Online Learning with Continuous Ranked Probability Score

Vladimir V'yugin
Institute for Information Transmission Problems
Skolkovo Institute of Science and Technology
(Moscow, Russia)
e-mail vyugin@iitp.ru
Vladimir Trunov
Institute for Information Transmission Problems
(Moscow, Russia)
e-mail trunov@iitp.ru

Abstract

Probabilistic forecasts in the form of probability distributions over future events have become popular in several fields of statistical science. The dissimilarity between a probability forecast and an outcome is measured by a loss function (scoring rule). Popular example of scoring rule for continuous outcomes is the continuous ranked probability score (CRPS). We consider the case where several competing methods produce online predictions in the form of probability distribution functions. In this paper, the problem of combining probabilistic forecasts is considered in the prediction with expert advice framework. We show that CRPS is a mixable loss function and then the time independent upper bound for the regret of the Vovk's aggregating algorithm using CRPS as a loss function can be obtained. We present the results of numerical experiments illustrating the proposed methods.

1 Introduction

Probabilistic forecasts in the form of probability distributions over future events have become popular in several fields including meteorology, hydrology, economics, and demography (see discussion in Jordan et al. 2018). 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. 2018).

The dissimilarity between a probability forecast and an outcome is measured by a loss function (scoring rule). Popular examples of scoring rules for continuous outcomes include the logarithmic score and the continuous ranked probability score. The logarithmic score (Good 1952) is defined as $\mathrm{LogS}(F, y) = -\log(F(y))$, where F is a probability distribution function, is a proper scoring rule relative to the class of probability distributions with densities. The continuous ranked probability score (CRPS) is defined as

$$\mathrm{CRPS}(F, y) = \int (F(u) - 1_{u \geq y})^2 du,$$

where $F(u)$ is a probability distribution function, and y is an outcome – a real number.[1]

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

We solve this problem in the prediction with expert advice (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 1994, Freund and Schapire 1997, Vovk 1990, Vovk 1998, Cesa-Bianchi and Lugosi 2006 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 a form of probability distribution function computed using the experts probabilistic predictions.

The quality of the experts and of the learner predictions is measured by the continuous ranked probability score as a loss function. At each time step t any 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 a difference between the cumulated loss of the aggregating algorithm and the cumulated loss of the best expert.

There are a lot of papers on probabilistic predictions and on CRPS scoring rule (some of them are Brier 1950, Bröcker et al. 2007, Bröcker et al. 2008, Bröcker 2012, Jordan et al. 2018, Raftery et al. 2005). Most of them referred to the ensemble interpretation models. In particular, Bröcker

[1] Also, $1_{u \geq y} = 1$ if $u \geq y$ and it is 0 otherwise.

(2012) established a relation between the CRPS score and the quantile score with non-uniform levels.

In some cases, experts use for their predictions data models (probability distributions) which are defined explicitly in an analytic form. In this paper, we propose the rules for aggregation of such the data models. We present the formulas for direct calculation of the aggregated probability distribution function given probability distribution functions presented by the experts.

The proposed rules work both in the case of analytical models and in the case when empirical distribution functions (ensemble forecasts) are used.

Thorey et al. (2017) used the online exponentiated gradient method for aggregating probabilistic forecasts with the CRPS as a loss function. They pointed that in this case the theoretical guarantee (upper bound) for the regret is $O(\sqrt{T \ln N})$, where N is the number of the experts and T is the length of time interval.

In this paper we obtain a more tight upper bound of the regret for a special case when the outcomes and the probability distributions are located in a finite interval $[a, b]$ of real line. We show that the loss function $\mathrm{CRPS}(F, y)$ is mixable in sense of Vovk (1998) and apply the aggregating algorithm to obtain the time independent upper bound $\frac{b-a}{2} \ln N$ for the regret.[2]

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 theirs forecasts. After that, a current outcome is revealed by the nature. The validity of the forecasts of the experts and of the learner is measured using CRPS score and the Vovk (1998) aggregating algorithm and Adamskiy et al. (2017) method for aggregating vector valued forecasts. In Section 2 some details of these methods are presented.

In Section 3 we present a method for computing the aggregated probability distribution functions given the probability distribution functions presented by the experts.

We demonstrate the effectiveness of the proposed methods in Section 4, where the results of numerical experiments 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] The complete definitions are given in Section 2.

Aggregating algorithm. We consider the learning with a loss functions $\lambda(\gamma, \omega)$, where $\omega \in \{0, 1\}$ is an outcome, and $\gamma \in [0, 1]$ is a forecast. A set of experts $E = \{1, \ldots, N\}$ be given.

In the online setting, the following game is considered. At any round $t = 1, 2, \ldots$ each expert $i \in E$ presents a forecast $f_{i,t}$, then the learner presents its forecast f_t, after that, an outcome ω_t will be revealed. Each expert i suffers the loss $\lambda(f_{i,t}, \omega_t)$ and the learner suffers the loss $\lambda(f_t, \omega_t)$.

The Vovk's aggregating algorithm (Vovk 1990, Vovk 1998) is the base algorithm for computing the learner predictions. This algorithm assign weights $w_{i,t}$ for the experts using the weight update rule: $w_{i,1} = \frac{1}{N}$,

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

The normalized weights are defined

$$w^*_{i,t} = \frac{w_{i,t}}{\sum\limits_{j=1}^{N} w_{j,t}}.$$

The main tool is a superprediction function

$$g_t(\omega) = -\frac{1}{\eta} \ln \sum_{i=1}^{N} e^{-\eta \lambda(f_{i,t}, \omega)} w^*_{i,t}.$$

By Vovk (1998) a loss function is called η-mixable if for any probability distribution $\mathbf{w}^*_t = (w^*_{1,t}, \ldots, w^*_{N,t})$ on the set of experts and for any predictions $\mathbf{c}_t = (f_{1,t}, \ldots, f_{N,t})$ of the experts there exists a forecast f_t such that

$$\lambda(f_t, \omega) \leq g_t(\omega) \text{ for all } \omega. \tag{2}$$

We fix some rule of calculating f_t and write $f_t = \text{Subst}(\mathbf{c}_t, \mathbf{w}^*_t)$. Such a function is called substitution function.

For the η-mixable loss function $\lambda(\gamma, \omega)$, where $\omega \in \{0, 1\}$ is an outcome, and $\gamma \in [0, 1]$ is a forecast, the corresponding forecast can be defined as

$$f_t = \text{Subst}(g_t) = \frac{1}{2} - \frac{1}{2\eta} \ln \frac{\sum\limits_{i=1}^{N} w^*_{i,t} e^{-\eta \lambda(f_{i,t}, 0)}}{\sum\limits_{i=1}^{N} w^*_{i,t} e^{-\eta \lambda(f_{i,t}, 1)}}. \tag{3}$$

The square loss function $\lambda(\gamma, \omega) = (\omega - \gamma)^2$ is η-mixable for any η such that $0 < \eta \leq 2$.

Note that if a loss function $\lambda(\gamma, \omega)$ is η-mixable then the loss function $\Delta\lambda(\gamma, \omega)$ is $\frac{\eta}{\Delta}$-mixable for each $\Delta > 0$. We refer the reader for details to Vovk (1990), Vovk (1998), and Vovk (2001).

The mixability is a generalization of the notion of the exponential concavity. A loss function $\lambda(\gamma, \omega)$ is called η-exponential concave if for each ω the function $\exp(-\eta\lambda(\gamma, \omega))$ is concave by γ (see Cesa-Bianchi and Lugosi 2006). For such a function the inequality $\lambda(f_t, \omega) \leq g_t(\omega)$ also holds for all ω, where

$$f_t = \sum_{i=1}^{N} w_{i,t}^* f_{i,t}. \tag{4}$$

The square loss function is η-exponential concave for $0 < \eta < \frac{1}{2}$.

Regret analysis. Let $H_T = \sum\limits_{t=1}^{T} \lambda(f_t, \omega_t)$ be the cumulated loss of the learner and $L_T^i = \sum\limits_{t=1}^{T} \lambda(f_{i,t}, \omega_t)$ be the cumulated 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.

Let $W_t = \sum\limits_{i=1}^{N} w_{i,t}$. By definition $g_t(\omega_t) = -\frac{1}{\eta} \ln \frac{W_{t+1}}{W_t}$, where $W_1 = 1$. By the weight update rule (1) we obtain $w_{i,t+1} = \frac{1}{N} e^{-\eta L_t^i}$

Then by (2) and by telescoping, we obtain

$$H_T \leq \sum_{t=1}^{T} g_t(\omega_t) = -\frac{1}{\eta} \ln W_{T+1} \leq L_T^i + \frac{\ln N}{\eta} \tag{5}$$

for any expert i.

Vector-valued predictions. Consider a case where 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 $\mathbf{f}_{i,t} = (f_{i,t}^1, \ldots, f_{i,t}^d)$ and the learner presents a vector of forecasts $\mathbf{f}_t = (f_t^1, \ldots, f_t^d)$. After that, a vector $\omega_t = (\omega_t^1, \ldots, \omega_t^d)$ of outcomes will be revealed and the experts and the learner suffer losses.

A method for computing d-dimensional forecasts of the learner was presented by Adamskiy et al. (2017). Apply the aggregation rule to each coordinate separately: define $f_t^s = \mathrm{Subst}(\mathbf{c}_t^s, \mathbf{w}_t^*)$ for $1 \leq s \leq d$, where $\mathbf{c}_t^s = (f_{1,t}^s, \ldots, f_{N,t}^s)$ and $\mathbf{w}_t^* = (w_{i,1}^*, \ldots, w_{i,N}^*)$.

Rewrite the inequality $\lambda(f_t^s, \omega) \le g_t(\omega)$ as

$$e^{\eta\lambda(f_t^s,\omega)} \ge \sum_{i=1}^{N} e^{-\eta\lambda(f_{i,t}^s,\omega)} w_{i,t}^* \tag{6}$$

for $1 \le s \le d$. This inequality is valid for all ω.

When a sequence $\omega_t = (\omega_t^1, \ldots, \omega_t^d)$ of outcomes will be revealed, the experts and the learner suffer losses $\lambda(\omega_t, \mathbf{f}_{i,t}) = \sum_{s=1}^{d} \lambda(f_{i,t}^s, \omega_t^s)$ and $\lambda(\mathbf{f}_t, \omega_t) = \sum_{s=1}^{d} \lambda(f_t^s, \omega_t^s)$, where $\mathbf{f}_t = (f_t^1, \ldots, f_t^d)$ and $\mathbf{f}_{i,t} = (f_{i,t}^1, \ldots, f_{i,t}^d)$. We call this λ with vector valued arguments generalized loss function.

Multiplying the inequalities (6) for $s = 1, \ldots, d$, where $\omega = \omega_t^s$, we obtain

$$e^{-\eta \sum_{s=1}^{d} \lambda(f_t^s,\omega_t^s)} \ge \prod_{s=1}^{d} \sum_{i=1}^{N} e^{-\eta\lambda(f_{i,t}^s,\omega_t^s)} w_{i,t}^*. \tag{7}$$

The generalized Hölder inequality says that

$$\|F_1 F_2 \cdots F_d\|_r \le \|F_1\|_{q_1} \|F_2\|_{q_2} \cdots \|F_d\|_{q_d},$$

where $\frac{1}{q_1} + \ldots + \frac{1}{q_d} = \frac{1}{r}$, $q_s \in (0, +\infty)$ and $F_s \in L^{q_s}$ for $1 \le s \le d$. Let $q_s = 1$ for all $1 \le s \le d$, then $r = 1/d$. Let $F_{i,s} = e^{-\eta\lambda(f_{i,t}^s,\omega_t^s)}$ for $s = 1, \ldots, d$ and $\|F_s\|_1 = E_{i\sim\mathbf{w}^*}[F_{i,s}] = \sum_{i=1}^{N} F_{i,s} w_{i,t}^*$. Then

$$e^{-\eta\frac{1}{d}\sum_{s=1}^{d} \lambda(f_t^s,\omega_t^s)} \ge \sum_{i=1}^{N} e^{-\eta\frac{1}{d}\sum_{i=1}^{N} \lambda(f_{i,t}^s,\omega_t^s)} w_{i,t}^*.$$

or, equivalently,

$$e^{-\frac{\eta}{d}\lambda(\mathbf{f}_t,\omega_t)} \ge \sum_{i=1}^{N} e^{-\frac{\eta}{d}\lambda(\mathbf{f}_{i,t},\omega_t)} w_{i,t}^*. \tag{8}$$

The inequality (8) means that the generalized loss function $\lambda(\mathbf{f}_t, \omega_t)$ is $\frac{\eta}{d}$-mixable.

3 Aggregation of probability forecasts

Let the range of y be an interval $[a, b]$ of the real line. Suppose that the prediction for a label y is a probability distribution function $F : [a, b] \to [0, 1]$.[3]

[3] A probability distribution function is a non-decreasing function $F(y)$ defined on this interval such that $F(a) = 0$ and $F(b) = 1$.

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

$$\mathrm{CRPS}(F, y) = \int_a^b (F(u) - 1_{u \geq y})^2 du, \tag{9}$$

where 1 stands for the indicator function (Matheson and Winkler 1976).

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 also define the auxiliary representation of y – a binary variable $\omega_{u,y} = 1_{u \geq y}$.

We consider a game of prediction with expert advice, where the forecasts of the experts and of the learner are probability distribution functions. At any round (step) t of the game each expert $i \in \{1, \ldots, N\}$ presents its forecast – a probability distribution function $F_t^i(u)$ and the learner presents its forecast $F_t(u)$. After that, Nature presents an outcome $y \in [a, b]$ and the experts and the learner suffer losses $\mathrm{CRPS}(F_t^i, y)$ and $\mathrm{CRPS}(F_t, y)$. The goal of the learner is to predict such that its cumulated loss is less or equal to the loss of the best expert up to some regret.

We consider such a game as a "limit" of a sequence of games with the vector valued forecasts. To do this, we approximate any probability distribution functions $F(y)$ by the picewise-constant functions $L(y)$. Any such function L is defined by the points $z_0, z_1, z_2, \ldots, z_d$ and the values $f_0, f_1, f_2, \ldots, f_d$, where $a = z_0 < z_1 < z_2 < \ldots < z_d = b$ and $0 = f_0 < f_1 < f_2 < \ldots < f_d = 1$. By definition $L(y) = f_1$ for $z_0 \leq y \leq z_1$, $L(y) = f_2$ for $z_1 \leq y < z_2$, ..., $L(y) = f_d$ for $z_{d-1} \leq y \leq z_d$. Also, assume that $z_{i+1} - z_i = \Delta$ for all $0 \leq i < d$. By definition $\Delta = \frac{b-a}{d}$.

In this case the continuous ranked probability score is equal to

$$\mathrm{CRPS}(L, y) = \Delta \sum_{i=1}^{d} (f_i - \omega_{y,i})^2,$$

where $\omega_{y,i} = 1_{z_i \geq y} \in \{0, 1\}$ for $1 \leq i \leq d$. We identify the function L with the vector $\mathbf{f} = (f_1, f_2, \ldots, f_d)$.

Since the square loss function $\lambda(\gamma, \omega) = (\gamma - \omega)^2$ is 2-mixable, where $\omega \in \{0, 1\}$, by results of Section 2 the corresponding generalized loss function $\sum_{i=1}^{d} (f_i - \omega_i)^2$ is $\frac{2}{d}$-mixable and then the function

$$\lambda(\mathbf{f}, \omega) = \Delta \sum_{i=1}^{d} (f_i - \omega_i)^2$$

is $\frac{2}{d\Delta} = \frac{2}{b-a}$-mixable, where $\omega = (\omega_1, \ldots, \omega_d)$, $\omega_i \in \{0, 1\}$, and $\mathbf{f} = (f_1, \ldots, f_d)$. Also, $\mathrm{CRPS}(L, y) = \lambda(\mathbf{f}, \omega_y)$, where $\omega_y = (\omega_{y,1}, \ldots, \omega_{y,d})$.

Let the grid size Δ (and points $z_1, \dots z_d$) be fixed. Consider the following protocol for computing picewise-constant approximations L_t:

Protocol 1

Define $w_{i,1} = \frac{1}{N}$ for $1 \le i \le N$.

FOR $t = 1, \dots, T$

1. Receive the expert predictions $\mathbf{f}_{i,t} = (f^1_{i,t}, \dots f^d_{i,t})$ (values of a picewise constant function $L_{i,t}$ in the corresponding intervals), where $1 \le i \le N$.
2. Compute a picewise-constant function L_t defined by the forecasts $\mathbf{f}_t = (f^1_t, \dots, f^d_t)$ of the aggregating algorithm (the learner), where f^s_t is defined by (3), namely,
$$f^s_t = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^N w_{i,t} e^{-2(f^s_{i,t})^2}}{\sum_{i=1}^N w_{i,t} e^{-2(1-f^s_{i,t})^2}} \tag{10}$$
for $1 \le s \le d$.[4]
3. Observe the true outcome y_t and compute the score
$\mathrm{CRPS}(L^i_t, y_t) = \Delta \sum_{s=1}^d (f^s_t - \omega^s_t)^2$ for the experts $1 \le i \le N$
and the score
$\mathrm{CRPS}(L_t, y_t) = \Delta \sum_{s=1}^d (f^s_t - \omega^s_t)^2$ for the learner, where $\omega^s_t = 1_{z_s \ge y_t}$.
4. Update the weights of the experts $1 \le i \le N$
$$w_{i,t+1} = w_{i,t} e^{-\frac{2}{b-a} \mathrm{CRPS}(L^i_t, y_t)}.$$

ENDFOR

Using the analysis of Section 2, we obtain by (5)

$$\sum_{t=1}^T \mathrm{CRPS}(L_t, y_t) \le \sum_{t=1}^T \mathrm{CRPS}(L^i_t, y_t) + \frac{b-a}{2} \ln N \tag{11}$$

for any i.

Given t (and a, b), letting $\Delta \to 0$ (or, equivalently, $d \to \infty$) in (10), we obtain the expression for computing the learner forecast $F_t(u)$ given the forecasts $F^i_t(u)$ of the experts $1 \le i \le N$.

$$F_t(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^N w_{i,t} e^{-2(F^i_t(u))^2}}{\sum_{i=1}^N w_{i,t} e^{-2(1-F^i_t(u))^2}}, \tag{12}$$

where $w_{i,t+1} = w_{i,t} e^{-\frac{2}{b-a} \mathrm{CRPS}(F^i_t, y_t)}$ is computing recursively. Easy to verify that $F_t(u)$ is a probability distribution function.

[4]The same goes for the rule (4).

Theorem 1 *For any i and T,*

$$\sum_{t=1}^{T} \mathrm{CRPS}(F_t, y_t) \le \sum_{t=1}^{T} \mathrm{CRPS}(F_t^i, y_t) + \frac{b-a}{2} \ln N. \tag{13}$$

Proof. Since given i and T the inequality (11) holds for any grid size and the regret does not depend on Δ, letting $\Delta \to 0$, we obtain the similar inequality (13) for the limit quantities. $\triangle$

The square loss function is also η-exponential concave for $0 < \eta < \frac{1}{2}$. In this case (12) can be replaced with

$$F_t(u) = \sum_{i=1}^{N} w_{i,t}^* F_t^i(u), \tag{14}$$

where $w_{i,t}^* = \frac{w_{i,t}}{\sum\limits_{j=1}^{N} w_{j,t}}$ are normalized weights. The corresponding weights are computing recursively

$$w_{i,t+1} = w_{i,t} e^{-\frac{1}{2(b-a)} \mathrm{CRPS}(F_t^i, y_t)}. \tag{15}$$

In this case the bound (13) is replaced with

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

Let us present the protocol for computing the probability forecast of the learner given probability forecasts of the experts.

Protocol 2

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 $F_t^i(u)$, where $1 \le i \le N$.
2. Present the learner forecast – the probability distribution function $F_t(u)$:

$$F_t(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^{N} w_{i,t} e^{-2(F_t^i(u))^2}}{\sum_{i=1}^{N} w_{i,t} e^{-2(1-F_t^i(u))^2}}. \tag{16}$$

3. Observe the true outcome y_t and compute the score
   $\mathrm{CRPS}(F_t^i, y_t) = \int_a^b (F_t^i(u) - 1_{u \ge y_t})^2 du$ for the experts $1 \le i \le N$
   and the score
   $\mathrm{CRPS}(F_t, y_t) = \int_a^b (F_t(u) - 1_{u \ge y_t})^2 du$ for the learner.

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

$$w_{i,t+1} = w_{i,t} e^{-\frac{2}{b-a}\mathrm{CRPS}(F_t^i, y_t)} \tag{17}$$

ENDFOR

For exponential concave functions, the rules (16) and (17) be replaced with (14) and (15).

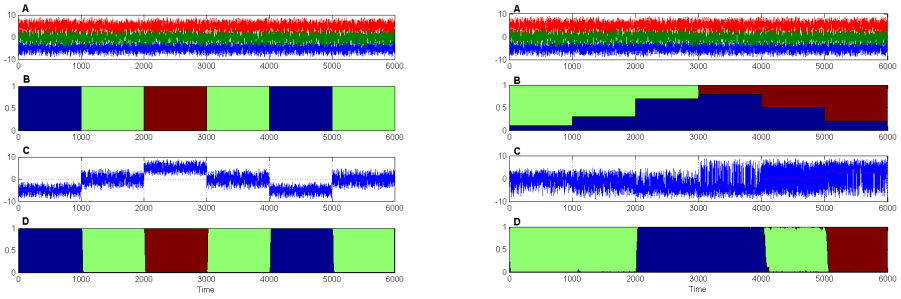

Figure 1: The stages of numerical experiments and the results of experts aggregation for two models (model 1 – left, model 2 - right). (A) – realizations of the trajectories for the three data generating distribtions; (B) – weights of the distributions assigned by the data generating model; (C) – sequence generated by the mixing model; (D) – weights of the experts assigned online by the aggregating algorithm (using the rule (17)).

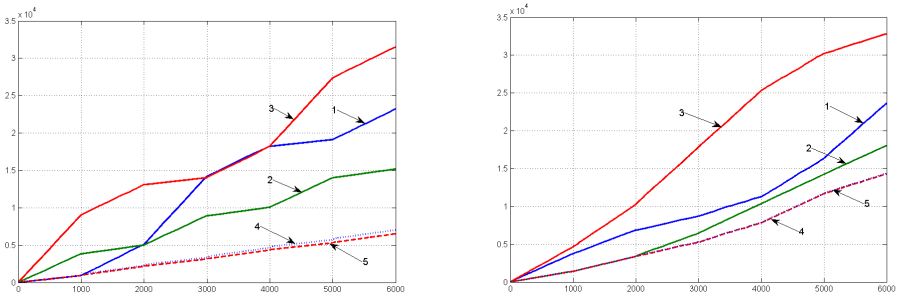

Figure 2: The cumulated losses of the experts (lines 1-3) and of the aggregating algorithm for both data generative models (Model 1 – left, Model 2 - right) and for both methods of computing forecasts: line 4 – for the rule (14) and line 5 – for the rule (16).

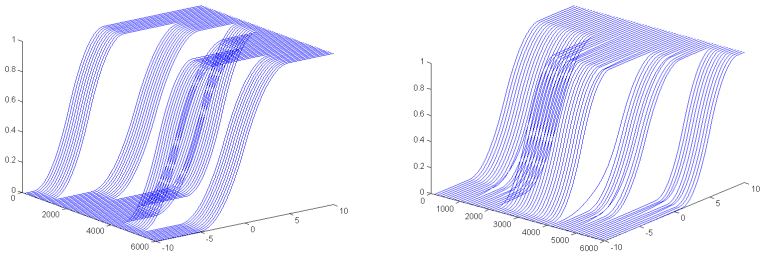

Figure 3: Empirical distribution functions obtained online as a result of aggregation of the distributions of three experts (at each 100th step) by the rule (16) for both data generative models.

4 Experiments

In this section we present the results of experiments which were performed on synthetic data. The initial data was obtained as a result of sampling from a mixture of three probability distributions with symmetric triangular densities, where the weights of the components of the mixture change with time. 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 and of the corresponding data generative models – Models 1 and 2.

There are three experts $i = 1, 2, 3$, each of which assumes that the time series under study is obtained as a result of sampling from the probability distribution with the fixed symmetric 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. In these experiments, we have used Fixed Share modification (see Herbster and Warmuth 1998) of Protocol 2, where we replace the rule (17) with the two-level scheme

$$w_{i,t}^{\mu} = \frac{w_{i,t} e^{-\frac{2}{b-a}\mathrm{CRPS}(F_t^i, y_t)}}{\sum\limits_{j=1}^{N} w_{j,t} e^{-\frac{2}{b-a}\mathrm{CRPS}(F_t^j, y_t)}},$$

$$w_{i,t+1} = \frac{\alpha}{N} + (1-\alpha) w_{i,t}^{\mu},$$

where $0 < \alpha < 1$. We set $\alpha = 0.001$ in our experiments.[5]

[5] In this case, using a suitable choice of the parameter α, we can obtain a bound

Figure 1 shows the main stages of the models formation (Model 1 – left, Model 2 - right) and the results of aggregation of the 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 relative weights that were used in mixing the probability distributions. Section C shows the result of sampling from the mixture model. The diagram of Section D shows the weights of the experts assigned by the corresponding Fixed Share algorithm in the online aggregating process.

Figure 2 shows the cumulated losses of the experts and the cumulated loss of the aggregating algorithm for both models (Model 1 – left, Model 2 - right) and for both methods of computing forecasts – by the rule (14) and by the rule (16).

Figure 3 shows in 3D format the empirical distribution functions obtained online by Protocol 2 for both data generative models.

5 Conclusion

In this paper, the problem of aggregating the probabilistic forecasts is considered. In this case, a popular example of proper scoring rule for continuous outcomes is the continuous ranked probability score CRPS.

We present the theoretical analysis of the continuous ranked probability score CRPS in the prediction with expert advice framework and illustrate these results with computer experiments.

We have proved that the CRPS loss function is mixable and exponential concave. Basing on these properties, we propose two methods for calculating the predictions of the Vovk (1998) aggregating algorithm. The time independent upper bounds for the regret of the aggregating algorithm were obtained for both methods.

The obvious disadvantage of these results is that they are valid only for the outcomes and distribution functions localized in finite intervals of the real line.

We present the results of numerical experiments based on the proposed methods and algorithms. These results show that two methods of computing forecasts lead to similar empirical cumulative losses while the rule (12) results in four times more regret bound than (14).

$O(\ln(TN))$ for the regret of the corresponding algorithm.

References

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

G.W. Brier. Verification of forecasts expressed in terms of probabilities. Mon. Weather Rev. 78. 1-3, 1950.

J. Bröcker, L.A. Smith. Scoring probabilistic forecasts: The importance of being proper. Weather and Forecasting 22. 382-388, 2007.

J. Bröcker, L.A. Smith. From ensemble forecasts to predictive distribution functions. Tellus A 60. 663-678, 2008.

J. Bröcker. Evaluating raw ensembles with the continuous ranked probability score. Q. J. R. Meteorol. Soc. 138, 1611-1617, July 2012 B

N. Cesa-Bianchi, G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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.

I.J. Good. Rational Decisions. Journal of the Royal Statistical Society B, 14(1), 107-114, 1952. `https://www.jstor.org/stable/2984087`.

M. Herbster, M. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

A. Jordan, F. Krüger, S. Lerch. Evaluating Probabilistic Forecasts with scoringRules, arXiv:1709.04743.

N. Littlestone, M. Warmuth. The weighted majority algorithm. *Information and Computation*. 108:212–261, 1994.

J.E. Matheson, R.L. Winkler. Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10), 1087-1096, 1976. doi:10.1287/mnsc.22.10.1087.

A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155-1174, 2005.

J. Thorey, V. Mallet and P. Baudin. Online learning with the Continuous Ranked Probability Score for ensemble forecasting. *Quarterly Journal of the Royal Meteorological Society*, 143, 521-529, January 2017 A DOI:10.1002/qj.2940

V. Vovk, Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, 371–383. San Mateo, CA, Morgan Kaufmann, 1990.

V. Vovk, A game of prediction with expert advice. *Journal of Computer and System Sciences*. 56(2), 153–173, 1998.

V. Vovk. Competitive on-line statistics. *International Statistical Review* 69, 213–248, 2001.

Vladimir Vovk, Jieli Shen, Valery Manokhin, Min-ge Xie. Nonparametric predictive distributions based on conformal prediction *Machine Learning*. ISSN: 0885-6125 (Print) 1573-0565 (Online)