Online aggregation of conformal predictive systems

Vladimir Trunov and Valdimir V'yugin

Institute for Information Transmission Problems, Moscow

COPA 2023
September 13-15, 2023
Limassol, Cyprus

logo

The goal of the work

- At each moment of time several competing conformal predictive systems (experts) give their predictions in the form of probability distribution functions.
- Probabilistic forecasts of the experts are combined by an aggregation algorithm into one probabilistic forecast at each step of the forecasting process, while expert forecasts can be used partially.
- The developed methods are used to solve the well-known problem of predicting the load of an electrical network online. Numerical experiments have shown the agreement of predictions with real data.

Problem setting

We consider methods for predicting the test labels y of objects x, where $x \in \mathscr{R}^k$, (in the simplest case $k = 1$) and $y \in \mathscr{R}$. It is assumed that "object-label" pairs (x, y) are generated by some probability source (distribution), moreover, the pairs (x, y) are independent and identically distributed (iid). A weaker hypothesis on data exchangeability can also be used as the main assumption. We refer to such an assumption as to main assumption or main hypothesis. The specific form of this probability distribution may be unknown to us and will not be used in what follows.

There are many methods for point, interval and probabilistic prediction. The first part of this work is related to improving the quality and reliability of known methods, based on a recently proposed non-parametric approach called conformal predicting systems.

The contribution

- Two methods for constructing online Mondrian conformal predictive systems are proposed and tested. In this case, we propose fuzzy Mondrian partition of the data.
- A method for online aggregating of conformal predictive systems in the prediction of expert advice framework using experts' competence levels is presented.
- The algorithm has been developed for obtaining probabilistic forecasts online based on the proposed methods.
- We generalize the aggregating algorithm for the case when expert predictions are provided with levels of competence. The concept of discounted regret is introduced, its upper bound is obtained.
- The algorithm has been developed for obtaining probabilistic forecasts online based on the proposed methods.

Split conformal prediction systems

The entire sample is divided into a training set $z_1^n = z_1, \ldots, z_n$ and a calibration set $\tilde{z}_1^m = \tilde{z}_1, \ldots, \tilde{z}_m$.
Based on the training sample $z_1^n = z_1, \ldots, z_n$, basic algorithm is built that can make point predictions $\hat{y} = f_{z_1^n}(x)$ for every x.
Based on the calibration sample, the conformity measure $A(z_1^n, (x, y))$ can be defined, and the conformity counters α_i of elements of the calibration sample and the counter α^y of an arbitrary test pair (x, y):

$$\alpha_i = A(z_1^n, \tilde{z}_i), \text{ for } i = 1, \ldots, m.$$
$$\alpha^y = A(z_1^n, (x, y)).$$

A game of prediction with expert advice

Ω – a set of outcomes, Γ – a set of of forecasts,
$E = \{1, \ldots, N\}$ – a set of the experts,
$\lambda(f, y)$ – a loss function, $f \in \Gamma$, $y \in \Omega$.

FOR $t = 1, \ldots, T$

1. Receive the experts' predictions $f_{i,t}$, where $1 \leq i \leq N$.
2. Learner presents a 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

Aggregation of probabilistic forecasts. Loss function CRPS.

Let the set of outcomes in Protocol 1 be the interval $\Omega = [a, b]$ of the real line, where $a < b$, and the set of predictions Γ be the set of all probability distribution functions on this interval: $F : [a, b] \to [0, 1]$. The continuous ranked probability score (CRPS loss function) is defined as

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

where $y \in [a, b]$ is an outcome and $H(x)$ is the Heaviside function: $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \geq 0$.
Consider a probability forecasting game with expert advice. At each step t, each expert $i \in \{1, \ldots, N\}$ presents a forecast which is a probability distribution function $F_{i,t}(u)$, Forecaster presents his prediction $F_t(u)$. After that, an outcome $y_t \in [a, b]$ is revealed and the experts and Forecaster suffer losses $\mathrm{CRPS}(F_{i,t}, y_t)$ and $\mathrm{CRPS}(F_t, y_t)$.

Regret

$H_T = \sum_{t=1}^{T} \lambda(f_t, y_t)$ – the accumulated loss of Forecaster,

$L_T^i = \sum_{t=1}^{T} \lambda(f_{i,t}, y_t)$ – the accumulated loss of an expert i.

$R_T^i = H_T - L_T^i$ – regret with respect to an expert i.

$R_T = H_T - \min_i L_T^i$ – regret with respect to the best expert.

The goal of the learner is to minimize the regret.

Forecaster's strategy

Assign weights $w_{i,t}$ for the experts $i \in E$:
weights update rule: $w_{i,1} = \frac{1}{N}$,

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

where $\eta > 0$ is a learning rate.
The normalized weights are defined $w^*_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}}$. "Mix" expert's prediction according to their weights. Simplest method is weighted average - WA: $f_t = \sum_{i=1}^{N} w^*_{i,t} f_{i,t}$.

Vovk's aggregating algorithm AA

AA computes a forecast of the learner using a more special way of mixing: The main tool of AA is a superprediction function

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

AA computes a forecast f_t such that
$\lambda(f_t, y) \leq c g_t(y)$ for all y, where $c \geq 1$ is a constant (small as possible).

Mixability: $c = 1$

Let $\mathbf{p} = (p_1, \ldots, p_N)$ be a probability distributions on the set E of the experts: $\sum_{i=1}^{N} p_i = 1$ and $p_i \geq 0$ for all i.

By Vovk a loss function is called η-mixable if for any probability distribution $\mathbf{p} = (p_1, \ldots, p_N)$ and for any predictions $\mathbf{f} = (f_1, \ldots, f_N)$ of the experts a forecast f exists such that

$$\lambda(f, y) \leq g(y) \text{ for all } y, \tag{2}$$

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

We fix some rule $f = \mathrm{Subst}(\mathbf{f}, \mathbf{p})$ for calculating the forecast f (substitution function).

A loss function is η-exponentially concave if (2) is valid for the weighted average $f = \sum_{i=1}^{N} p_i f_i$.

AA with competence levels

Protocol 2

FOR $t = 1, \ldots, T$

1. Get $f_{i,t}$ expert predictions and competence levels $p_{i,t}$, where $1 \leq i \leq N$.
2. Define Forecaster's prediction $f_t = \mathrm{Subst}(\mathbf{f}_t, \mathbf{w}_t^*))$, where $\mathbf{w}_t^* = (w_{1,t}^*, \ldots, w_{N,t}^*)$ are normalized weights defined by

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

3. Get the true value of the outcome y_t and calculate the loss $l_{i,t} = \lambda(f_{i,t}, y_t)$ of experts and the loss of Forecaster $\lambda(f_t, y_t)$.
4. Update the experts' weights:

$$w_{i,t+1} = w_{i,t} e^{-\eta(p_{i,t}\lambda(f_{i,t},y_t)+(1-p_{i,t})\lambda(f_t,y_t))}. \tag{3}$$

ENDFOR

Performance bound for AA

Theorem

For any $1 \leq i \leq N$ *the upper bound of the total discounted regret relative to any expert* i*:*

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

Game with probabilistic predictions

Protocol 2

FOR $t = 1, \ldots, T$

1. Receive the experts' predictions – the probability distribution functions: $F_t^1(u), \ldots, F_t^N(u)$.
2. Learner presents its forecast – the probability distribution function $F_t(u)$:
3. Observe the true outcome $y_t \in [a, b]$ 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

Mixability of CRPS

Theorem

1. $\mathrm{CRPS}(F, y)$ *is* $\frac{2}{b-a}$*-mixable loss function, where* $y \in [a, b]$.
2. *The learner's forecast* $F(u)$ *given the forecasts* $F^i(u)$ *of the experts* $1 \leq i \leq N$ *and a probability distribution* $\mathbf{p} = (p_1, \ldots, p_N)$ *on the set of all experts can be computed by the rule*

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

Game of probabilistic prediction for CRPS using AA

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

FOR $t = 1, \ldots, T$

1. Receive the expert predictions $F_t^i(u)$, where $1 \le i \le N$.
2. Learner presents the forecast $F_t(u)$:

$$F_t(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^N w_{i,t} e^{-2(F_t^i(u))^2}}{\sum_{i=1}^N w_{i,t} e^{-2(1-F_t^i(u))^2}}. \tag{5}$$

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

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

ENDFOR

Experts training.

The entire array of historical data, consisting of pairs (x_t, y_t), where x_t is the temperature, y_t is the network load at time t, is divided by intervals of time segments (season, time of day), that are the areas of competence of the respective experts. The data split used is essentially Mondrian categories.[1] In each area of competence, the data is divided into a training sample $z_1^n = z_1, \ldots, z_n$, and a calibration sample $\tilde{z}_1^m = \tilde{z}_1, \ldots, \tilde{z}_m$, where $z_i = (x_{t_i}, y_{t_i})$ for $1 \leq i \leq n$ and $\tilde{z}_i = (\tilde{x}_{t_i}, \tilde{y}_{t_i})$ for $1 \leq i \leq m$.
Each expert is trained on its own training set z_1^n.
We compare results with the Reference model for experts construction GMM – approximation of a two-dimensional point clouds using several Gaussian components.

[1] The Mondrian partition breaks the data into regions of homogeneity, within which the main hypothesis gets more evidence.

Calibration

Two methods for forming calibration sample are used. With the CP method, a part of the training sample is allocated, which serves as a calibration sample at all subsequent steps.[2] With the CP+ method, the initial calibration sample is replenished at each step with new parts from the area of competence observed at time t by the expert.[3]

---

[2] At the same time, the main assumption is preserved that the elements of the calibration sample and the test value must be independently and identically distributed with respect to some probability distribution on pairs (x, y) corresponding to the the expert's area of competence. The specific form of this distribution is not taken into account.

[3] This method makes it possible to take into account possible local violations of the basic assumption.

Fuzzy Mondrian partitions.

When aggregating the predictive distributions of the experts, we somewhat expand the concept of Mondrian partitioning – by specifying a partition using real values p_i, we define fuzzy sets in which conformal predictive systems are applied and aggregated. The conducted comparative experiments (below) show that in this way we achieve more accurate results in forecasting.

Algorithm 3

FOR $i = 1, \ldots N$ **Preprocessing loop*
Using the ith training sample z_1^n we build a regression rule (algorithm) $y = f_i(x)$.
ENDFOR **End of preprocessing loop*

Define $w_{i,1} = \frac{1}{N}$ for $1 \leq i \leq N$.
FOR $t = 1, \ldots, T$ **Main Loop*
We get the testing object x_t and define the probabilistic forecasts of experts – probability distribution functions $F_{i,t}(y)$ for $i = 1, \ldots N$.
FOR $i = 1, \ldots N$ **Construction of the experts' conformal distributions.*

Construction of the predictive conformal probability distribution function

:

Let us fix the calibration sample $\tilde{z}_1^m = \tilde{z}_1, \ldots, \tilde{z}_m$ from the area of competence of the corresponding expert i, $\tilde{z}_s = (\tilde{x}_s, \tilde{y}_s)$ for $1 \leq s \leq m$.

We use the conformity measure $A(z_1^n, (x, y)) = y - \hat{y}$, where $\hat{y} = f_i(x)$ is the label prediction computed by the regression algorithm.

Calculate the conformity counters α_s for $s = 1, \ldots, m$: $\alpha_s = A(z_1^n, (\tilde{x}_s, \tilde{y}_s))$ and arrange them in ascending order:

$$\alpha_{(1)} < \cdots < \alpha_{(k)}.$$

Let $n_j = |\{s : \alpha_s = \alpha_{(j)}\}|$ for $j = 1, \ldots, k$.

Define also $m_j = \sup\{y : \alpha^y < \alpha_{(j)}\}$ and $M_j = \inf\{y : \alpha^y > \alpha_{(j)}\}$, where $\alpha^y = A(z_1^n, (x, y))$.

Definition of the predictive conformal probability distribution function

:

$$Q_{z_1^n,\tilde{z}_1^m,x_t,\tau}(y) = \begin{cases} \frac{\tau}{m+1} \text{ if } y < m_1, \\ \frac{n_1+\cdots+n_{j-1}+\tau n_j+\tau}{m+1} \text{ if } m_j < y < M_j, j = 1,\ldots k, \\ \frac{n_1+\cdots+n_j+\tau}{m+1} \text{ if } M_j < y < m_{j+1}, j = 1,\ldots k-1, \\ \frac{n_1+\cdots+n_k+\tau}{m+1} = \frac{m+\tau}{m+1} \text{ if } y > M_k. \end{cases}$$

Denote $F_{i,t}(y) = Q_{z_1^n,\tilde{z}_1^m,x_t,\tau}(y)$ the conformal probability distribution function of the expert i.
ENDFOR **End of loop for constructing the experts' conformal distributions*

Aggregation of the experts' probability distribution functions.

Get the competence levels $p_{i,t}$ of the experts, $1 \leq i \leq N$. Define the probability distribution function of Forecaster by the rule

$$F_t(y) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^{N} w_{i,t}^{p} e^{-2(F_{i,t}(y))^2}}{\sum_{i=1}^{N} w_{i,t}^{p} e^{-2(1-F_{i,t}(y))^2}} \tag{6}$$

for AA algorithm, or by the rule

$$F_t(y) = \sum_{i=1}^{N} w_{i,t}^{p} F_{i,t}(y) \tag{7}$$

for WA algorithm, where

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

Observe outcome y_t and compute losses $\mathrm{CRPS}(F_{i,t}, y_t)$ of the experts $1 \leq i \leq N$, as well as the loss $\mathrm{CRPS}(F_t, y_t)$ of Forecaster.

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

$$w_{i,t+1} = w_{i,t} e^{-\eta(p_{i,t}\mathrm{CRPS}(F_{i,t},y_t)+(1-p_{i,t})\mathrm{CRPS}(F_t,y_t))}, \tag{8}$$

where $\eta = \frac{2}{b-a}$ for AA and $\eta = \frac{1}{2(b-a)}$ for WA.
ENDFOR **End of the main loop*

Probabilistic forecasting of hourly electrical loads

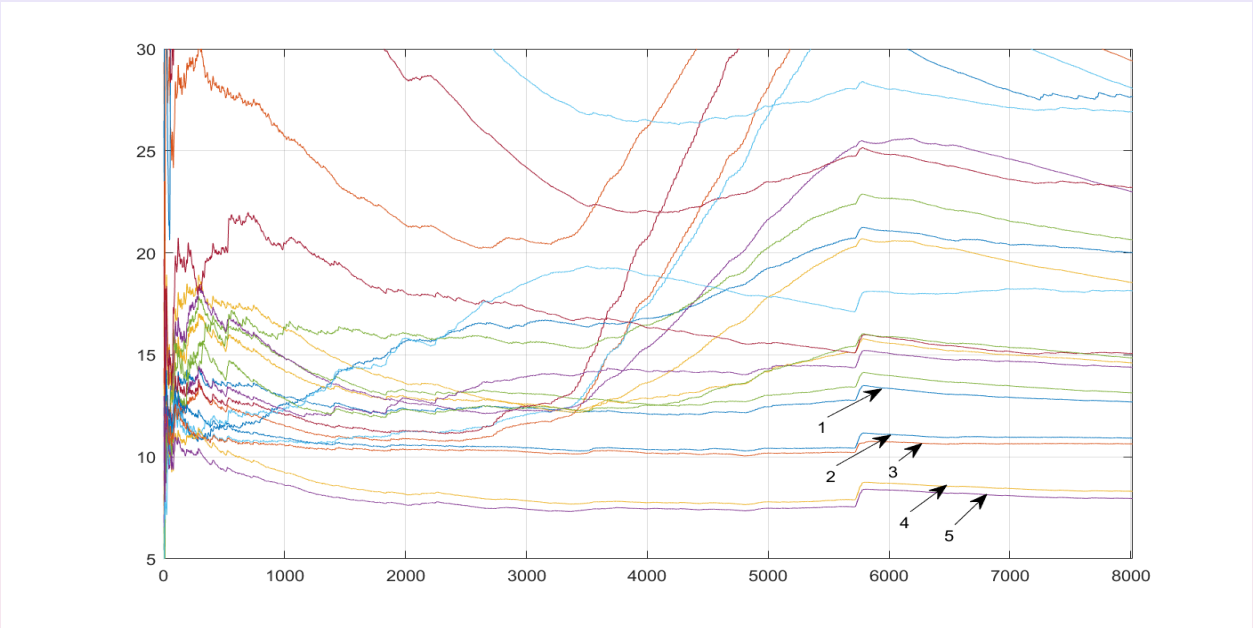

Figure: Average values of accumulated CRPS losses of experts, constructed by the CP method, and the loss of aggregators: 2-WA and 3-AA of these experts, used without taking into account their levels of competence For comparison, the same figure shows the losses of aggregators 4-WA and 5-AA, taking into account the levels of expert's competence. 1–The loss of the AnyTime expert.

logo

Probabilistic forecasting of hourly electrical loads

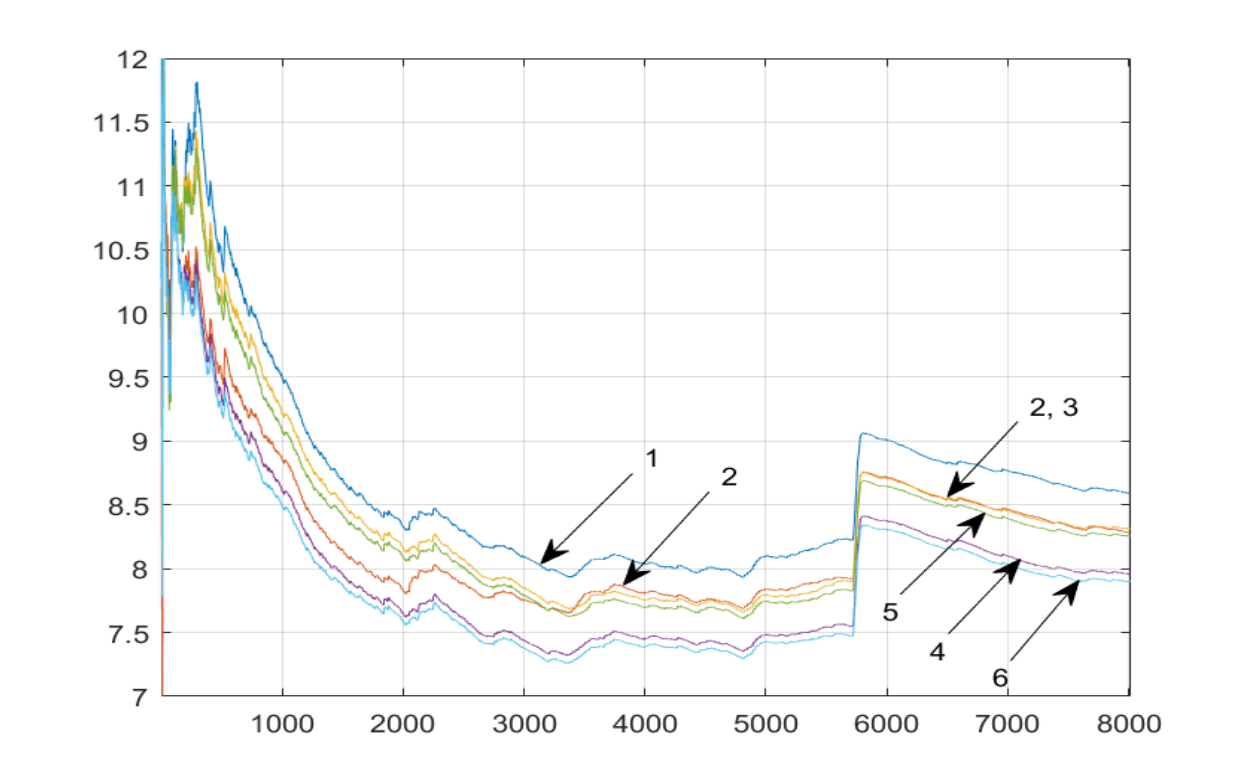

Figure: Average values of accumulated CRPS losses of aggregators AA and WA for experts constructed by GMM, CP and CP+ methods: 1-WA((GMM),2-AA(GMM), 3-WA(CP), 4-AA(CP), 5-WA(CP+), 6-AA(CP+) aggregators used with experts' levels of competence.

logo

Discounted Regret curves of the AA Algorithm for all experts build by the CP Method

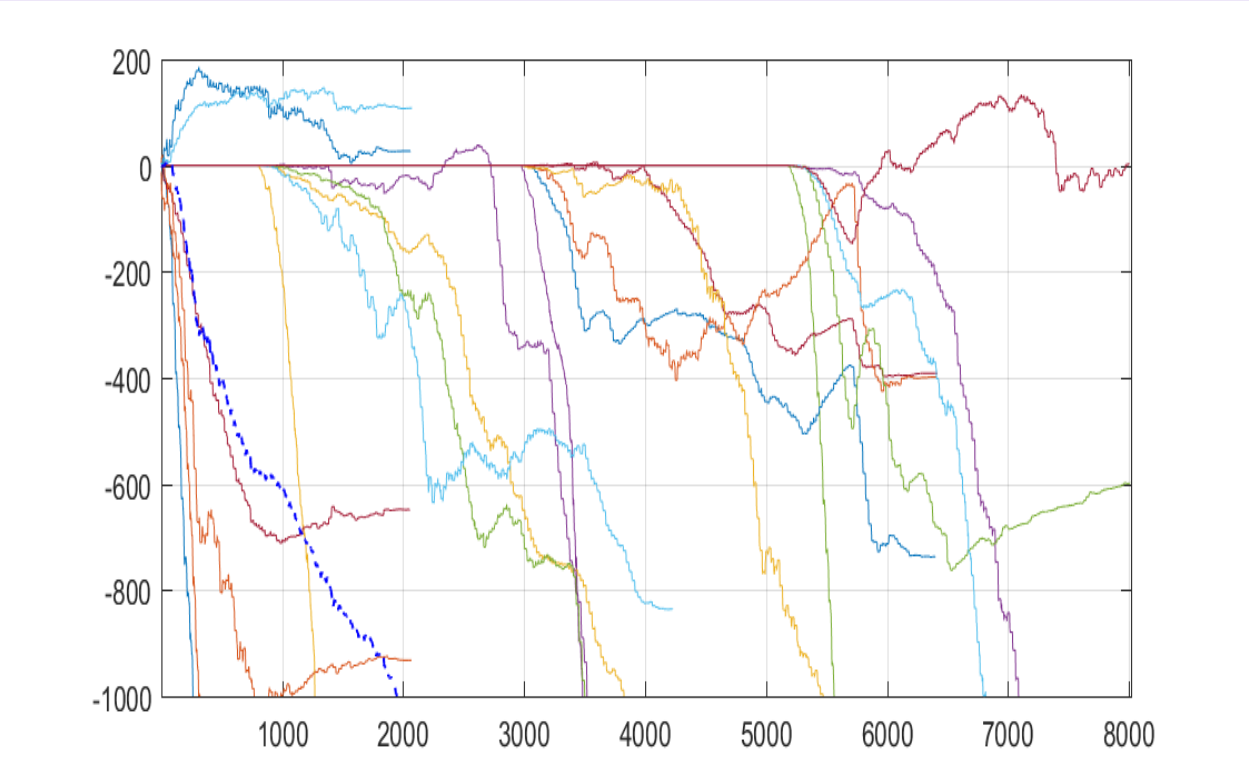

Figure: Discounted regrets of the AA algorithm with respect to Experts constructed by the CP method.

logo

Densities of distributions built by the AA algorithm.

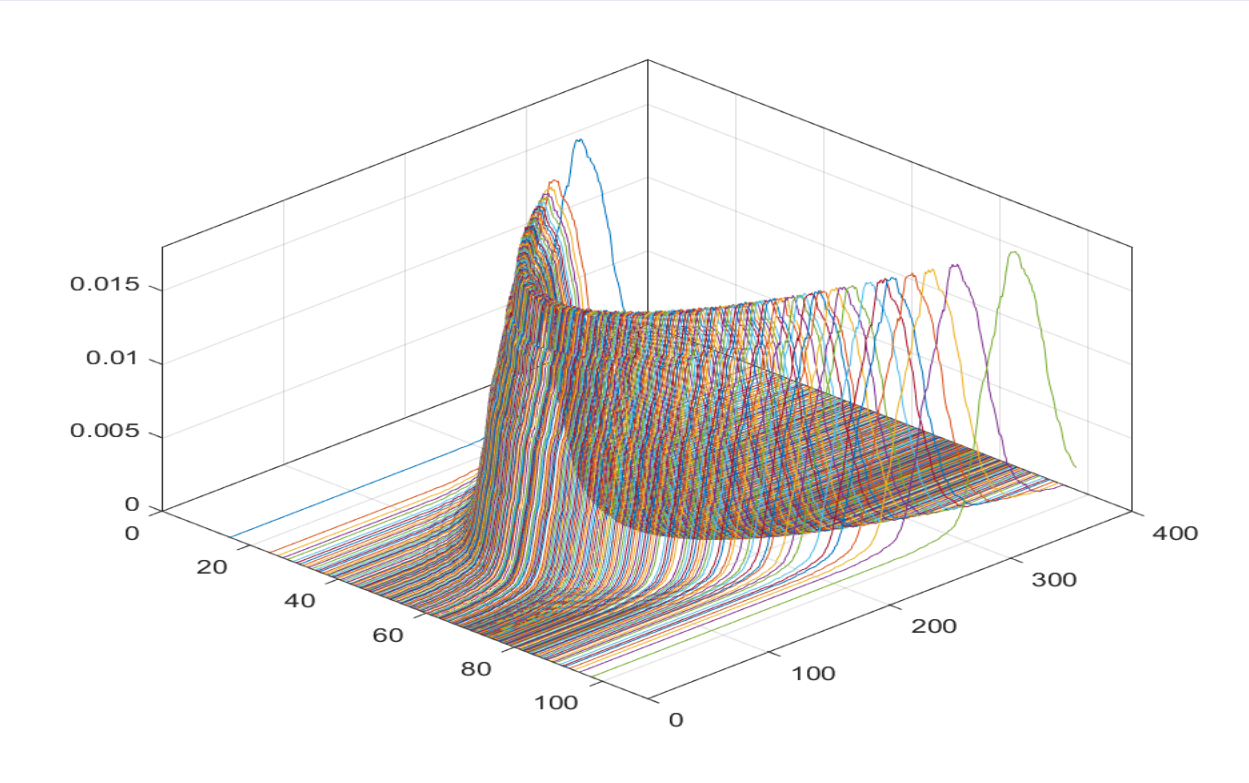

Figure: Densities of distributions built by the AA algorithm when aggregating Experts defined by the CP method.

logo