Influence of hyperparameters on online aggregation with countable experts

Sergey Kunin-Bogoiavlenskii

Moscow Institute of Physics and Technology

Course: My first scientific paper
(Strijov's practice)/Group 125

Expert: R. D. Zukhba

Consultant: A. V. Zukhba

2024

Goal of research

> Prediction is very difficult,
> especially if it's about the future.
>
> *Niels Bohr*

Goal

Examining the influence of hyperparameters on the performance of the aggregation algorithm with a countable number of experts

Targets

1. Time series generator implementation
2. Aggregating algorithm implementation
3. Experiments with various hyperparameters

Literature

- V. V'yugin, V. Trunov. 2023. Prediction of Locally Stationary Data Using Prediction with Expert Advice. `http://www.jip.ru/2023/470-487-2023.pdf`
- O. Bousquet, M. Warmuth. 2002. `https://www.jmlr.org/papers/volume3/bousquet02b/bousquet02b.pdf`
- N. Cesa-Bianchi, G. Lugosi. 2006. Prediction, Learning, and Games. `https://ii.uni.wroc.pl/~lukstafi/pmwiki/uploads/AGT/Prediction_Learning_and_Games.pdf`.
- Hyndman, R. J. & Athanasopoulos, G., 2nd edition. 2018. Forecasting: Principles and Practice. `https://otexts.com/fpp2/`

Problem statement

> There are two kinds of forecasters: those who don't know, and those who don't know they don't know.
>
> *John Kenneth Galbraith*

Data

It is assumed that there are multiple generators, whose structure is unknown to the predictors. The time series is obtained by merging segments, each produced by one of the generators. These segments are called areas of stationarity, and can be studied using machine learning methods.

Gerators implemented:

- Linear
- ARMA

Problem statement

Terms

- X — signals space
- Y — responses space
- $\mathcal{N}$ — set of experts, indexed by natural numbers
- D — desicion space, to which predictions belong
- $\lambda : \mathrm{D} \times \mathrm{Y} \to \mathbb{R}_+$ — nonnegative loss function
- $L^i_T = \sum\limits_{t=1}^{T} l^i_t$ — cumulative loss of expert i during the first T steps
- $H_T = \sum\limits_{t=1}^{T} h_t$ — master's cumulative loss during the first T steps
- $R_T = H_T - L_T$ — master's regret relative to the best partition, where L_T is the cumulative loss of the best partition.

Problem statement

Algorithm

```
FOR $t = 1, 2, \ldots :$
  1. Expert $f^t$ initialization
  2. Experts' predictions $f_t^i = f_t^i(x_t),\ 1 \leq i \leq t$
  3. Master's prediction evaluation $\gamma_t = \mathrm{Subst}(\mathbf{f_t}, \widehat{\mathbf{w}}_\mathbf{t})$
  4. Computation of master's loss $h_t = \lambda(p_t, y_t)$ and experts' losses $l_t^i$
  5. **Loss Update** weights modification
  6. **Mixing Update** weights modification
ENDFOR
```

Experiments

Metric — R_T, the regret

Initialization weights

Default weights: $w_1^i = \frac{1}{(i+1)\ln^2(i+1)}$

Experimental: $\frac{1}{i^\alpha}$, $\frac{1}{c}$, $\frac{1}{(i+4)\ln(i+4)\ln^2\ln(i+4)}$, etc.

Noise

Different noise variance leads to diverse ability of experts to train, which opens curious quialities of the master algorithm

Window size

As the algorithm does not know the locations of generator switches, finding an optimal training window is also a challenge.

Experiments

Mixing update scheme

$\widetilde{w}^i_{t+1} = \sum_{q=1}^{t} \beta_t(q) \widetilde{w}^i_q$

- Start Vector Share - default scheme in GMPP
- Uniform Past Share
- Decaying Past Share
- Increasing Past Share - new proposed scheme:

$$\beta_t(q) = \begin{cases} \alpha_t (t-q)^\gamma \frac{1}{Z_t}, & 1 \leq q < t \\ 1 - \alpha_t, & q = t \end{cases}$$

,with $Z_t = \sum_{q=1}^{t-1} (t-q)^\gamma, \gamma > 0.$

Mixing update coefficients

Default coefficient: $\alpha_t = \frac{1}{t+1}$

Experimental: $\frac{1}{(t+1)^\beta}$, $\frac{1}{c}$, $\frac{1}{(t+c)}$, $\frac{1}{e^{t/3}}$, etc.

Loss plot

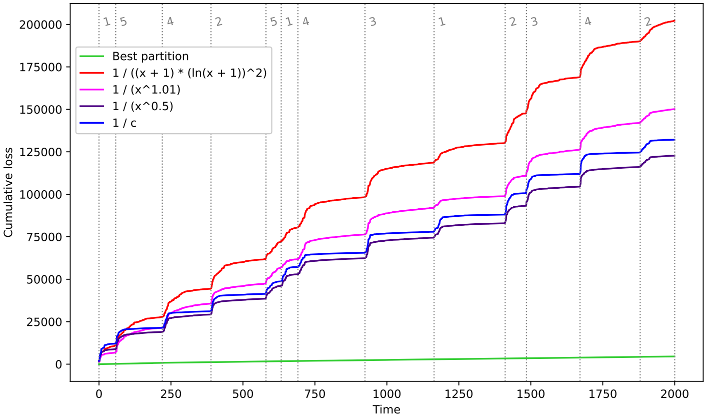

Figure: Total loss for different weight functions (alpha function is defaut, $\sigma^2 = 1$, window size = 10)

Table

Table: Regret with different Mixing Update schemes and Noise Variance (alpha function is defaut, weight function is $1/x^{1.01}$, window size = 10)

	Mixing scheme		
Noise variance	increasing past	start	uniform past
0.10	114564.74	131578.01	115927.25
1	110438.09	132268.30	110569.83
2	105398.29	136043.26	103136.06
5	92343.75	144554.62	83630.45
6	89032.63	146382.86	25178.60
7	29417.82	144827.02	-9208.74
8	-13307.43	147510.57	-55905.83
10	-120166.34	90089.56	-377184.32
12	-1123130.94	-420588.94	-1354731.42
15	-1508510.73	-1205600.59	-1862352.43

Loss plot

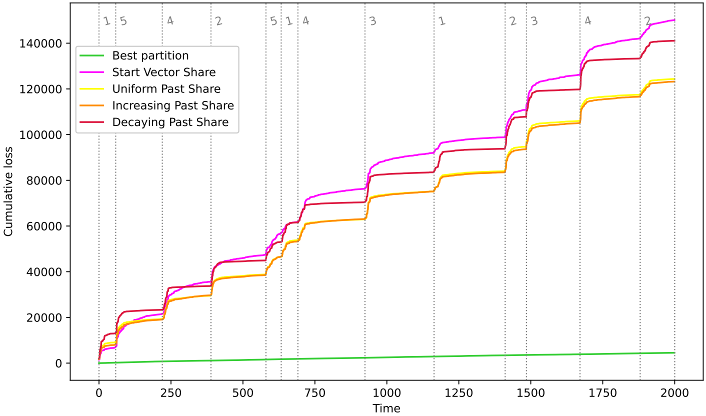

Figure:

Total loss with different mixing schemes
(default alpha function, weight function is $1/x^{1.01}$, window size $= 10$)

Table

Table:
Regret with different Alpha and Weight Functions in Start Vector Share Update scheme
($\sigma^2 = 1$, window size $= 10$)

	Weight function		
Alpha function	$\frac{1}{(x+1)\ln^2(x+1)}$	$\frac{1}{x^{1.01}}$	$\frac{1}{x^{0.5}}$
$1/(t+1)$	175594.64	132268.30	**108630.68**
$1/(t+1)^{0.5}$	245494.10	207099.32	164079.31
$1/(t+1)^{1.5}$	130339.21	125699.71	130185.66
$1/(t+1)^2$	136029.09	134929.54	133876.06
$1/e^{t/3}$	136413.87	135409.06	134012.15
$1/(t+10)$	175411.17	132208.09	**108638.55**
$1/(t+100)$	173760.98	131623.56	**108732.56**
$1/(t+1000)$	162129.20	127165.47	110389.16

Conclusion

Summary

- Generators and algorithm implemented
- Correctness of the algorithm veryfied
- A series of experiments conducted
- Enhanced weight functions achieved
- New Mixing Update Scheme proposed

Further plans

- Understand what is going on with different length of time series
- Run experiments on real data
- Theoretically prove that the obtained function is the best.

The End