside the area for which the expert was tuned, its confidence values are equal to 1, and outside this area they decrease with time linearly from 1 to 0.

The method of specialized experts was first proposed by Freund et al. [13] and further developed by Chernov and Vovk [8], Devaine et al. [10], Gaillard et al. [14], Kalnishkan et al. [16]. With this approach, at each step t, a set of specialized experts $E_t \subseteq \{1, \dots, N\}$ be given. A specialized expert i issues its forecasts not at all steps $t = 1, 2, \dots$, but only when $i \in E_t$. At any step, the aggregating algorithm uses forecasts of only "active (non-sleeping)" experts.

[2] In what follows ω_t denotes a binary outcome.

We consider a more general case. At each time moment t, any expert's forecast $f_{i,t}$ is supplied by a confidence level which is a real number $p_{i,t} \in [0,1]$.

In particular, $p_{i,t} = 1$ means that the forecast of the expert i is used in full, whereas in the case of $p_{i,t} = 0$ it is not taken into account at all (the expert sleeps). In cases where $0 < p_{i,t} < 1$, the expert's forecast is partially taken into account. For example, when moving from one season to another, an expert tuned to the previous season gradually loses his predictive ability. Confidence value can be set by the expert itself or by the learner.

The dependence of $p_{i,t}$ on values of exogenous parameters can be predetermined by a specialist in the domain or can be constructed using regression analysis on historical data.

The setting of prediction with experts that use confidence values as numbers in the interval [0,1] was studied (for Hedge algorithm) by Blum and Mansour [2] and Gaillard et al. [14]. We modify this approach for AA algorithm.

Let $\lambda(f,y)$ be an η-mixable loss function. At each time moment t the forecasts $\mathbf{f}_t = (f_{1,t}, \dots f_{N,t})$ of the experts and confidence levels $\mathbf{p}_t = (p_{1,t}, \dots, p_{N,t})$ of these forecasts are revealed.

In this section we modify AA for the experts with confidence.

To take into account confidence levels, we use the fixed point method by Chernov and Vovk [8]. We associate with any confidence level $p_{i,t}$ a probability distribution $\mathbf{p}_{i,t} = (p_{i,t}, 1-p_{i,t})$ on a two element set. Define the auxiliary probabilistic forecast:

$$\tilde{f}_{i,t} = \begin{cases} f_{i,t} \text{ with probability } p_{i,t}, \\ f_t \text{ with probability } 1-p_{i,t}, \end{cases}$$

where f_t is a forecast of the learner.

First, we provide a justification of the algorithm presented below. Our goal is to define the forecast f_t such that

$$e^{-\eta\lambda(f_t,y)} \ge \sum_{i=1}^{N} \mathbb{E}_{\mathbf{p}_{i,t}}[e^{-\eta\lambda(\tilde{f}_{i,t},y)}] w_{i,t}^* \tag{9}$$

for each y, where $\mathbf{w}_t^* = (w_{1,t}^*, \dots, w_{N,t}^*)$ is the vector of normalized weights defined by (1) and (2).

Here $\mathbb{E}_{\mathbf{p}_{i,t}}$ is the mathematical expectation with respect to the probability distribution $\mathbf{p}_{i,t}$. We rewrite inequality (9) in a more detailed form:

$$e^{-\eta\lambda(f_t,y)} \ge \sum_{i=1}^{N} \mathbb{E}_{\mathbf{p}_{i,t}}[e^{-\eta\lambda(\tilde{f}_{i,t},y)}] w_{i,t}^* = \tag{10}$$

$$\sum_{i=1}^{N} p_{i,t} w_{i,t}^* e^{-\eta\lambda(f_{i,t},y)} + e^{-\eta\lambda(f_t,y)} \left(1 - \sum_{i=1}^{N} p_{i,t} w_{i,t}^*\right) \tag{11}$$

for all ω. Therefore, the inequality (9) is equivalent to the inequality

$$e^{-\eta\lambda(f_t,y)} \ge \sum_{i=1}^{N} w_{i,t}^p e^{-\eta\lambda(f_{i,t},y)}, \tag{12}$$

where

$$w_{i,t}^p = \frac{p_{i,t} w_{i,t}^*}{\sum_{j=1}^{N} p_{j,t} w_{j,t}^*} = \frac{p_{i,t} w_{i,t}}{\sum_{j=1}^{N} p_{j,t} w_{j,t}}. \tag{13}$$

According to the rule (6) for computing the forecast of AA, define $f_t = \mathrm{Subst}(\mathbf{f}_t, \mathbf{w}_t^p)$. Then (12) and its equivalent (10) are valid. Here Subst is the substitution function, $\mathbf{w}_t^p = (w_{i,1}^p, \dots, w_{i,N}^p)$ and $\mathbf{f}_t = (f_{1,t}, \dots f_{i,N})$.. Let us refine Protocol 1 in the form of Algorithm 1a which is the algorithm AA with confidence. This algorithm presents a strategy for the learner in Protocols 1.

Algorithm 1a

FOR $t = 1, \dots, T$

1. Receive the experts' predictions $f_{i,t}$ and confidence levels $p_{i,t}$, where $1 \le i \le N$.
2. Present the learner's forecast $f_t = \mathrm{Subst}(\mathbf{f}_t, \mathbf{w}_t^p)$, where normalized weights $\mathbf{w}_t^p = (w_{1,t}^p, \dots, w_{N,t}^p)$ are defined by (13).
3. Observe the true outcome y_t and compute the losses $l_{i,t} = \lambda(f_{i,t}, y_t)$ of the experts and the loss $\lambda(f_t, y_t)$ of the learner.
4. Update the weights (of the virtual experts) by the rule

$$w_{i,t+1} = w_{i,t} e^{-\eta(p_{i,t}\lambda(f_{i,t},y_t)+(1-p_{i,t})\lambda(f_t,y_t))} \tag{14}$$

ENDFOR

Let $l_{i,t} = \lambda(f_{i,t}, y_t)$ be the loss of an expert i and $h_t = \lambda(f_t, y_t)$ be the loss of the learner at step t. Define the estimated loss of an expert i as $\tilde{l}_{i,t} = \lambda(\tilde{f}_{i,t}, y_t)$ and $\hat{l}_{i,t} = \mathbb{E}_{\mathbf{p}_{i,t}}[\tilde{l}_{i,t}]$ be its expectation. By the virtual expert i we mean the expert which suffers the loss $\hat{l}_{i,t}$.

Since by definition $\hat{l}_{i,t} = p_{i,t} l_{i,t} + (1 - p_{i,t}) h_t$, we have $h_t - \hat{l}_{i,t} = p_{i,t}(h_t - l_{i,t})$. We call the last quantity discounted excess loss with respect to an expert i at a time moment t and we will measure the performance of our algorithm by the cumulative discounted excess loss with respect to any expert i.

Theorem 1. *For any $1 \le i \le N$, the following upper bound for the cumulative excess loss (discounted regret) holds true:*

$$\sum_{t=1}^{T} p_{i,t}(h_t - l_{i,t}) \le \frac{\ln N}{\eta} \tag{15}$$

for all T.

Proof. By convexity of the exponent the inequality (9) implies

$$e^{-\eta\lambda(f_t,y)} \ge \sum_{i=1}^{N} e^{-\eta\mathbb{E}_{\mathbf{p}_{i,t}}[\lambda(\tilde{f}_{i,t},y)]} w_{i,t}^* = \sum_{i=1}^{N} e^{-\eta\hat{l}_{i,t}} w_{i,t}^*. \tag{16}$$

Rewrite the update rule (14) as $w_{i,t+1} = w_{i,t} e^{-\eta\hat{l}_{i,t}}$. Using the regret analysis for AA in A.1, we obtain

$$\sum_{t=1}^{T} h_t \le \sum_{t=1}^{T} \hat{l}_{i,t} + \frac{\ln N}{\eta}$$

for any i. Since $h_t - \hat{l}_{i,t} = p_{i,t}(h_t - l_{i,t})$, the inequality (15) follows. □

4. Aggregation of probability forecasts

Let the set of outcomes in Protocol 1 be an interval $\Omega = [a, b]$ of the real line for some $a < b$ and the set of forecasts Γ be a set of all probability distribution functions $F : [a, b] \to [0, 1]$.[3]

The quality of the prediction F in view of the actual outcome y is often measured by the continuous ranked probability score (loss function)

$$\mathrm{CRPS}(F, y) = \int_a^b (F(u) - H(u - y))^2 du, \tag{17}$$

where $H(x)$ is the Heaviside function: $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \ge 0$ (Epstein [11], Matheson and Winkler [21], etc).

For simplicity, we consider in this definition integration over a finite interval. Such definition is closer to practical applications and allows a more elementary theoretical analysis. More general definition includes a density $\mu(u)$ and integration over the real line:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{+\infty} (F(u) - H(u - y))^2 \mu(u) du. \tag{18}$$

The definition (17) is a special case of this definition (up to a factor), where $\mu(u) = \frac{1}{b-a}$ for $u \in [a, b]$ and $\mu(u) = 0$ otherwise. It can be proved that the function (18) is η-mixable for $0 < \eta \le 2$ and η-exponentially concave for $0 < \eta \le \frac{1}{2}$ (see Korotin et al. [18]).

The CRPS score measures the difference between the forecast F and a perfect forecast $H(u - y)$ which puts all mass on the verification y. The lowest possible value 0 is attained when F is concentrated at y, and in all other cases $\mathrm{CRPS}(F, y)$ will be positive.

We consider a game of prediction with expert advice, where the forecasts of the experts and of the learner are (cumulative) probability distribution functions. At any step t of the game each expert $i \in \{1, \dots, N\}$ presents its forecast – a probability distribution function $F_{i,t}(u)$ and the learner presents its forecast $F_t(u)$.[4] After an outcome $y_t \in [a, b]$ have been revealed and the experts and the learner suffer losses $\mathrm{CRPS}(F_{i,t}, y_t)$ and $\mathrm{CRPS}(F_t, y_t)$.

The corresponding game of probabilistic prediction is defined by the following protocol.

Protocol 2

FOR $t = 1, \dots, T$

1. Receive the experts' predictions – the probability distribution functions $F_{i,t}(u)$ for $1 \le i \le N$.
2. Present the learner's forecast – the probability distribution function $F_t(u)$.
3. Observe the true outcome y_t and compute the scores $\mathrm{CRPS}(F_{i,t}, y_t) = \int_a^b (F_{i,t}(u) - H(u - y_t))^2 du$ of the experts $1 \le i \le N$ and the score $\mathrm{CRPS}(F_t, y_t) = \int_a^b (F_t(u) - H(u - y_t))^2 du$ of the learner.

ENDFOR

The goal of the learner is to predict in such a way that independently of which outcomes are revealed and the experts' predictions are presented, its accumulated loss $H_T = \sum_{t=1}^{T} \mathrm{CRPS}(F_t, y_t)$ is asymptotically less than the loss $L_T^i = \sum_{t=1}^{T} \mathrm{CRPS}(F_{i,t}, y_t)$ of the best expert i up to some regret and $H_T - \min_i L_T^i = o(T)$ as $T \to \infty$.

First, we show that CRPS loss function (and the corresponding game) is mixable.

Theorem 2. *The continuous ranked probability score $\mathrm{CRPS}(F, y)$ is $\frac{2}{b-a}$-mixable loss function. The corresponding learner's forecast $F(u)$ given the forecasts $F_i(u)$ of the experts $1 \le i \le N$ and a probability distribution $\mathbf{q} = (q_1, \dots, q_N)$ on the set of all experts can be computed by the rule*[5]

$$F(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^{N} q_i e^{-2(F_i(u))^2}}{\sum_{i=1}^{N} q_i e^{-2(1-F_i(u))^2}}, \tag{19}$$

Proof. We approximate any probability distribution function $F(u)$ by a piecewise-constant function that takes a finite number of values on a uniform grid of arguments. Accordingly, the forecasts of the experts and of the learner will take the form of d-dimensional vectors, where d is a positive integer number. We apply AA to the d-dimensional forecasts, then we consider the limit $d \to \infty$.

[3] A probability distribution function is a non-decreasing function $F(y)$ defined on this interval such that $F(a) = 0$ and $F(b) = 1$. Also, it is right-continuous and has the left limit at each point.

[4] For simplicity of presentation, we consider the case where the set of the experts is finite. In case of infinite E, the sums by i should be replaced by integrals with respect to the corresponding probability distributions on the set of experts. In this case the choice of initial weights on the set of the experts is a non-trivial problem.

[5] Ii is easy to verify that $F(u)$ is a probability distribution function.

Adamskiy et al. [1] generalize the AA for the case of d-dimensional forecasts, where d is a positive integer number. Let an η-mixable loss function $\lambda(f,y)$ be given, where $\eta>0$, $f\in\Gamma$ and $y\in\Omega$. Let $\mathbf{f}=(f^1,\ldots,f^d)\in\Gamma^d$ be a d-dimensional forecast and $\mathbf{y}=(y^1,\ldots,y^d)\in\Omega^d$ be a d-dimensional outcome. The generalized loss function is defined $\lambda(\mathbf{f},\mathbf{y})=\sum_{s=1}^{d}\lambda(f^s,y^s)$; we call $\lambda(f,y)$ its source function.

The corresponding (generalized) game can be presented by Protocol 1 where at each step t the experts and the learner present d-dimensional forecasts: at any round $t=1,2,\ldots$ each expert $i\in\{1,\ldots,N\}$ presents a vector of forecasts $\mathbf{f}_{i,t}=(f^1_{i,t},\ldots,f^d_{i,t})$ and the learner presents a vector of forecasts $\mathbf{f}_t=(f^1_t,\ldots,f^d_t)$. After that, a vector $\mathbf{y}_t=(y^1_t,\ldots,y^d_t)$ of outcomes will be revealed and the experts and the learner suffer losses $\lambda(\mathbf{f}_{i,t},\mathbf{y}_t)=\sum_{s=1}^{d}\lambda(f^s_{i,t},\mathbf{y}^s_t)$ and $\lambda(\mathbf{f}_t,\mathbf{y}_t)=\sum_{s=1}^{d}\lambda(f^s_t,y^s_t)$.

Adamskiy et al. [1] proved that the generalized loss function (game) is mixable. □

Lemma 1. *The generalized loss function $\lambda(\mathbf{f},\mathbf{y})$ is $\frac{\eta}{d}$-mixable if the source loss function $\lambda(f,y)$ is η-mixable.*

We reproduce the proof in A.2 for completeness of presentation.

We now turn to the proof of Theorem 2. We approximate any probability distribution function $F(y)$ by piecewise-constant functions $F_d(y)$, where $d=1,2,\ldots$. Any such function F_d is defined by the points $z_0,z_1,z_2,\ldots,z_d$ and the values $f_0=F(z_0)$, $f_1=F(z_1)$, $f_2=F(z_2)$, $\ldots$, $f_d=F(z_d)$, where $a=z_0<z_1<z_2<\ldots<z_d=b$ and $0=f_0\le f_1\le f_2\le\cdots\le f_d=1$. By definition $F_d(y)=f_i$ for $z_{i-1}<y\le z_i$, where $1\le i\le d$. Also, assume that $z_{i+1}-z_i=\Delta$ for all $0\le i<d$. By definition $\Delta=\frac{b-a}{d}$. Since $F(u)\le F_d(u)$ for all u,

$$|\mathrm{CRPS}(F,y)-\mathrm{CRPS}(F_d,y)|\le\int_a^y(F_d^2(u)-F^2(u))du+\int_y^b((1-F(u))^2-(1-F_d(u))^2)du \tag{20}$$

for any $y\in[a,b]$. Let $z_{k-1}<y\le z_k$, where $1\le k\le d$. Then

$$\int_a^y(F_d^2(u)-F^2(u))du\le\sum_{i=0}^{k-1}\int_{z_i}^{z_{i+1}}(F_d^2(u)-F^2(u))du\le\Delta\sum_{i=0}^{k-1}F_d^2(z_{i+1})-F_d^2(z_i)=\Delta(F_d^2(z_k)-F_d^2(a))\le\Delta.$$

The second integral in (20) is also bounded by Δ. Hence,

$$|\mathrm{CRPS}(F,y)-\mathrm{CRPS}(F_d,y)|\le 2\Delta. \tag{21}$$

Define an auxiliary representation of y, which is a binary variable $\omega_{y,s}=1_{z_s\ge y}\in\{0,1\}$ for $1\le s\le d$ and $\boldsymbol{\omega}_y=(\omega_{y,1},\ldots,\omega_{y,d})$, where $1_{z_s\ge y}=H(z_s-y)$.

Consider any $y\in[a,b]$. It is easy to see that for each $1\le s\le d$ the uniform measure of all $u\in[z_{s-1},z_s]$ such that $1_{z_s\ge y}\ne 1_{u\ge y}$ is less or equal to Δ if $y\in[z_{s-1},z_s]$ and $1_{z_s\ge y}=1_{u\ge y}$ for all $u\in[z_{s-1},z_s]$ otherwise. Since $0\le f_s\le 1$ for all s, this implies that

$$\left|\mathrm{CRPS}(F_d,y)-\Delta\sum_{s=1}^{d}(f_s-\omega_{y,s})^2\right|=\left|\int_{z_{k-1}}^{z_k}(f_k-1_{u\ge y})^2du-\Delta(f_k-\omega_{y,k})^2\right|\le\Delta\left|f_k^2-(f_k-1)^2\right|=\Delta|2f_k-1|\le\Delta, \tag{22}$$

where $y\in(z_{k-1},z_k]$. Let us study the generalized loss function

$$\lambda(\mathbf{f},\boldsymbol{\omega})=\Delta\sum_{s=1}^{d}(f_s-\omega_s)^2, \tag{23}$$

where $\mathbf{f}=(f_1,\ldots,f_d)$, $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_d)$ and $\omega_s\in\{0,1\}$ for $1\le s\le d$.

The key observation is that the deterioration of the learning rate for the generalized loss function (it gets divided by the dimension d of vector-valued forecasts) is exactly offset by the decrease in the weight of each component of the vector-valued prediction as the grid-size decreases.

Since the square loss function $\lambda(f,\omega)=(\gamma-\omega)^2$ is 2-mixable, where $f\in[0,1]$ and $\omega\in\{0,1\}$, by results of Section 2 the corresponding generalized loss function $\sum_{s=1}^{d}(f_s-\omega^s)^2$ is $\frac{2}{d}$-mixable and then the loss function (23) is $\frac{2}{d\Delta}=\frac{2}{b-a}$-mixable independently of what grid-size is used.[6]

Let $F_i(u)$ be the probability distribution functions presented by the experts $1\le i\le N$ and $\mathbf{f}_i=(f_{i,1},\ldots,f_{i,d})$, where $f_{i,s}=F_i(z_s)$ for $1\le s\le d$. By (A.4)

$$e^{-\frac{2}{(b-a)}\lambda(\mathbf{f},\boldsymbol{\omega})}\ge\sum_{i=1}^{N}e^{-\frac{2}{b-a}\lambda(\mathbf{f}_i,\boldsymbol{\omega})}q_i \tag{24}$$

for each $\boldsymbol{\omega}\in\{0,1\}^d$ (including $\boldsymbol{\omega}=\boldsymbol{\omega}_y$ for any $y\in[a,b]$), where the forecast $\mathbf{f}=(f_1,\ldots,f_d)$ can be defined as

$$f_s=\frac{1}{2}-\frac{1}{4}\ln\frac{\sum_{i=1}^{N}q_ie^{-2(f_{i,s})^2}}{\sum_{i=1}^{N}q_ie^{-2(1-f_{i,s})^2}} \tag{25}$$

for each $1\le s\le d$.

By letting the grid-size $\Delta\to 0$ (or, equivalently, $d\to\infty$) in (22), (24), where $\boldsymbol{\omega}=\boldsymbol{\omega}_y$, and in (21), we obtain for any $y\in[a,b]$,

$$e^{-\frac{2}{(b-a)}\mathrm{CRPS}(F,y)}\ge\sum_{i=1}^{N}e^{-\frac{2}{b-a}\mathrm{CRPS}(F_i,y)}q_i, \tag{26}$$

where $F(u)$ is the limit form of (25) defined by

$$F(u)=\frac{1}{2}-\frac{1}{4}\ln\frac{\sum_{i=1}^{N}q_ie^{-2(F_i(u))^2}}{\sum_{i=1}^{N}q_ie^{-2(1-F_i(u))^2}}$$

for each $u\in[a,b]$.

The inequality (26) means that the loss function $\mathrm{CRPS}(F,y)$ is $\frac{2}{b-a}$-mixable. □

Let us refine the protocol 2 of the game with probabilistic predictions for the case when the rule (19) for AA is used. This algorithm presents a strategy for the learner in Protocol 2.

Algorithm 3

Define $w_{i,1}=\frac{1}{N}$ for $1\le i\le N$.
FOR $t=1,\ldots,T$

1. Receive the expert predictions – the probability distribution functions $F_{i,t}(u)$, where $1\le i\le N$.
2. Present the learner forecast – the probability distribution function $F_t(u)$:
$$F_t(u)=\frac{1}{2}-\frac{1}{4}\ln\frac{\sum_{i=1}^{N}w^*_{i,t}e^{-2(F_{i,t}(u))^2}}{\sum_{i=1}^{N}w^*_{i,t}e^{-2(1-F_{i,t}(u))^2}}, \tag{27}$$
where $w^*_{i,t}=\frac{w_{i,t}}{\sum_{j=1}^{N}w_{j,t}}$.
3. Observe the true outcome y_t and compute the score $\mathrm{CRPS}(F_{i,t},y_t)$ for the experts $1\le i\le N$ and the score $\mathrm{CRPS}(F_t,y_t)$ for the learner.

[6] This also means that in numerical experiments, when calculating forecasts of the learner, we can use the same learning rate, regardless of the accuracy of the presentation of expert forecasts.

4. Update the weights of the experts $1 \le i \le N$

$$w_{i,t+1} = w_{i,t} e^{-\frac{2}{b-a}\mathrm{CRPS}(F_{i,t}, y_t)} \tag{28}$$

ENDFOR

The performance bound of Algorithm 3 is presented in the following theorem.

Theorem 3. *For each T,*

$$\sum_{t=1}^{T} \mathrm{CRPS}(F_t, y_t) \le \min_{1\le i\le N} \sum_{t=1}^{T} \mathrm{CRPS}(F_{i,t}, y_t) + \frac{b-a}{2} \ln N. \tag{29}$$

Proof. The bound (29) is a direct corollary of the regret analysis of A.1 and the bound (A.1). □

The square loss function is also η-exponentially concave for $0 < \eta \le \frac{1}{2}$ (see Cesa-Bianchi and Lugosi [9]). In this case (27) can be replaced with the forecast WA

$$F_t(u) = \sum_{i=1}^{N} w^*_{i,t} F_{i,t}(u), \tag{30}$$

where $w^*_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}}$ are normalized weights. The corresponding weights are computed recursively

$$w_{i,t+1} = w_{i,t} e^{-\frac{1}{2(b-a)}\mathrm{CRPS}(F_{i,t}, y_t)}. \tag{31}$$

Using Lemma 1 and Theorem 3, we conclude that in this case the bound (29) can be replaced with

$$\sum_{t=1}^{T} \mathrm{CRPS}(F_t, y_t) \le \min_{1\le i\le N} \sum_{t=1}^{T} \mathrm{CRPS}(F_{i,t}, y_t) + 2(b-a) \ln N.$$

The proof is similar to the proof of Theorem 3. □

4.1. Aggregation of probabilistic predictions with confidence

In Section 5.2 (below), we present results of numerical experiments with the real data and when probabilistic predictions of the experts are supplied with the levels of confidence. In this case we use Algorithm 3a as a strategy of the learner, that is a modification of Algorithm 3.

At each round, only some of the experts output a prediction while the other ones are inactive. Each expert is expected to provide accurate forecasts mostly under given external conditions that can be known beforehand, namely, the experts are specialized to a season, temperature forecast, and time of the day.

We define a smooth extension of the domain of any expert. Thus, each expert competes with other experts working at overlapping intervals.

The aggregating algorithms AA and WA allow us to combine the probabilistic predictions of the specialized experts with overlapping domains of theirs competence.

Algorithm 3a (Strategy for the learner)

Define $w_{i,1} = \frac{1}{N}$ for $1 \le i \le N$.
FOR $t = 1, \ldots, T$

1. Receive the expert predictions – the probability distribution functions $F_{i,t}(u)$ and confidence levels $p_{i,t}$, where $1 \le i \le N$.
2. Present the learner forecast – the probability distribution function $F_t(u)$ which is defined by the rule

$$F_t(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^{N} w^p_{i,t} e^{-2(F_{i,t}(u))^2}}{\sum_{i=1}^{N} w^p_{i,t} e^{-2(1-F_{i,t}(u))^2}} \tag{32}$$

for AA or by the rule

$$F_t(u) = \sum_{i=1}^{N} w^p_{i,t} F_{i,t}(u) \tag{33}$$

for WA, where $w^p_{i,t} = \frac{p_{i,t} w_{i,t}}{\sum_{j=1}^{N} p_{j,t} w_{j,t}}$.

3. Observe the true outcome y_t and compute the score $\mathrm{CRPS}(F_{i,t}, y_t)$ for the experts $1 \le i \le N$ and the score $\mathrm{CRPS}(F_t, y_t)$ for the learner.
4. Update the weights of the (virtual) experts $1 \le i \le N$

$$w_{i,t+1} = w_{i,t} e^{-\eta(p_{i,t}\mathrm{CRPS}(F_{i,t}, y_t) + (1-p_{i,t})\mathrm{CRPS}(F_t, y_t))}, \tag{34}$$

where $\eta = \frac{2}{b-a}$ for AA and $\eta = \frac{1}{2(b-a)}$ for WA.

ENDFOR

The performance of the algorithm is presented by the inequality (15) of Theorem 1, where $h_t = \mathrm{CRPS}(F_t, y_t)$, $l_{i,t} = \mathrm{CRPS}(F_{i,t}, y_t)$ and $\eta = \frac{2}{b-a}$ if the rule (32) for computing the learner's forecast was used and $\eta = \frac{1}{2(b-a)}$ if the rule (33) was used.

The proposed rules (32) for AA and (33) for WA can be used when the probability distributions presented by the experts are given in the closed form (i.e., distributions given by analytical formulas). For this case, numerical methods can be used to calculate the integrals (CRPS) with any degree of accuracy given in advance (see also Footnote 6).

5. Experiments

In this section we apply our proposed algorithm on synthetic data and on electricity consumption data, and compare its performance for several predictive models. We use Algorithm 3 in the experiments with synthetic data and Algorithm 3a for the electricity consumption data.

To optimize the losses in our mixing schemes, we used the mixing past posteriors modification of Algorithms 3 and 3a, see A.3.

The algorithms and the data are presented at GitHub: https://github.com/VladimirVyugin, Project "Online-Aggregation-of-Probability-Forecasts -With-Confidence"

5.1. Synthetic data

In this section we present the results of experiments with AA and WA on synthetic data. The data for experiments were obtained by sampling from a mixture of the three distinct probability distributions with the triangular densities. The time interval is made up of several segments of the same length, and the weights of the components of the mixture depend on time. We use two methods of mixing of the three distinct initial probability distributions. By Method 1, only one generating probability distribution is a leader at each segment (i.e., its weight is equal to one). By Method 2, the weights of the mixture components vary smoothly over time (as shown in section B of Fig. 1).

Fig. 1 shows the main stages of data mixing (Method 1 – left, Method 2 - right) and the results of aggregation of the experts models. Section A of the figure shows the realizations of the trajectories of the three data generating distributions. The diagram in Section B displays the actual prior probabilities (relative weights) that were used for mixing of the probability distributions. Section C shows the result of sampling from the mixture distribution.

There are three experts $i = 1, 2, 3$, which assume that the time series under study is obtained as a result of sampling from the probability distribution with the fixed triangular density with given peak and base. Each expert evaluates the similarity of the testing point of the series with its distribution using CRPS score.

We also compare two rules of aggregation of the experts' forecasts, AA (27) and the weighted average WA (30). The diagrams of Sections D and E of Fig. 1 show the weights of the experts assigned by the corresponding algorithm in the online aggregating process using rules (27) and (30).

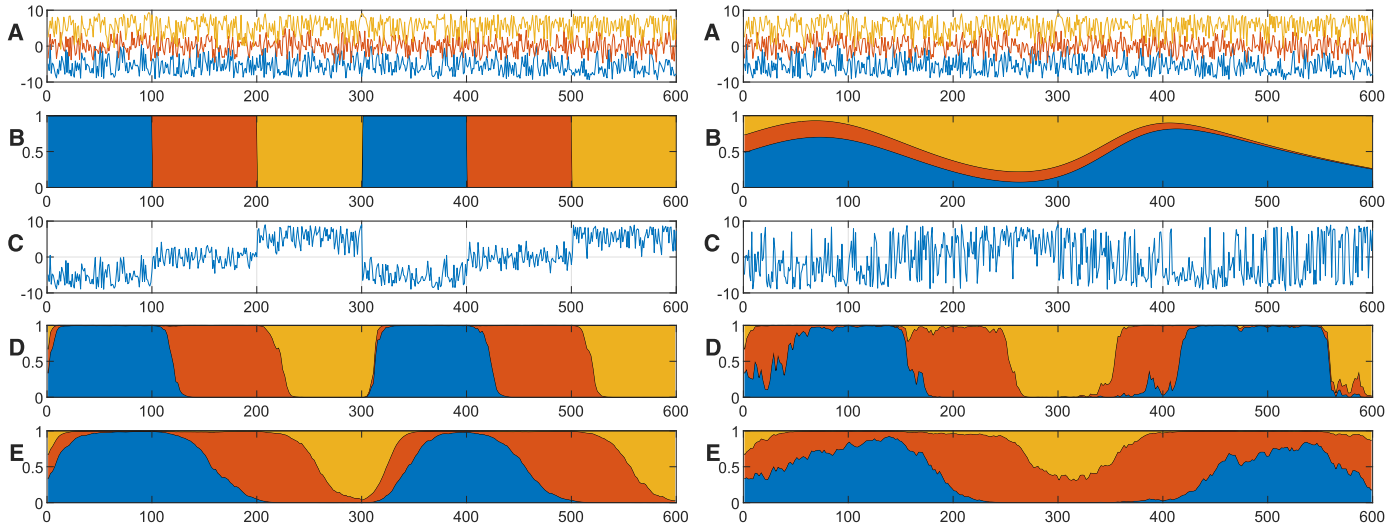

Fig. 1. The stages of numerical experiments and the results of experts' aggregation for two initial synthetic data mixing methods (Method 1 – left, Method 2 - right). (A) Realizations of the trajectories for the three initial data generating distributions; (B) weights of the distributions assigned by the data mixing method; (C) sequence sampled from the distributions defined by Method 1 and Method 2; (D) weights of the experts assigned online by AA using the rule (28); (E) weights of the experts assigned online by WA using the rule (31). .

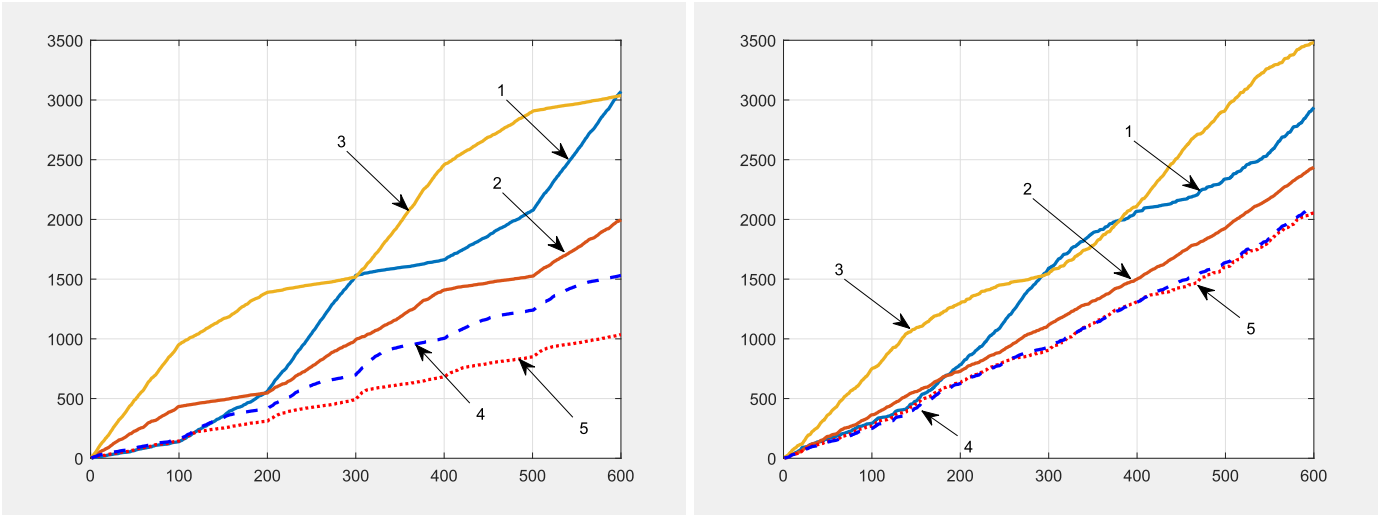

Fig. 2. The accumulated losses of the experts (lines 1–3) and of the aggregating algorithm for both initial data mixing methods (Method 1 – left, Method 2 - right) and for both methods of computing aggregated forecasts: line 4 – for WA (the rule (30)) and line 5 – for AA (the rule (27)). We note an advantage of AA over WA in the case of data generating Method 1, in which there is a rapid change in leadership of the data generating distributions.

Fig. 2 shows the accumulated losses of the experts and the accumulated losses of the aggregating algorithm for both data generating methods (Method 1 – left, Method 2 - right) and for both methods of computing the aggregated forecasts – by the rule (27) and by the rule (30). We note an advantage of rule (27) over the rule (30) in the case of data generating Method 1, in which there is a rapid change in leadership of the data generating models.

Fig. 3 shows in 3D format the empirical distribution functions obtained online by Algorithm 3 for both data generating models and the rule (27).

5.2. Probabilistic forecasting of electrical loads

The second group of numerical experiments on probabilistic forecasting were performed with the data of the 2014 (GEFCOM 2014,Track Load, Hong et al. [23]). The time series were divided into training (about 5 years) and testing (about 1 year) samples.

The main unit of the training sample includes data on hourly electrical load and data on hourly temperature measurements for all days of training period.

The training sample shows the dependence of electrical loads on temperature which looks differently during different seasons and time of the day. Therefore, each expert is trained on its specific domain where the specific relationship between temperature and electrical load is observed. We use the corresponding point clouds of "temperature–loads" to define the probability distribution function of the expert.

The scatter diagrams "Load - Temperature" for several sets of calendar parameters (four seasons of the year and four consecutive intervals of the day, each for 6 hours) are presented in Fig. 4. The diagrams are constructed according to the training part of the sample.

Fig. 4 shows the nature of the relationship between potential predictors and response. These data show the dependence of electrical loads on temperature. For each of the scattering diagrams

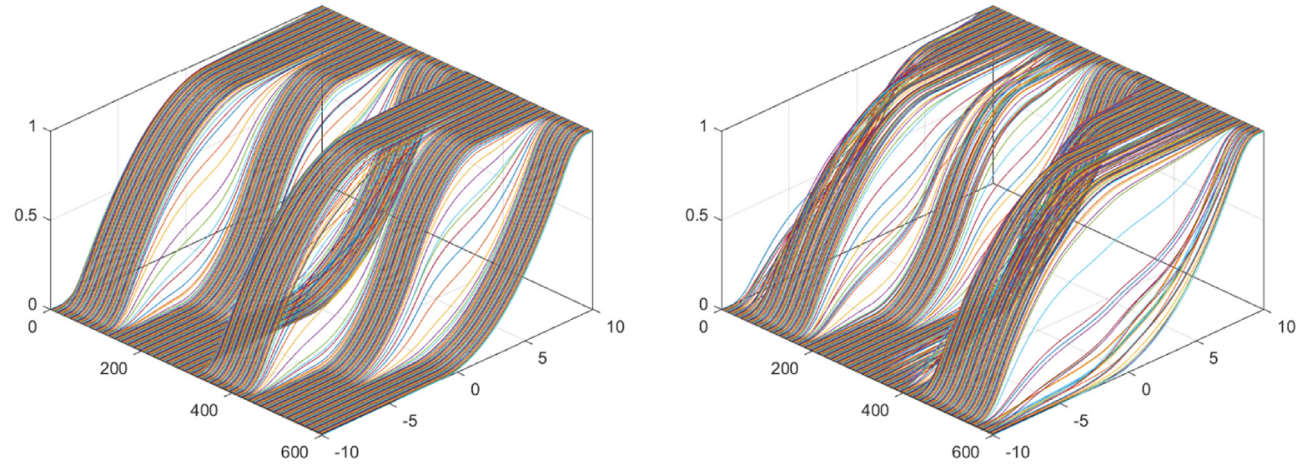

Fig. 3. Empirical distribution functions obtained online as a result of aggregation of the distributions of three experts by the rule (27) for both data generating methods.

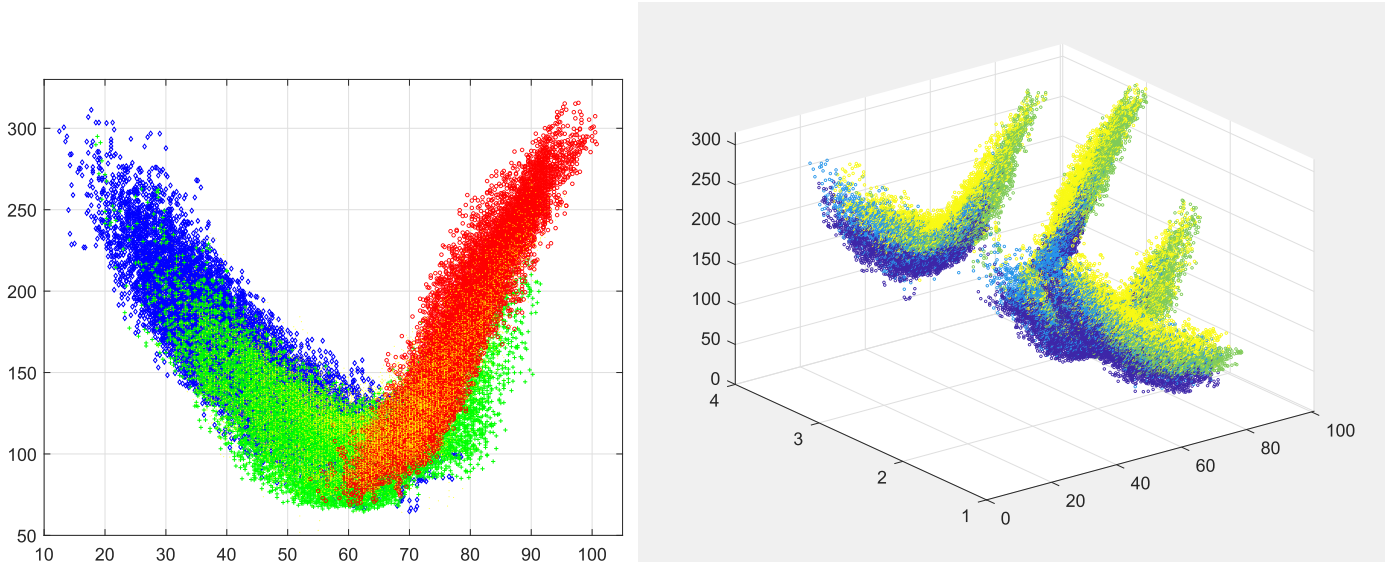

Fig. 4. Scatter plots of hourly temperature and electrical loads for all days of training period: Left figure – all data marked by seasons; Right figure – data grouped by seasons (Winter, Spring, Summer, Autumn) and time of the day marked in color (Night, Morning, Day, Evening).

presented, two or three temperature intervals can be distinguished in such a way that within each interval the point cloud has a simple ellipsoidal shape. This provides the basis for using a mixture of normal distributions for the probabilistic forecast of the expected electrical load according to the short-term temperature forecast.

Scatter patterns on Fig. 4 can serve as the basis for determining the pool of the experts. Each of them learns (a predictive probabilistic model) at sample points related to a predefined calendar segment, for example, "Winter&Morning", etc. These segments should cover all possible combinations of calendar indicators present in the data.

A set of 21 specialized experts is defined by dividing the calendar space into domains where the relationship between temperature and electrical load can be described by a simple and relatively uniform dependence. To define an expert, a combined sample of historical data consisting of the initial sample of "temperature–load" ensemble, as well as its competence area (season, time of the day) was determined. Each expert represents the temperature dependence of the probabilistic distribution of the magnitude of the electrical load within a certain domain. These domains represent four daily periods (morning, afternoon, evening, night) for each season (winter, spring, summer, autumn). There are 16 such experts in total, they have numbers 6–21.

The anytime Expert 1 corresponds to the left part of Fig. 4, Experts 2–5 correspond to four seasons (see right part of Fig. 4). Experts 6–21 correspond to the colored parts of the plots on the right part of Fig. 4. To construct the probability distribution of any expert, we use the method of Gaussian Mixture Models (GMM), which is applied to the corresponding ensemble of "temperature–load". This probabilistic model of any expert is presented as a mixture of normal distributions. The number of components in a Gaussian mixture is preselected (from 1 to 3) depending on the complexity of the scattering cloud shape for "temperature–load" pair constructed from the training sample.

The main parameter of any expert's model (algorithm) is the temperature forecast. Therefore, the predictive performance of our algorithm extends as far as the temperature forecast allows.

In the experiments, which are presented in Figs. 5–8, a particular forecasting problem is considered, that is the short-term forecasting of a probability distribution function for one hour in advance. We use the current temperature as its forecast on one hour ahead.

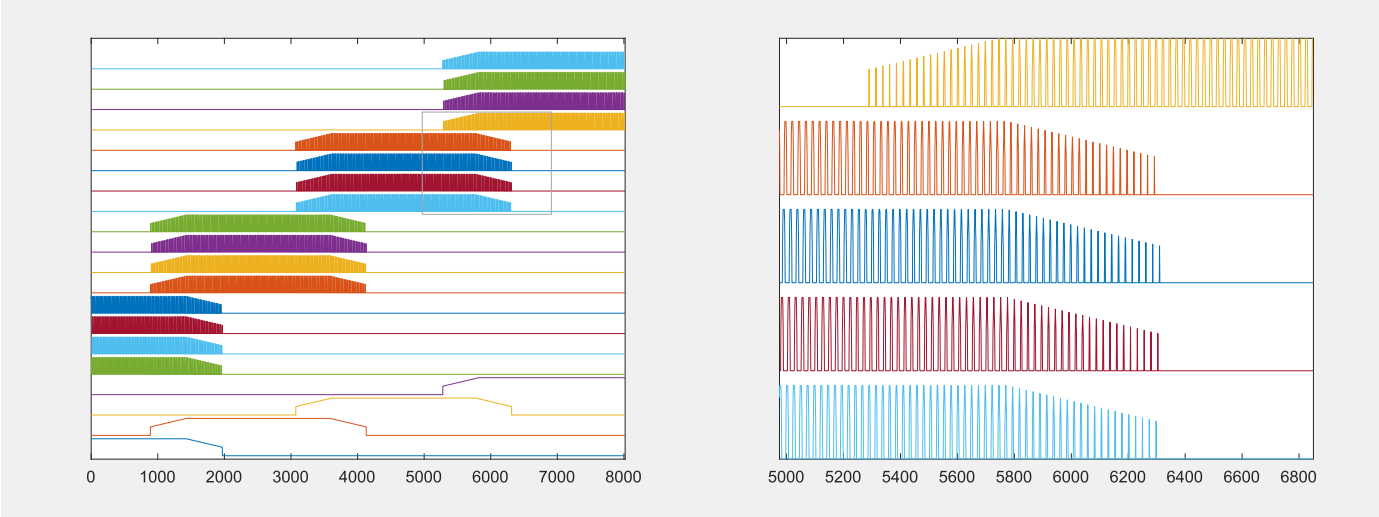

Fig. 5. Left part: confidence levels for Experts 2–5 (season experts) and 6–21 ("season&time of the day"). Right part: enlarged fragment. Each block is the result of overlaying the confidence levels of the corresponding easonal expert with the confidence levels of the day experts. The horizontal axis displays time, the blocks are vertically spaced.

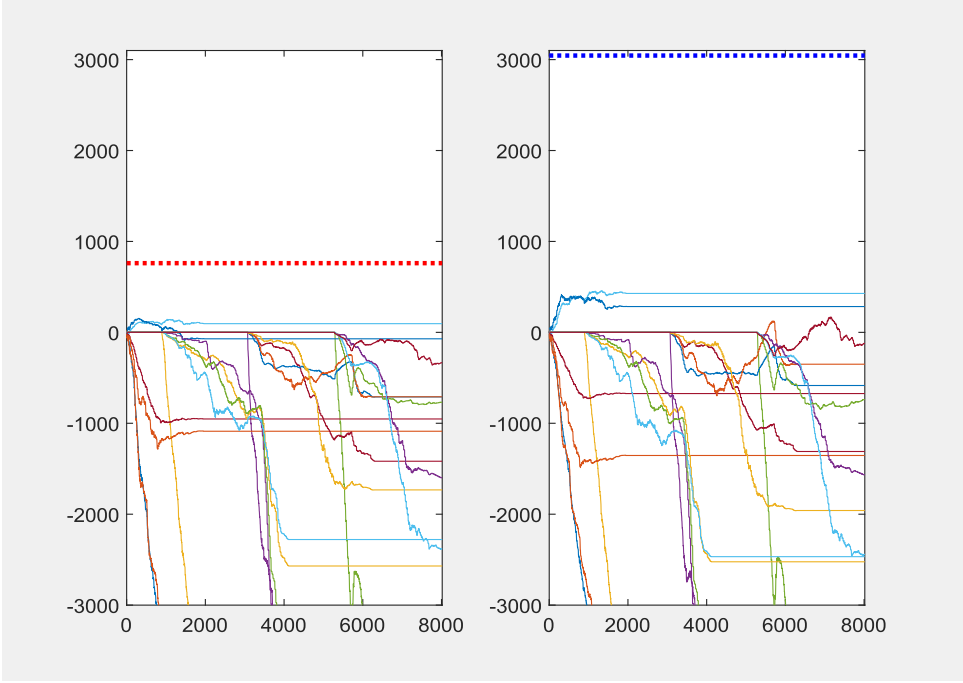

Fig. 6. Discounted regret curves for AA (left) and WA (right) with respect to each of 21 specialized experts. The dotted lines above represent the theoretical bounds for the regret.

Each expert is trained on its specific domain of time interval. The scope of each expert is determined by its confidence values. Moving from one domain to another, an expert, which was tuned to the previous domain, gradually loses his predictive abilities. To take this into account, when forecasting, we define a smooth extension of the domain of any expert. Inside the area for which the expert was tuned, its confidence values are equal to 1, and outside this area they decrease linearly from 1 to 0; moreover, the area of decrease for a seasonal expert is equal to half of the duration of the season, and the area of decrease for a daily expert is equal to two hours (the specific domain of any daily expert is equal to six hours).

Confidence levels of Seasonal Experts 2–5, as well as corresponding Experts 6–21, are presented as blocks on Fig. 5. Each block is the result of overlaying the confidence levels of the corresponding seasonal expert with the confidence levels of the experts.[7]

The constructed experts and methods of their aggregation were tested on the testing sample. Temperature and hourly electrical loads for the testing period are presented on Fig. 7.

When forecasting, the smooth areas of expert competence are chosen wider than those areas in which these experts were trained. Thus, each expert competes with other experts working at overlapping intervals using the corresponding algorithm for combining experts with confidence levels from Section 4.1, like it was done for computing the pointwise forecasts by V'yugin and Trunov [29].

[7] These values are simply multiplied.

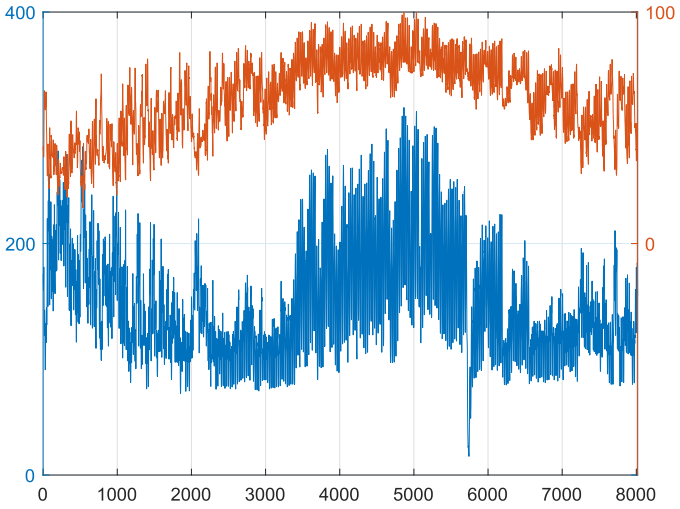

Fig. 7. Temperature (top graph) and hourly electrical loads (bottom graph) for the testing period. The left vertical axis is the load value, the right vertical axis is the temperature in Fahrenheit scale. There is a jump of consumption between 5000 and 6000 hours of testing period, which is then reflected in the results of the forecasting algorithms.

The regret curves $T \to \sum_{t=1}^{T} p_{i,t}(h_t - l_{i,t})$ for AA and WA with respect to each of 21 specialized experts are presented in Fig. 6. The dotted lines above represent the theoretical bounds for the regret (see the inequality (15)).

Two ways of aggregation of the experts by AA and WA were tested. In the first method of aggregation, confidence levels of all experts were equal to 1. In the second way, algorithms AA and WA use specialized experts, where theirs confidence levels are set externally. Non-zero confidences correspond to the training intervals of specialized experts, but are somewhat wider and monotonically decrease to zero outside these intervals (see example in Fig. 5).

To justify the role of confidence parameters, the comparative experiments were conducted. Their results are presented in Fig. 8. The accumulated losses and their time averages are presented in Fig. 8. These curves show that specialized experts, which were trained only for certain types of data, quickly lose their effectiveness in other types of data areas and generally suffer large losses. An exception is Expert 1, which was trained on all types of data, but the aggregating algorithms AA and WA with confidence essentially outperform it.

Other experiments study the effects of smooth and constant confidence levels. During the first experiment, all confidence values for each expert were equal to 1: curves 2 and 3 (in Fig. 8) represent results of their aggregation by AA and WA, where confidence levels of the experts are set to 1. In the second experiment, AA and WA algorithms used the experts predictions within the levels of their confidence: curves 4–5 represent results of aggregation by WA and AA algorithms using non-trivial overlapping smooth confidence levels.

We also test the binary case, where confidence levels of the experts take only values 0 or 1 (sleeping and non-sleeping experts): curves 6–7 represent results of aggregation by WA and AA for the binary case where the expertise areas of the experts do not overlap.

The results of the experiments show that the use of smooth confidence levels of specialized experts increases the efficiency of the process of online adaptation compared to those cases where confidence values are binary or when they are not used at all (when they are always equal to 1).

These results also show that AA in all experiments outperforms WA.

6. Conclusion

In this paper, the problem of aggregating the probabilistic forecasts is considered. In this case, Continuous Ranked Probability Score (CRPS) is a popular among practitioners example of proper scoring rule for continuous outcomes. We incorporate this loss function in PEA framework and present its theoretical analysis. We have proved that the CRPS loss function is mixable. This implies that all machinery of the Vovk aggregating algorithm can be applied to this loss function. Basing on mixability of CRPS, we analyze two methods for calculating the predictions using the aggregating algorithm (AA) and the weighted average of forecasts of the experts (WA). The time-independent upper bounds for the regret were obtained for both methods.

We illustrate the theoretical results with computer experiments. In Section 5.1 we test the performance of two methods of aggregation, AA and WA, on synthetic data. We use three probabilis-

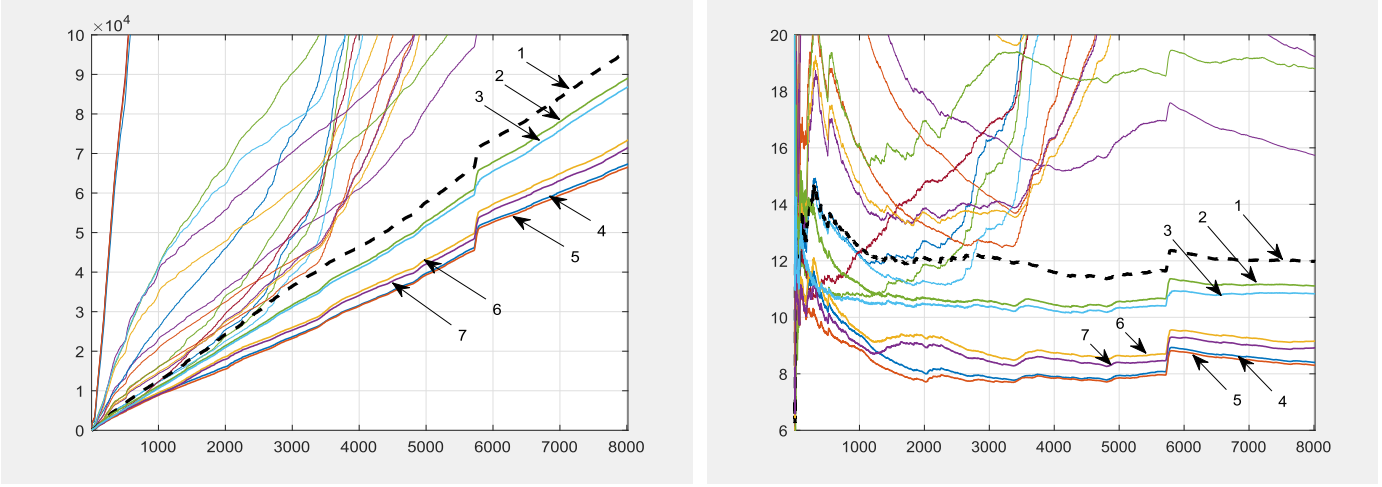

Fig. 8. Comparative study of learning with/without specialization of the experts. Accumulated losses (left) and their time averages (right): of all 21 specialized experts working any time (there is some difference with curves on Fig. 6, where the discounted regrets are presented); 1– losses of the anytime expert trained on the entire sample; 2 and 3 – results of aggregation by WA and AA, where confidence levels of the experts are set to 1; 4–5 – results of aggregation by WA and AA algorithms using non-trivial overlapping smooth confidence levels; 6–7 – the same for the case where the expertise areas of the experts do not overlap (sleeping and non-sleeping experts). AA is always slightly outperforms WA.

tic models for generating data. The same models are used as experts. Our experiments show how quickly the mixing algorithms can adapt to the data generation strategy (see Fig. 1).

These results show that two methods of computing forecasts AA and WA lead to similar empirical cumulative losses while the rule (27) for AA results in four times less regret bound than (30) for WA. We note a significantly better performance of method AA over method WA (30) in the case where there is a rapid change in leadership of the data generating models.

We have incorporated a smooth generalization of the method of specialized experts into the aggregating algorithm, which allows us to combine the probabilistic predictions of the specialized experts with overlapping domains of theirs competence.

This paper applies our approach to a popular problem of predicting electricity consumption using Gaussian mixture models as experts. We propose a technology for developing specialized experts and learning their probability distributions using ensembles of learning samples.

A set of 21 specialized experts is defined by dividing the calendar space into domains where the relationship between temperature and electrical load can be described by a simple and relatively uniform dependence. The main parameter of any expert's model (algorithm) is the temperature forecast. Therefore, the predictive performance of our algorithm extends as far as the temperature forecast allows. The problem of predicting temperature for several hours in advance is beyond the scope of this study and is a separate problem that may be the subject of future research.

The results of these experiments show that the use of smooth confidence levels of specialized experts increases the efficiency of the process of online adaptation compared to those cases where confidence values are binary or when they are not used at all.

The proposed methods are closely related to the so called ensemble forecasting (Thorey et al. [24]). In practice, the output of physical process models are usually not probabilities, but rather ensembles. Ensemble forecasts are based on a set of physical models. Each model may have its own physical formulation, numerical formulation and input data. An ensemble is a collection of model trajectories generated using different initial conditions of model equations. Consequently, the individual ensemble members represent likely scenarios of the future physical system development, consistent with the currently available incomplete information. It is possible to apply the aggregation methods developed directly to the data represented in the form of ensembles.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This paper is an extended version of COPA 2019 (Conformal and Probabilistic Prediction with Applications) paper by V'yugin and Trunov [30]. This work was partially supported by the Russian Foundation for Basic Research, project 20-01-00203. The authors are grateful to Vladimir Vovk and Yuri Kalnishkan for useful discussions. The authors thank the anonymous reviewers, whose comments significantly improved the presentation of this work.

Appendix A. Auxiliary results

A1. Regret analysis for AA

Assume that a loss function $\lambda(f,y)$ is η-mixable. Let $\mathbf{w}_t^* = (w_{1,t}^*,\dots,w_{N,t}^*)$ be the normalized weights and $\mathbf{f}_t = (f_{1,t},\dots,f_{N,t})$ be the experts' forecasts at step t. Define in Protocol 1 the learner's forecast $f_t = \mathrm{Subst}(\mathbf{f}_t,\mathbf{w}_t^*)$. By (4) $\lambda(f_t,y_t) \le g_t(y_t)$ for all t, where $g_t(y)$ is defined by (3).

Let $H_T = \sum_{t=1}^{T}\lambda(f_t,y_t)$ be the accumulated loss of the learner and $L_T^i = \sum_{t=1}^{T}\lambda(f_{i,t},y_t)$ be the accumulated loss of an expert i. By definition $g_t(y_t) = -\frac{1}{\eta}\ln\frac{W_{t+1}}{W_t}$, where $W_t = \sum_{i=1}^{N} w_{i,t}$ and $W_1 = 1$. By the weight update rule (1), we obtain $w_{i,t+1} = \frac{1}{N}e^{-\eta L_t^i}$.

By telescoping, we obtain the time-independent bound

$$H_T \le \sum_{t=1}^{T} g_t(y_t) = -\frac{1}{\eta}\ln W_{T+1} \le L_T^i + \frac{\ln N}{\eta} \tag{A.1}$$

for any expert i regardless of which sequence of outcomes is observed.

A2. Proof of lemma 1

Proof. Let the forecasts $\mathbf{c}_i = (c_i^1,\dots,c_i^d)$ of the experts $1 \le i \le N$ and a probability distribution $\mathbf{p} = (p_1,\dots,p_N)$ on the set of the experts be given.

Since the loss function $\lambda(f,y)$ is η-mixable, we can apply the aggregation rule to each sth column $\mathbf{e}^s = (c_1^s,\dots,c_N^s)$ of coordinates separately: define $f^s = \mathrm{Subst}(\mathbf{e}^s,\mathbf{p})$ for $1 \le s \le d$. Rewrite the inequality (4):

$$e^{-\eta\lambda(f^s,y)} \ge \sum_{i=1}^{N} e^{-\eta\lambda(c_i^s,y)} p_i \tag{A.2}$$

for $1 \le s \le d$ and for any y.

Let $\mathbf{y} = (y^1,\dots,y^d)$ be a vector of outcomes. Multiplying the inequalities (A.2) for $s = 1,\dots,d$ and $y = y^s$, we obtain

$$e^{-\eta\sum_{s=1}^{d}\lambda(f^s,y^s)} \ge \prod_{s=1}^{d}\sum_{i=1}^{N} e^{-\eta\lambda(c_i^s,y^s)} p_i. \tag{A.3}$$

The generalized Hölder inequality says that

$$\|G_1 G_2\cdots G_d\|_r \le \|G_1\|_{q_1}\|G_2\|_{q_2}\cdots\|G_d\|_{q_d},$$

where $\frac{1}{q_1}+\cdots+\frac{1}{q_d} = \frac{1}{r}$, $q_s \in (0,+\infty)$ and $G_s \in L^{q_s}$ for $1 \le s \le d$ (Lo'eve [20]). Let $q_s = 1$ for all $1 \le s \le d$, then $r = 1/d$. Let $G_s(i) = e^{-\eta\lambda(c_i^s,y^s)}$ for $s = 1,\dots,d$ and $\|G_s\|_1 = \mathbb{E}_{i\sim\mathbf{p}}[G_s(i)] = \sum_{i=1}^{N} G_s(i)p_i$. Then using the inequality (A.3), we obtain

$$e^{-\eta\sum_{s=1}^{d}\lambda(f^s,y^s)} \ge \left(\sum_{i=1}^{N} e^{-\eta\frac{1}{d}\sum_{s=1}^{d}\lambda(c_i^s,y^s)} p_i\right)^d.$$

or, equivalently,

$$e^{-\frac{\eta}{d}\lambda(\mathbf{f},\mathbf{y})} \ge \sum_{i=1}^{N} e^{-\frac{\eta}{d}\lambda(\mathbf{c}_i,\mathbf{y})} p_i \tag{A.4}$$

for all $\mathbf{y} = (y^1,\dots,y^d)$, where $\mathbf{f} = (f^1,\dots,f^d)$.

The inequality (A.4) means that the generalized loss function $\lambda(\mathbf{f},\mathbf{y})$ is $\frac{\eta}{d}$-mixable.

By (1), the weights update rule for generalized loss function in Protocol 1 is

$$w_{i,t+1} = w_{i,t}e^{-\frac{\eta}{d}\lambda(\mathbf{f}_{i,t},\mathbf{y}_t)} \text{ for } t = 1,2,\dots,$$

where $\eta > 0$ is a learning rate for the source function. The normalized weights $\mathbf{w}_t^* = (w_{i,t}^*,\dots,w_{i,t}^*)$ are defined by (2). At any round t, the learner forecast $\mathbf{f}_t = (f_t^1,\dots,f_t^d)$ is defined as $f_t^s = \mathrm{Subst}(\mathbf{e}_t^s,\mathbf{w}_t^*)$ for each $s = 1,\dots,d$, where $\mathbf{e}_t^s = (f_{1,t}^s,\dots,f_{N,t}^s)$. □

Table A.1
Some values of the parameter α and the corresponding accumulated losses of Algorithm 3, when the first synthetic data generation model (Method 1) was used. The values of the losses are normalized relative to the losses of the algorithm WA for $\alpha = 0$.

α	0	0.0001	0.001	0.005	0.01	0.05	0.1	0.2
AA	0.984	0.596	0.542	0.513	0.508	0.564	0.657	0.824
WA	1.000	0.958	0.869	0.759	0.728	0.816	0.957	1.115

A3. Mixing past posteriors

We have used mixing past posteriors modification of Algorithms 3 and 3a (see Fig. 1 and mixing scheme Fixed Share Update (to start vector) on Table 1 by [3]), where the rules (28) and (34) are replaced with

$$w_{i,t+1} = \frac{\alpha}{N} + (1-\alpha)\frac{\tilde{w}_{i,t}}{\sum_{j=1}^{N}\tilde{w}_{j,t}}, \text{ where}$$
$$\tilde{w}_{i,t} = w_{i,t}e^{-\eta(p_{i,t}\mathrm{CRPS}(F_{i,t},y_t)+(1-p_{i,t})\mathrm{CRPS}(F_t,y_t))}.$$

The value of parameter α was not optimized. Some values of the parameter α and the corresponding accumulative losses of Algorithm 3 for the first synthetic data generation model (Method 1) are presented on Table A.1

The loss values given in the table show that in this particular case, a significant decrease in losses occurs already at the first nonzero value of the parameter α. There is a jump in losses at the first nonzero tested value $\alpha = 0.0001$, after which their change was insignificant. We have chosen the value $\alpha = 0.001$ within the interval of relative stabilization of the corresponding losses. Optimization of the parameter value α can serve as a subject for further research.

References

[1] D. Adamskiy, T. Bellotti, R. Dzhamtyrova, Y. Kalnishkan, Aggregating algorithm for prediction of packs, Mach Learnhttps://link.springer.com/article/10.1007/s10994-018-5769-2,(arXiv:1710.08114 [cs.LG]).
[2] A. Blum, Y. Mansour, From external to internal regret, Journal of Machine Learning Research 8 (2007) 1307–1324.
[3] O. Bousquet, M. Warmuth, Tracking a small set of experts by mixing past posteriors, Journal of Machine Learning Research 3 (2002) 363–396.
[4] G.W. Brier, Verification of forecasts expressed in terms of probabilities, Mon. Weather Rev. 78 (1950) 1–3.
[5] J. Bröcker, L.A. Smith, Scoring probabilistic forecasts: the importance of being proper, Weather Forecasting 22 (2007) 382–388.
[6] J. Bröcker, L.A. Smith, From ensemble forecasts to predictive distribution functions, Tellus A 60 (2008) 663–678.
[7] J. Bröcker, Evaluating raw ensembles with the continuous ranked probability score, Q. J. R. Meteorol. Soc. 138 (2012) 1611–1617.
[8] A. Chernov, V. Vovk, Prediction with expert evaluators advice. in algorithmic learning theory, ALT 2009, Proceedings, volume 5809 of LNCS, pages 8–22. Springer (2009).
[9] N. Cesa-Bianchi, G. Lugosi, Prediction, learning, and games, Cambridge University Press, 2006.
[10] M. Devaine, P. Gaillard, Y. Goude, G. Stoltz, Forecasting electricity consumption by aggregating specialized experts, Mach Learn 90 (2) (2013) 231–260.
[11] E.S. Epstein, A scoring system for probability forecasts of ranked categories, J. Appl. Meteorol. Climatol. 8 (1969) 985–987.
[12] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J Comput Syst Sci 55 (1997) 119–139.
[13] Y. Freund, R.E. Schapire, Y. Singer, M.K. Warmuth, Using and combining predictors that specialize. in:, Proc. 29th Annual ACM Symposium on Theory of Computing334–343
[14] P. Gaillard, G. Stoltz, T. van Erven., A second-order bound with excess losses, JMLR: Workshop and Conference Proceedings 35 (2014) 1–21.
[15] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Amer. Statist. Assoc. 102 (2007) 359–378.
[16] Y. Kalnishkan, D. Adamskiy, A. Chernov, T. Scarfe, Specialist experts for prediction with side information, IEEE International Conference on Data Mining Workshop (ICDMW).IEEE (2015) 1470–1477.
[17] J. Kivinen, M.K. Warmuth, Averaging expert prediction. in paul fisher and hans ulrich simon, editors, Computational Learning Theory: 4th European Conference (EuroColt '99) (1999) 153–167. Springer
[18] A. Korotin, V. V'yugin, E. Burnaev, Integral mixabilty: a tool for efficient online aggregation of functional and probabilistic forecasts, arXiv:1912.07048 [cs.LG] (2019). https://arxiv.org/abs/1912.07048
[19] N. Littlestone, M. Warmuth, The weighted majority algorithm, Information and Computation 108 (1994) 212–261.
[20] M. Lo'eve, Probability theory, I. Springer, 4th edition (1977).
[21] J.E. Matheson, R.L. Winkler, Scoring rules for continuous probability distributions, Manage Sci 22 (10) (1976) 1087–1096, doi:10.1287/mnsc.22.10.1087.
[22] A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev. 133 (2005) 1155–1174.
[23] T. Hong, P. Pinson, S. Fanc, H. Zareipour, A. Troccoli, R.J. Hyndman, Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond, Int J Forecast 32 (2016) 896–913.
[24] J. Thorey, V. Mallet, P. Baudin, Online learning with the continuous ranked probability score for ensemble forecasting, Q. J. R. Meteorolog. Soc. 143 (2017) 521–529. 10.1002/qj.2940
[25] V. Vovk, Aggregating strategies. in m. fulk and j. case, editors, Proceedings of the 3rd Annual Workshop on Computational Learning Theory (1990) 371–383. San Mateo, CA, Morgan Kaufmann
[26] V. Vovk, A game of prediction with expert advice, J Comput Syst Sci 56 (2) (1998) 153–173.
[27] V. Vovk, Competitive on-line statistics, International Statistical Review 69 (2001) 213–248.
[28] V. Vovk, J. Shen, V. Manokhin, Min-ge xie. nonparametric predictive distributions based on conformal prediction, Mach Learn 108 (3) (2019) 445–474, doi:10.1007/s10994-018-5755-8.
[29] V. V'yugin, V. Trunov, Online aggregation of unbounded losses using shifting experts with confidence., Mach Learn 108 (3) (2019) 425–444, doi:10.1007/s10994-018-5751-z.
[30] V. V'yugin, V. Trunov, Online learning with continuous ranked probability score, Proceedings of Machine Learning Research 105 (2019) 163–177.

Vladimir Vyugin Vladimir graduated from Moscow State University (MSU) with (Diploma in Mathematics) in 1971. In 1976 he obtained his Ph.D. degree (Physical and mathematical sciences: Mathematical Logic and Algorithms theory) at MSU. Thesis title: 'Structure of upper semilattices of computable numberings'. In 2002 he obtained his Dr.Sci. Degree at MSU. Thesis title: 'Applications of Kolmogorov's theory of algorithmic complexity to the logical foundations of probability theory'. Since 1975 Vladimir has been working on various topics including algorithmic complexity and randomness, computable numbering and online learning among others. In 1996 Vladimir joined IITP as a senior researcher and currently is the head of the Laboratory No.1 of Theory of information transmission and control. In 2018 Vladimir has joined the Skolkovo Institute of Science and Technology as a Senior Research Scientist.

Vladimir Trunov Vladimir Trunov obtained his MSc in Applied Physics and Mathematics from the Moscow Institute of Physics and Technology in 1970. He defended his Ph.D. thesis in Foundations of Computer Science at the Institute for Information Transmission Problem RAS (IITP RAS) in 1986, Vladimir stayed with the Institute as the senior researcher of the Laboratory No.1 of Theory of information transmission and control.