Non-Stationary Online Structured Prediction with Surrogate Losses

Shinsaku Sakaue

CyberAgent / NII / RIKEN AIP

Han Bao

ISM / Tohoku University / RIKEN AIP

Yuzhou Cao

Nanyang Technological University

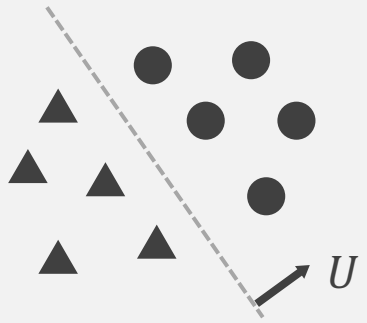
Online Structured Prediction

Example: online classification

For $t = 1, \ldots, T$

Observe x_t and predict $\hat{y}_t \in \{1, \ldots, d\}$

Incur $\mathbb{1}(\hat{y}_t \neq y_t)$ and observe y_t



Existing guarantee

Finite!

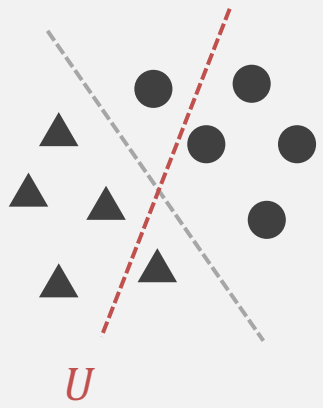
For any fixed estimator U, $\mathbb{E}[\sum_t \text{target}_t(\hat{y}_t)] = \sum_t \text{surrogate}_t(U) + O(\|U\|^2)$

- Rosenblatt (1958): binary and linearly separable
- Van der Hoeven (2020): multiclass, not necessarily separable
- Sakaue et al. (2024): structured prediction, with multiclass as a special case

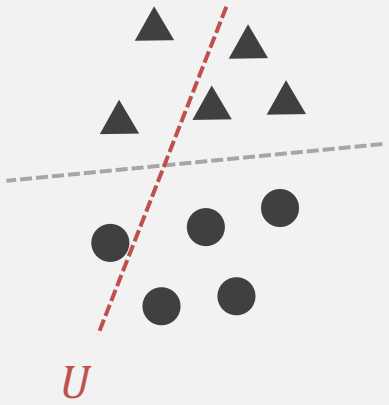
Challenges of Non-Stationary Environments

Example: Ad click prediction (trend shifts)

Daytime: $t = 1, \dots, T/2$



Nighttime: $t = T/2 + 1, \dots, T$



Existing guarantee

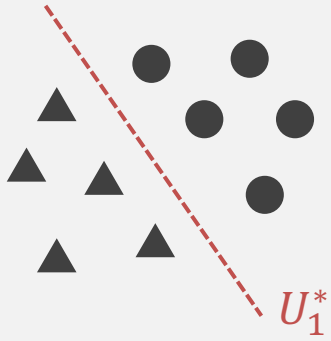
$$\mathbb{E}[\textstyle\sum_t \text{target}_t(\hat{y}_t)] = \sum_t \text{surrogate}_t(U) + O(\|U\|^2)$$

No single U can avoid $\Omega(T)$ cumulative surrogate loss ☹

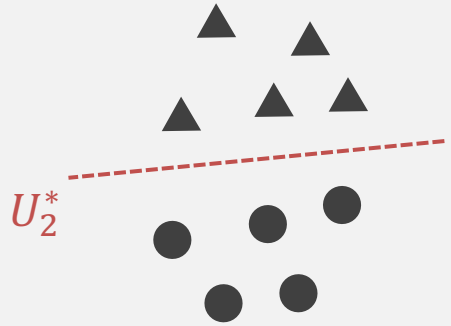
Our Result: "Small-Surrogate-Loss + Path-Length" Bound

Example: Ad click prediction (trend shifts)

Daytime: $t = 1, \dots, T/2$



Nighttime: $t = T/2 + 1, \dots, T$



Main result

For any $U_1, \dots, U_T$, $\mathbb{E}[\sum_t \text{target}_t(\hat{y}_t)] = \sum_t \text{surrogate}_t(U_t) + O(\sum_{t=2}^{T} \|U_t - U_{t-1}\|)$

In the example, first half $U_t = U_1^*$, latter half $U_t = U_2^*$ gives $\mathbb{E}[\sum_t \text{target}_t(\hat{y}_t)] = O(1)$!

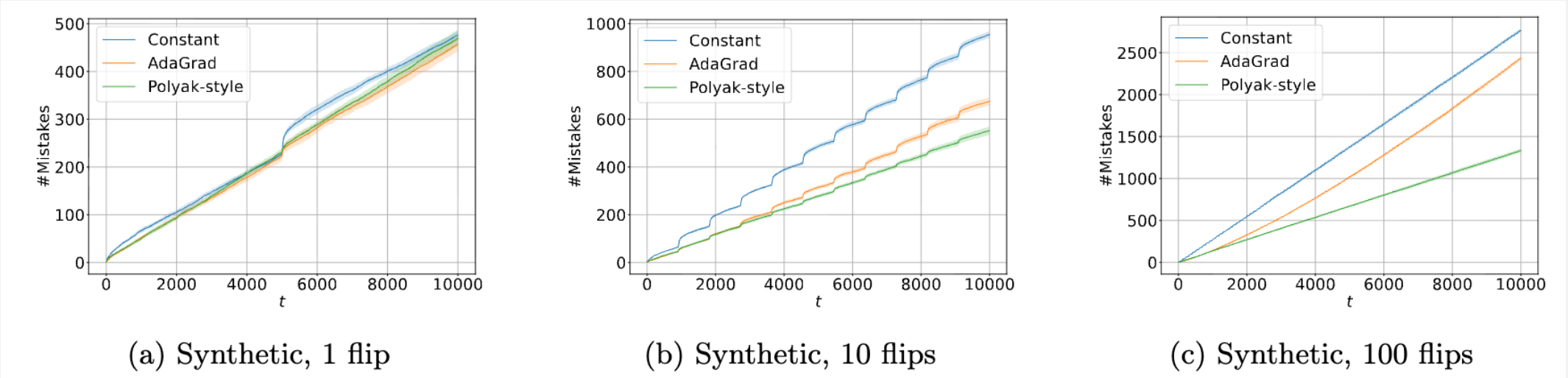
Matching lower bound: linear dependence on $\sum_t \text{surrogate}_t(U_t)$ and $\sum_{t=2}^{T} \|U_t - U_{t-1}\|$ is tight

Also extends to broader structured prediction tasks via convolutional Fenchel–Young losses

Method: Polyak-style Stepsize OGD (Plus Existing Decoding)

Run OGD to update estimator W_t with a monotone Polyak-style stepsize rule:

$$\eta_t = \min\left\{\frac{2(\text{surrogate}_t(W_t) - \mathbb{E}[\text{target}_t(\hat{y}_t)])}{\|\nabla \text{surrogate}_t(W_t)\|^2}, \eta_{t-1}\right\}$$



(a) Synthetic, 1 flip (b) Synthetic, 10 flips (c) Synthetic, 100 flips

Empirically, Polyak-style step size performs better as non-stationarity increases