Online Learning with Continuous Ranked Probability Score

Vladimir V. V'yugin VYUGIN@IITP.RU
Institute for Information Transmission Problems
Skolkovo Institute of Science and Technology
Moscow, Russia
Vladimir G. Trunov TRUNOV@IITP.RU
Institute for Information Transmission Problems
Moscow, Russia

Editor: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo and Evgueni Smirnov

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.

Keywords: On-line learning, Prediction with Expert Advice, Aggregating Algorithm, Probabilistic prediction, Continuous Ranked Probability Score, CRPS, Mixability

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. 2019).

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. Also, $1_{u \geq y} = 1$ if $u \geq y$ and it is 0 otherwise (see Epstein (1969)).

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 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's 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 the 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, Epstein (1969), Jordan et al. 2018, Raftery et al. 2005). Most of them referred to the ensemble interpretation models. In particular, Bröcker (2012) established a relation between the CRPS score and the quantile score with non-uniform levels.

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 the probability distributions functions. We present the formulas for direct calculation of the aggregated probability distribution function given probability distribution functions presented by the experts.

The proposed rules can be applied 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 ridge regression and online exponentiated gradient method for aggregating probabilistic forecasts with the CRPS as a loss function. They pointed that in this case the theoretical guarantees (upper bounds) for the regret are $O(\log T)$ for ridge regression and $O(\sqrt{T \ln N})$ for online exponentiated gradient descent, 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 Vovk's Aggregating Algorithm to obtain the time-independent upper bound $\frac{b-a}{2} \ln N$ for the regret.[1]

1. The complete definitions are given in Section 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 their 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. In Section 2 some details of these methods are presented.

In Section 3 we prove that the CRPS function is mixable and and then all machinery of the Vovk's Aggregating Algorithm can be applied. The proof is based on the method of prediction of packs by Adamskiy et al. (2017). We present a method for computing the aggregated probability distribution function given the probability distribution functions presented by the experts and prove a time-independent bound for the regret of the proposed algorithm.

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.

Online learning. Let a set of outcomes Ω and a set Γ of forecasts (decision space) be given.[2] 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 $E = \{1, \dots, N\}$. The following game of prediction with expert advice is considered. 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 $f_t \in \Gamma$, after that, an outcome $y_t \in \Omega$ will be revealed. Each expert i suffers the loss $\lambda(f_{i,t}, y_t)$ and the learner suffers the loss $\lambda(f_t, y_t)$, see 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 cumulated loss of the learner and $L_T^i = \sum_{t=1}^{T} \lambda(f_{i,t}, y_t)$ 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. The goal of the learner is to minimize the regret.

Aggregating Algorithm (AA). The Vovk's Aggregating algorithm (Vovk 1990, Vovk 1998) is the base algorithm for computing the learner predictions. This algorithm starting

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

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:

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

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

$$w^*_{i,t} = \frac{w_{i,t}}{\sum\limits_{j=1}^{N} w_{j,t}}. \tag{2}$$

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 $\mathbf{p} = (p_1, \ldots, p_N)$ on the set E of the experts: $\sum\limits_{i=1}^{N} p_i = 1$ and $p_i \geq 0$ for all i.

By Vovk (1998) a loss function is called η-mixable if for any probability distribution $\mathbf{p} = (p_1, \ldots, p_N)$ on the set E of experts and for any predictions $\mathbf{c} = (c_1, \ldots, c_N)$ of the experts there exists a forecast f such that

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

where

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

We fix some rule for calculating a forecast f and write $f = \text{Subst}(\mathbf{c}, \mathbf{p})$. 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 $c > 0$.

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

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

By telescoping, we obtain the time-independent bound

$$H_T \leq \sum_{t=1}^{T} g_t(y_t) = -\frac{1}{\eta} \ln W_{T+1} \leq L^i_T + \frac{\ln N}{\eta} \tag{6}$$

for any expert i. Therefore, there is a strategy for the learner that guarantees $R_T \le \frac{\ln N}{\eta}$ for all T.

Exponential concave loss functions. Assume that the set of all forecasts form a linear space. In this case, the mixability is a generalization of the notion of the exponential concavity. A loss function $\lambda(f, y)$ is called η-exponential concave if for each ω the function $\exp(-\eta\lambda(f, y))$ is concave by f for any y (see Cesa-Bianchi and Lugosi (2006)). For exponential concave loss function the inequality $\lambda(f, y) \le g(y)$ holds for all y by definition, where

$$f = \sum_{i=1}^{N} p_i c_i, \tag{7}$$

$p_1, \dots, p_N$ is a probability distribution on the set of experts, and $c_1, \dots, c_N$ are theirs forecasts.

For exponential 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 (6) for the regret.

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 $\omega \in \{0, 1\}$.[3] In this case, at any step t, the corresponding forecast f_t (in Protocol 1) can be defined as

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

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 (1990), Vovk (1998), and Vovk (2001).

The square loss function $\lambda(f, \omega) = (f - \omega)^2$, where $\omega \in \{0, 1\}$ and $f \in [0, 1]$, is η-exponential concave for any $0 < \eta \le \frac{1}{2}$ (see Cesa-Bianchi and Lugosi 2006).

Vector-valued predictions. Adamskiy et al. (2017) generalizes the AA for the case of d-dimensional forecasts, where d is a positive integer. Let an η-mixable loss function $\lambda(f, y)$ be given, where $\eta > 0$, $f \in \Gamma$ and $y \in \Omega$. Let $\mathbf{f} = (f^1, \dots, f^d) \in \Gamma^d$ be a d-dimensional forecast and $\mathbf{y} = (y^1, \dots, y^d) \in \Omega^d$ be a d-dimensional outcome. The generalized loss function is defined $\lambda(\mathbf{f}, \mathbf{y}) = \sum\limits_{s=1}^{d} \lambda(f^s, y^s)$; 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 $t = 1, 2, \dots$ each expert $i \in \{1, \dots, N\}$ presents a vector of forecasts $\mathbf{f}_{i,t} = (f_{i,t}^1, \dots, f_{i,t}^d)$ and the learner presents a vector of forecasts $\mathbf{f}_t = (f_t^1, \dots, f_t^d)$. After that, a vector $\mathbf{y}_t = (y_t^1, \dots, y_t^d)$ of outcomes will be revealed and the experts and the learner suffer losses $\lambda(\mathbf{f}_{i,t}, \mathbf{y}_t) = \sum\limits_{s=1}^{d} \lambda(f_{i,t}^s, \mathbf{y}_t^s)$ and $\lambda(\mathbf{f}_t, \mathbf{y}_t) = \sum\limits_{s=1}^{d} \lambda(f_t^s, y_t^s)$.

3. In what follows ω_t denotes a binary outcome.

Adamskiy et al. (2017) proved that the generalized game for AA is mixable. We rewrite this result for completeness of presentation.

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.*

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):

$$e^{-\eta\lambda(f^s,y} \ge \sum_{i=1}^N e^{-\eta\lambda(c_i^s,y)}p_i \text{ for all } y \tag{9}$$

for $1 \le s \le d$.

Let $\mathbf{y} = (y^1,\dots,y^d)$ be a vector of outcomes. Multiplying the inequalities (9) for $s = 1,\dots,d$ and $y = y^s$, we obtain

$$e^{-\eta\sum_{s=1}^d \lambda(f^s,y^s)} \ge \prod_{s=1}^d \sum_{i=1}^N e^{-\eta\lambda(c_i^s,y^s)}p_i \tag{10}$$

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

The generalized Hölder inequality says that

$$\|F_1F_2\cdots F_d\|_r \le \|F_1\|_{q_1}\|F_2\|_{q_2}\cdots\|F_d\|_{q_d},$$

where $\frac{1}{q_1}+\cdots+\frac{1}{q_d} = \frac{1}{r}$, $q_s \in (0,+\infty)$ and $F_s \in L^{q_s}$ for $1 \le s \le d$. Let $q_s = 1$ for all $1 \le s \le d$, then $r = 1/d$. Let $F_{i,s} = e^{-\eta\lambda(c_i^s,y^s)}$ for $s = 1,\dots,d$ and $\|F_s\|_1 = E_{i\sim\mathbf{p}}[F_{i,s}] = \sum_{i=1}^N F_{i,s}p_i$. Then

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

or, equivalently,

$$e^{-\frac{\eta}{d}\lambda(\mathbf{f},\mathbf{y})} \ge \sum_{i=1}^N e^{-\frac{\eta}{d}\lambda(\mathbf{c}_i,\mathbf{y})}p_i \tag{11}$$

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

The inequality (11) means that the generalized loss function $\lambda(\mathbf{f},\mathbf{y})$ is $\frac{\eta}{d}$-mixable. QED

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). The learner forecast $\mathbf{f}_t = (f_t^1,\dots,f_t^d)$ an any round t is defined: $f_t^s = \mathrm{Subst}(\mathbf{e}_t^s,\mathbf{w}_t^*)$ for each $s = 1,\dots,d$, where $\mathbf{e}_t^s = (f_{1,t}^s,\dots,f_{N,t}^s)$.

3. Aggregation of probability forecasts

Let in Protocol 1 the set of outcomes 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 $F : [a, b] \to [0, 1]$.[4]

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

$$\mathrm{CRPS}(F, y) = \int_a^b (F(u) - 1_{u \geq y})^2 du, \tag{12}$$

where 1 stands for the indicator function (Epstein (1969), Matheson and Winkler 1976 and so on).

The CRPS score measures the difference between the forecast F and a perfect forecast $1_{u \geq y}$ 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 probability distribution functions. At any step t of the game each expert $i \in \{1, \dots, N\}$ presents its forecast – a probability distribution function $F_t^i(u)$ and the learner presents its forecast $F_t(u)$.[5] After an outcome $y_t \in [a, b]$ be revealed and the experts and the learner suffer losses $\mathrm{CRPS}(F_t^i, y_t)$ 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 $F_t^i(u)$ for $1 \leq i \leq N$.
2. Present the learner's forecast – the probability distribution function $F_t(u)$:
3. Observe the true outcome y_t and compute the scores
   $\mathrm{CRPS}(F_t^i, y_t) = \int_a^b (F_t^i(u) - 1_{u \geq y_t})^2 du$ of the experts $1 \leq i \leq N$
   and the score
   $\mathrm{CRPS}(F_t, y_t) = \int_a^b (F_t(u) - 1_{u \geq y_t})^2 du$ of the learner.

ENDFOR

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

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

4. 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 left-continuous and has the right limit at each point.
5. 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.

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* $\mathbf{p} = (p_1, \dots, p_N)$ *on the set of all experts can be computed by the rule* [6]

$$F(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^N p_i e^{-2(F^i(u))^2}}{\sum_{i=1}^N p_i e^{-2(1-F^i(u))^2}}, \tag{13}$$

Proof. We approximate any probability distribution function $F(y)$ by the piecewise-constant functions $L_d(y)$, where $d = 1, 2, \dots$. Any such function L_d is defined by the points $z_0, z_1, z_2, \dots, z_d$ and the values $f_0 = F(z_0)$, $f_1 = F(z_1)$, $f_2 = F(z_2)$, $\dots$, $f_d = F(z_d)$, where $a = z_0 < z_1 < z_2 < \dots < z_d = b$ and $0 = f_0 \le f_1 \le f_2 < \dots \le f_d = 1$. By definition $L_d(y) = f_1$ for $z_0 \le y < z_1$, $L_d(y) = f_2$ for $z_1 \le y < z_2$, $\dots$, $L(y) = f_d$ for $z_{d-1} \le y < z_d$. Also, assume that $z_{i+1} - z_i = \Delta$ for all $0 \le i < d$. By definition $\Delta = \frac{b-a}{d}$.

We have

$$|\mathrm{CRPS}(F,y) - \mathrm{CRPS}(L_d,y)| \le$$
$$\int_a^y (L_d^2(u) - F^2(u))du + \int_y^b ((1-F^2(u))^2 - (1-L_d(u))^2)du \le 2\Delta \tag{14}$$

for any y, since each integral is bounded by Δ. Also, we take into account that by definition $F(u) \le L_d(u)$ for all u.

Define an auxiliary representation of y – a binary variable $\omega_y^s = 1_{z_s \ge y} \in \{0,1\}$ for $1 \le s \le d$ and $\boldsymbol{\omega}_y = (\omega_y^1, \dots, \omega_y^d)$.

Consider any $y \in [a,b]$. 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

$$\left| \mathrm{CRPS}(L_d,y) - \Delta \sum_{s=1}^d (f_s - \omega_y^s)^2 \right| \le 2\Delta \tag{15}$$

for all y. Let us study the generalized loss function

$$\lambda(\mathbf{f}, \boldsymbol{\omega}) = \Delta \sum_{s=1}^d (f_s - \omega^s)^2, \tag{16}$$

where $\mathbf{f} = (f_1, \dots, f_d)$, $\boldsymbol{\omega} = (\omega^1, \dots, \omega^d)$ 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 $\sum_{s=1}^d (f_s - \omega^s)^2$

6. Easy to verify that $F(u)$ is a probability distribution function.

is $\frac{2}{d}$-mixable and then the loss function (16) is $\frac{2}{d\Delta} = \frac{2}{b-a}$-mixable independently of that grid-size is used.

Let $F^i(u)$ be the probability distribution functions presented by the experts $1 \le i \le N$ and $\mathbf{f}_i = (f_i^1, \dots, f_i^d)$, where $f_i^s = F^i(z_s)$ for $1 \le s \le d$. By (11)

$$e^{-\frac{2}{(b-a)}\lambda(\mathbf{f},\boldsymbol{\omega})} \ge \sum_{i=1}^{N} e^{-\frac{2}{b-a}\lambda(\mathbf{f}_i,\boldsymbol{\omega})} p_i \tag{17}$$

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

$$f^s = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^{N} p_i e^{-2(f_i^s)^2}}{\sum_{i=1}^{N} p_i e^{-2(1-f_i^s)^2}} \tag{18}$$

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

By letting the grid-size $\Delta \to 0$ (or, equivalently, $d \to \infty$) in (15), (17), where $\boldsymbol{\omega} = \boldsymbol{\omega}_y$, and in (14), we obtain for any $y \in [a,b]$,

$$e^{-\frac{2}{(b-a)}\mathrm{CRPS}(F,y)} \ge \sum_{i=1}^{N} e^{-\frac{2}{b-a}\mathrm{CRPS}(F^i,y)} p_i, \tag{19}$$

where $F(u)$ is the limit form of (18) defined by

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

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

The inequality (19) means that the loss function $\mathrm{CRPS}(F,y)$ is $\frac{2}{b-a}$-mixable. QED

Let us specify the protocol 2 of the game with probabilistic predictions.

Protocol 3

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

FOR $t = 1, \dots, T$

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 $F_t(u)$:

$$F_t(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^{N} w_{i,t} e^{-2(F_t^i(u))^2}}{\sum_{i=1}^{N} w_{i,t} e^{-2(1-F_t^i(u))^2}}. \tag{20}$$

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 $1 \le i \le N$

$$w_{i,t+1} = w_{i,t} e^{-\frac{2}{b-a}\mathrm{CRPS}(F_t^i, y_t)} \tag{21}$$

ENDFOR

The performance bound of algorithm defined by Protocol 3 is presented in the following theorem.

Theorem 3 *For any i*

$$\sum_{t=1}^{T} \mathrm{CRPS}(F_t, y_t) \leq \sum_{t=1}^{T} \mathrm{CRPS}(F_t^i, y_t) + \frac{b-a}{2} \ln N \tag{22}$$

for each T.

Proof. The bound (22) is a direct corollary of the regret analysis of Section 2 and the bound (6). QED

The square loss function is also η-exponential concave for $0 < \eta \leq \frac{1}{2}$ (see Cesa-Bianchi and Lugosi (2006)). In this case (20) can be replaced with

$$F_t(u) = \sum_{i=1}^{N} w_{i,t}^* F_t^i(u), \tag{23}$$

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

$$w_{i,t+1} = w_{i,t} e^{-\frac{1}{2(b-a)} \mathrm{CRPS}(F_t^i, y_t)}. \tag{24}$$

Using results of Adamskiy et al. (2017) (presented in Section 2), we conclude that in this case the bound (22) can be replaced with

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

The proof is similar to the proof of Theorem 3.

4. Experiments

The proposed rules (20) and (21) can be used in the case 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.

The proposed methods are closely related to the so called ensemble forecasting (Thorey et al. (2017)). In practice, the output of physical process models usually not probabilities, but rather ensembles. Ensemble forecasts are based on a set of the expert's 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. The ensembles can be transformed into empirical probability distribution (or density) functions. Once the ensembles have been converted to probabilities, they are amenable to evaluation with probabilistic scoring rules like CRPS. (See discussion on evaluating ensembles in meteorology Bröcker (2012), Thorey et al. (2017)). When an uniform grid is used, instead of Protocol 3, you can use Protocol 3a given below.

The game with the vector-valued forecasts. We consider the game presented in Protocol 3 as a "limit" of a sequence of games with piecewise-constant distribution functions as the forecasts (vector-valued forecasts).

Protocol 3a

Define $w_{i,1} = \frac{1}{N}$ for $1 \le i \le N$ and and fix some grid-points $z_0, z_1, \dots, z_d$ in the interval $[a, b]$, $\Delta = z_s - z_{s-1}$ for all $1 \le s \le d$.

FOR $t = 1, \dots, T$

1. Receive the vectors $\mathbf{f}_{i,t} = (f^1_{i,t}, \dots f^d_{i,t})$ of the forecasts presented by the experts $1 \le i \le N$.
2. Compute the aggregated forecast $\mathbf{f}_t = (f^1_t, \dots, f^d_t)$ of the learner, where f^s_t is defined by (8), namely,

$$f^s_t = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^N w_{i,t} e^{-2(f^s_{i,t})^2}}{\sum_{i=1}^N w_{i,t} e^{-2(1-f^s_{i,t})^2}}$$

   for $1 \le s \le d$.
3. Observe the true outcome y_t and compute the losses $\lambda(\mathbf{f}_{i,t}, \boldsymbol{\omega}_{y_t}) = \Delta \sum_{s=1}^d (f^s_{i,t} - \omega^s_{y_t})^2$ of the experts $1 \le i \le N$, and the loss $\lambda(\mathbf{f}_t, \boldsymbol{\omega}_{y_t}) = \Delta \sum_{s=1}^d (f^s_t - \omega^s_{y_t})^2$ of the learner, where $\boldsymbol{\omega}_{y_t} = (\omega^1_{y_t}, \dots, \omega^d_{y_t})$ and $\omega^s_{y_t} = 1_{z_s \ge y_t}$.
4. Update the weights of the experts $1 \le i \le N$

$$w_{i,t+1} = w_{i,t} e^{-\frac{2}{b-a} \lambda(\mathbf{f}_{i,t}, \boldsymbol{\omega}_{y_t})}.$$

ENDFOR

Using the analysis of Section 2, we obtain by (6) time-independent and grid-independent bound for the regret

$$\sum_{t=1}^T \lambda(\mathbf{f}_t, \boldsymbol{\omega}_{y_t}) \le \sum_{t=1}^T \lambda(\mathbf{f}_{i,t}, \boldsymbol{\omega}_{y_t}) + \frac{b-a}{2} \ln N \tag{25}$$

for any i. Letting the grid-size d to infinity (or $\Delta \to 0$), we obtain the inequality (22) for the limit quantities.

Results of experiments. In this section we present the results of experiments which were performed on synthetic data. The initial data was 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. 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 Figure 1).

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 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 compare two rules of aggregations of the experts' forecasts: Vovk's AA (20) and the weighted average (23).

In these experiments, we have used Fixed Share modification (see Herbster and Warmuth 1998) of Protocol 3 and of its approximation – Protocol 3a, where we replace the rule (21) with the two-level scheme

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

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

where $0 < \alpha < 1$. We do the same for the rule (24). We set $\alpha = 0.001$ in our experiments.[7]

Figure 1 shows the main stages of data generating (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 relative weights that were used for mixing of the probability distributions. Section C shows the result of sampling from the mixture distribution. The diagram of Sections D and E show the weights of the experts assigned by the corresponding Fixed Share algorithm in the online aggregating process using rules (20) and (23).

Figure 2 shows the cumulated losses of the experts and the cumulated 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 (20) and by the rule (23). We note an advantage of rule (20) over the rule (23) in the case of data generating Method 1, in which there is a rapid change in leadership of the generating experts.

Figure 3 shows in 3D format the empirical distribution functions obtained online by Protocol 3 for both data generating methods and the rule (20).

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.

7. In this case, using a suitable choice of the parameter α, we can obtain a bound $O((k+1)\ln(TN))$ for the regret of the corresponding algorithm, where k is the number of swithing in the compound experts.

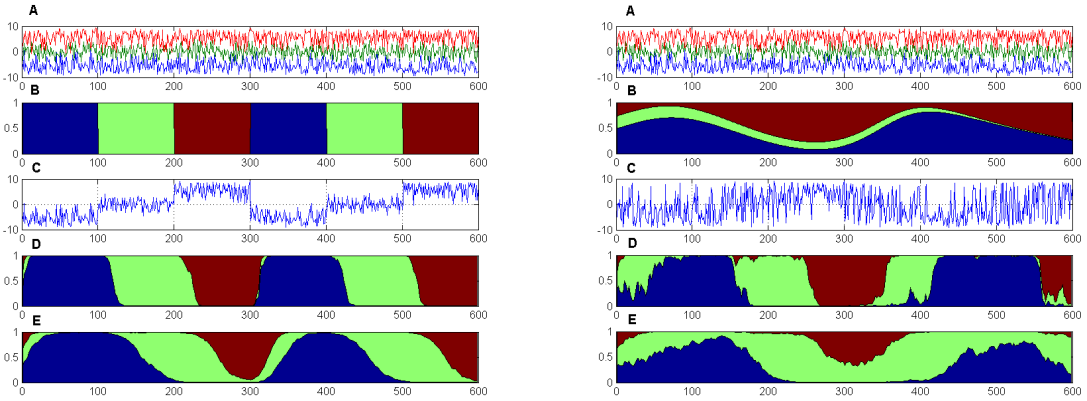

Figure 1: The stages of numerical experiments and the results of experts' aggregation for two data generation methods (Method 1 – left, Method 2 - right). (A) – realizations of the trajectories for the three data generating distributions; (B) – weights of the distributions assigned by the data generating method; (C) – sequence sampled from the distributions defined by Method 1 and Method 2; (D) – weights of the experts assigned online by the AA using the rule (21) and Fixed Share update; (E) – weights of the experts assigned online using the rule (24) and Fixed Share.

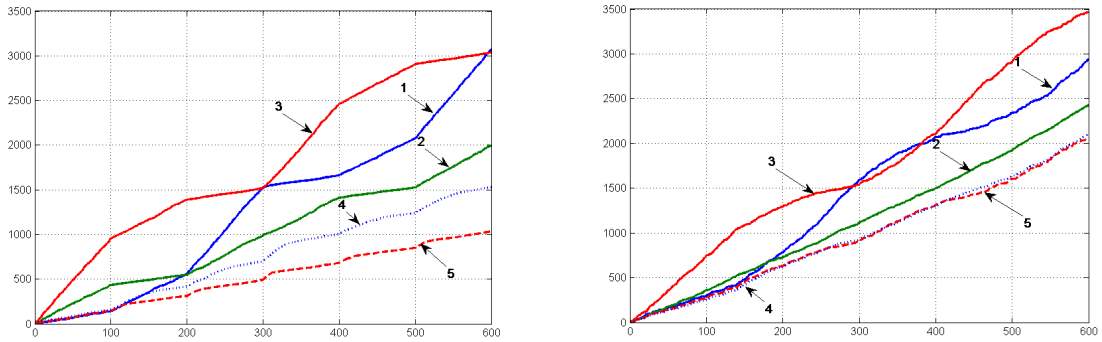

Figure 2: The cumulated losses of the experts (lines 1-3) and of the aggregating algorithm for both data generating methods (Method 1 – left, Method 2 - right) and for both methods of computing aggregated forecasts: line 4 – for the rule (23) and line 5 – for the rule (20). We note an advantage of rule (20) over rule (23) in the case of data generating Method 1, in which there is a rapid change in leadership of the data generating distributions.

We have proved that the CRPS loss function is mixable and and then all machinery of the Vovk's Aggregating Algorithm can be applied. The proof is an application of prediction of packs by Adamskiy et al. (2017): the probability distribution function can be approximated

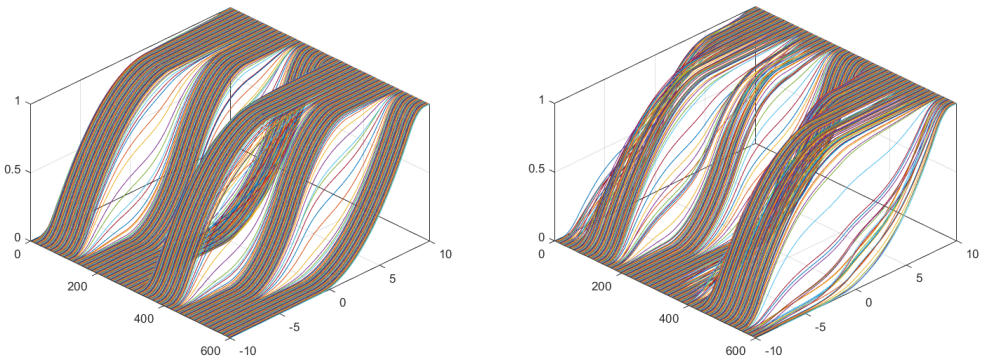

Figure 3: Empirical distribution functions obtained online as a result of aggregation of the distributions of three experts by the rule (20) for both data generating methods.

by a piecewise-constant function and further the method of aggregation of the generalized square loss function have been used.

Basing on mixability of CRPS, we propose two methods for calculating the predictions using the Vovk (1998) Aggregating Algorithm and simple mixture of the experts' forecasts. The time-independent upper bounds for the regret 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. It will be interesting to modify the algorithm and obtain regret bounds for unbounded outcomes. Also, does the learning rate $\frac{2}{b-a}$ is optimal is an open question.

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 (20) results in four times less regret bound than (23). We note a significantly best performance of method (20) over method (23) in the case where there is a rapid change in leadership of generating experts.

Acknowledgement

The authors are grateful to Vladimir Vovk and Yuri Kalnishkan for useful discussions that led to improving the presentation of the results.

References

D. Adamskiy, T. Bellotti, R. Dzhamtyrova, Y. Kalnishkan. Aggregating Algorithm for Prediction of Packs. *Machine Learning*, https://link.springer.com/article/10.1007/s10994-018-5769-2 (arXiv:1710.08114 [cs.LG]).

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.

E.S. Epstein. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol. Climatol.*, 8: 985–987, 1969.

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 scoring Rules, 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.

V. Vovk, J. Shen, V. Manokhin, Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108(3): 445-474, 2019. https://doi.org/10.1007/s10994-018-5755-8