

On the Value of Stochastic Side Information in Online Learning

Junzhang Jia, Xueting Wu, Jamie Evans, Jingge Zhu

The University of Melbourne



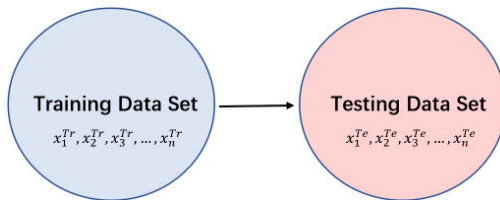
International Conference on Acoustics, Speech, and Signal Processing
June, 2023

Table of contents

- Introduction
- Main Results
- Examples
- Future work

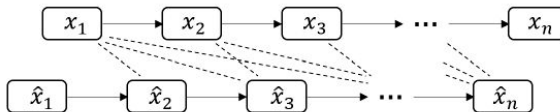
Online Learning

Traditional Learning



Universal Prediction(Online Learning)

Target Sequence



Prediction Sequence

Problem Setup

Given:

- An unknown target sequence $\mathcal{T}_n = (X_1^T, X_2^T, \dots, X_n^T)$.
- Decision space $\mathcal{D}$, a convex and nonempty subset of $\mathbb{R}$.
- Experts advice $\mathcal{F}^\theta = (f_1^\theta, f_2^\theta, \dots, f_n^\theta) \in \mathcal{D}^n$, $\theta \in \{1, 2, \dots, N\}$.
- Sequential side information $\mathcal{S}_n = (X_1^S, X_2^S, \dots, X_n^S)$.
- Non-negative loss: $\ell : \mathcal{D} \times \mathcal{X} \mapsto \mathbb{R}^+$.

Output:

- At each time t , the forecaster predicts the current target instance X_t^T with all previous observations $\mathcal{T}_{t-1}$ and the corresponding SSI $\mathcal{S}_t$ up to time t .
- The forecaster output sequence $\Xi_n := (p_1, p_2, \dots, p_n) \in \mathcal{D}^n$.

Minimax Regret

Definition (Minimax Expected Regret)

$$R := \inf_{\Xi_n} \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{L(\Xi_n(\mathcal{S}_n, \mathcal{T}_n), \mathcal{T}_n) - \min_{\theta} L(\mathcal{F}^{\theta}, \mathcal{T}_n)\}.$$

Ξ_n - The forecaster (algorithm) prediction sequence

$\mathcal{T}_n$ - The target sequence

$\mathcal{F}^{\theta}$ - The expert prediction sequence

$\mathcal{S}_n$ - The sequential side information

$$L(A_n, B_n) = \sum_{t=1}^n \ell(a_t, b_t)$$

for any $A_n := (a_1, a_2, \dots, a_n)$ and $B_n := (b_1, b_2, \dots, b_n)$

Algorithm

Algorithm 1 Exp3 with SSI

- 1: Initialize the weights for the SSI w_1^S and all experts w_1^θ to be 1;
 - 2: **for** $t = 1$ to n **do**
 - 3: Receive X_t^S ;
 - 4: $p_t = \frac{w_t^S \hat{X}_t^T(X_t^S) + \sum_{\theta=1}^N w_t^\theta f_t^\theta}{w_t^S + \sum_{\theta=1}^N w_t^\theta}$;
 - 5: Receive X_t^T ;
 - 6: The experts suffer the loss $\ell(f_t^\theta, X_t^T)$;
 - 7: The SSI suffers the loss $\ell(\hat{X}_t^T(X_t^S), X_t^T)$;
 - 8: $w_{t+1}^S = w_t^S e^{-\eta \ell(\hat{T}_n(\mathcal{S}_n), X_t^T)}$;
 - 9: **for** $\theta = 1$ to N **do**
 - 10: $w_{t+1}^\theta = w_t^\theta e^{-\eta \ell(f_t^\theta, X_t^T)}$
 - 11: **end for**
 - 12: **end for**
 - 13: **return** $\Xi_n = (p_1, p_2, \dots, p_n)$
-

$$\hat{X}_t^T(X_t^S) = \underset{X_t^T}{\operatorname{argmax}} P(X_t^S | X_t^T)$$

$$\hat{T}_n(\mathcal{S}_n) := (\hat{X}_t^T(X_1^S), \hat{X}_t^T(X_2^S), \dots, \hat{X}_t^T(X_n^S))$$

Main Results

Theorem (Upper Bound)

The minimax expected regret for predicting the target sequence with length n :

$$R \leq \sqrt{\frac{n}{2} \log(N+1)} + \min\{F(n), 0\}. \quad (1)$$

where $F(n)$ depends on the problems' specification.

- The improvement and the trade-off
- The possibility of negative regret

Main Results

Theorem (Lower Bound)

Consider $\mathcal{X} = \{0, 1\}$ and $\mathcal{D} = [0, 1]$, and ℓ is the absolute loss $\ell(x, y) = |x - y|$, then

$$R \geq \sqrt{\frac{n}{2} \log(N+1)} + \left(\xi^* - \frac{1}{2}\right) n, \quad (2)$$

where ξ^* is a constant related to side information.

- Matched negative regret when ξ^* is smaller than $\frac{1}{2}$.

Examples

Binary Symmetric Channel:

$$R \leq \sqrt{\frac{n}{2} \log(N+1)} + \min\{0, nc_1\} \quad (3)$$

$$R \geq \sqrt{\frac{n}{2} \log(N+1)} + nc_2. \quad (4)$$

Zero-mean Gaussian Channel:

$$R \leq \sqrt{\frac{n}{2} \log(N+1)} + \min\{0, nc_3\} \quad (5)$$

$$R \geq \sqrt{\frac{n}{2} \log(N+1)} + nc_4. \quad (6)$$

Future Work

1. To narrow the gap between the upper and lower bound
 - Design a stronger algorithm
 - Derive a tighter lower bound
2. To generalize the problem
 - Extend the work to a continuous instance space
 - Consider more complicated expert class