Efficient Non-stationary Online Learning by ***Wavelets*** with Applications to Online Distribution Shift Adaptation

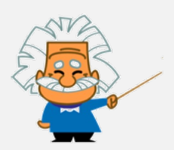
Yu-Yang Qian, Peng Zhao, Yu-Jie Zhang, Masashi Sugiyama, Zhi-Hua Zhou

Learning And Mining from DatA · 東京大学 THE UNIVERSITY OF TOKYO · RIKEN · AIP

Background

☐ Protocol of the **Online Learning** problem:

At each round $t = 1, \dots, T$:



(1) the learner submits a prediction $\boldsymbol{\theta}_t \in \Theta$;

⇩

(2) simultaneously, the environment picks loss $f_t : \Theta \mapsto \mathbb{R}$;

⇩

(3) the learner suffers loss $f_t(\boldsymbol{\theta}_t)$ and updates model.

☐ Previous Performance Measures:

Worst-case Dynamic Regret: compare with the function minimizers

$$\mathbf{Reg}_T^{\mathbf{d}}(\{f_t, \boldsymbol{\theta}_t^\star\}_{t=1}^T) \triangleq \sum_{t=1}^T f_t(\boldsymbol{\theta}_t) - \sum_{t=1}^T f_t(\boldsymbol{\theta}_t^\star)$$

$(\boldsymbol{\theta}_t^\star \in \arg\min_{\boldsymbol{\theta}\in\Theta} f_t(\boldsymbol{\theta}))$

However: May lead to ***overfitting*** to sample randomness.

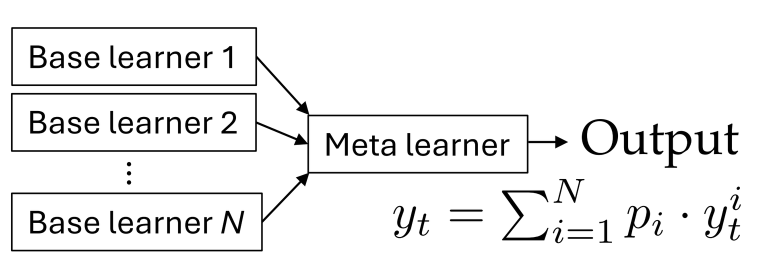
- ➢ Consider online supervised learning with loss $f_t(\boldsymbol{\theta}) = \sum_{(\mathbf{x},y)\in S_t} \ell(y; \mathbf{x}^\top\boldsymbol{\theta})$
- ➢ Only obtain f_t, but the *expected* $F_t(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}[\ell(y; \mathbf{x}^\top\boldsymbol{\theta})]$ is our goal

Universal Dynamic Regret: compare with *any* comparators $\mathbf{u}_1, \dots, \mathbf{u}_T$

$$\mathbf{Reg}_T^{\mathbf{d}}(\{f_t, \mathbf{u}_t\}_{t=1}^T) \triangleq \sum_{t=1}^T f_t(\boldsymbol{\theta}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t)$$

⇨ hold ***universally*** for arbitrary comparator sequence.

However: typically need to deploy a two-layer ***online-ensemble*** [1].



➢ need to maintain multiple (≈ $\log(T)$) base models, when using complex models (such as DNNs), may become ***computational costly.***

Our Measure: Underlying Dynamic Regret

Motivation: how to attain optimal dynamic regret ***without*** deploying an ensemble of multiple base models?

weaker → stronger

Worst-case Dynamic regret | ***Underlying Dynamic regret*** | Universal Dynamic regret

⇩

Underlying Dynamic Regret: compare with ground-truth minimizers

$$\mathbf{Reg}_T^{\mathbf{d}}(\{f_t, \mathring{\mathbf{u}}_t\}_{t=1}^T) \triangleq \sum_{t=1}^T f_t(\boldsymbol{\theta}_t) - \sum_{t=1}^T f_t(\mathring{\mathbf{u}}_t)$$

where $\mathring{\mathbf{u}}_t \in \Theta$ is the ground-truth comparator characterizing the ***underlying distribution*** at round t.

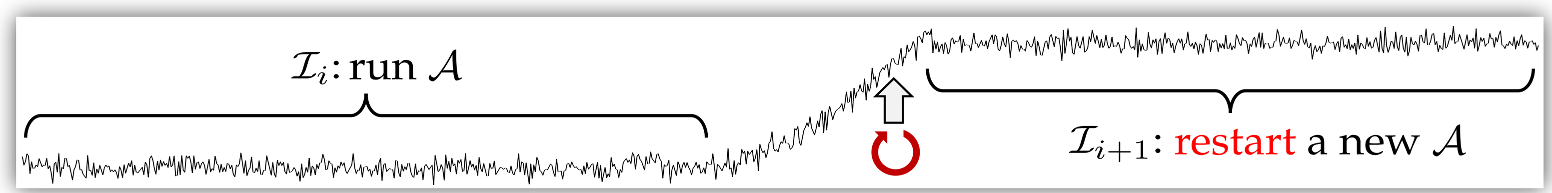
Although $\mathring{\mathbf{u}}_t$ is not observable, we can obtain a ***noisy estimation*** $\widetilde{\mathbf{u}}_t$

Assumption 1 (observation model). *The learner observes $\widetilde{\mathbf{u}}_t$ satisfying $\mathbb{E}[\widetilde{\mathbf{u}}_t] = \mathring{\mathbf{u}}_t$, with a bounded variance of σ^2, i.e., $\mathbb{V}[\widetilde{\mathbf{u}}_t] = \frac{1}{d}\|\widetilde{\mathbf{u}}_t - \mathring{\mathbf{u}}_t\|_2^2 \le \sigma^2$.*

- ➢ $\widetilde{\mathbf{u}}_t$ can be obtained by construct an unbiased estimator;
- ➢ Sufficiently ***general*** to encompass many real learning problem of interest: online label/covariate shift, etc.;
- ➢ By focusing on the specific structure of the stochastic comparator, we achieve a tight regret bound with a ***single-layer algorithm.***

Method: Wavelet Detection-Restart Framework

☐ *Overview* of our framework:



- ➢ "Denoise" the noisy observation $\widetilde{\mathbf{u}}_t$ to approximate $\mathring{\mathbf{u}}_t$;
- ➢ Restart the algorithm once the abrupt changes are ***detected.***

☐ Step 1: Detect environment non-stationarity based on ***wavelets*** [2]:

Algorithm 1: Detection Module by Wavelets

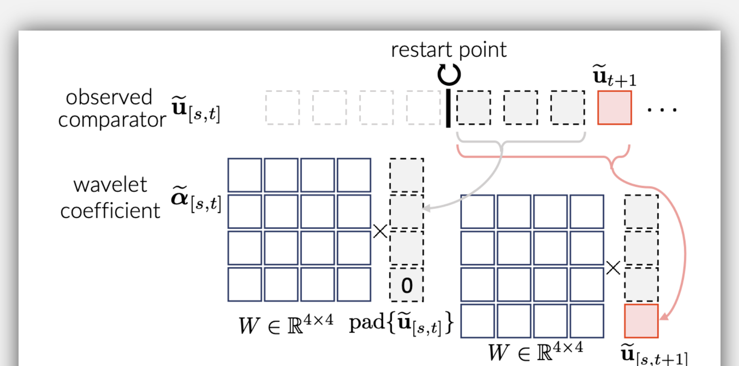
Input: Restart threshold γ; online algorithm $\mathcal{A}$.
Initialize: coefficient matrix $\widetilde{\boldsymbol{\alpha}} = \mathbf{0}$, time $s = 1$;
for $t = 1, \dots, T$ **do**
  Update coefficient matrix $\widetilde{\boldsymbol{\alpha}}_{[s,t]}$ as Sec 3.2;
  if $\|\delta_\gamma(\widetilde{\boldsymbol{\alpha}}_{[s,t]})\|_{\mathrm{F}} > \gamma$ **then**   ($\gamma \approx 1/\text{Variance}$)
    Restart the online algorithm $\mathcal{A}$;
    Reset coefficient matrix $\widetilde{\boldsymbol{\alpha}} = \mathbf{0}$, set $s = t + 1$;
  Output the prediction $\boldsymbol{\theta}_t$ using $\mathcal{A}$;
  Suffer loss $f_t(\boldsymbol{\theta}_t)$, observe $\widetilde{\mathbf{u}}_t$, and update $\mathcal{A}$;
end

- ➢ Maintain only ***one*** learner $\mathcal{A}$
- ➢ Maintain wavelet coefficients of the empirical sequence $\{\widetilde{\mathbf{u}}_t\}$
- ➢ ***Restart*** $\mathcal{A}$ once the norm of coefficients exceed threshold

☐ Step 2: ***Efficiently*** calculate wavelet coefficients in an ***online*** manner:

Traditional Computation

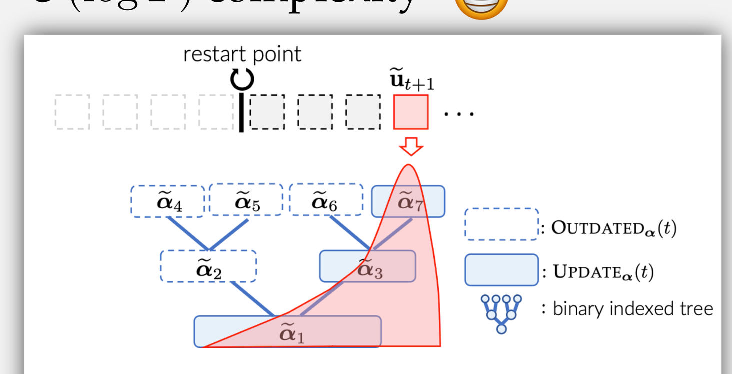
- Matrix multiplication $\widetilde{\boldsymbol{\alpha}}_{[s,t]} = W_{[s,t]}^\top \widetilde{\mathbf{u}}_{[s,t]}$
- Store all data, and recalculate all coeffi.
- $\mathcal{O}(T)$ complexity ☹



(a) traditional matrix computation

our ***Streaming Wavelet Operator***

- Use a binary indexed tree
- Only lazily update *a portion of* coeffi.
- Only maintain the ***norm*** information
- $\mathcal{O}(\log T)$ complexity 😄



(b) streaming wavelet operator

☐ Theoretical Guarantees of ***dynamic regret***:

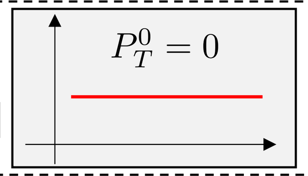
Theorem 1. With prob. at least $1-2/T$, using our detection-restart framework in Algorithm 1 with a $\mathcal{A}$ satisfying certain requirement, we have

- ➢ for ***convex*** function: $\mathbf{Reg}_T^{\mathbf{d}} \le \widetilde{\mathcal{O}}\left(T^{\frac{k+2}{2k+3}}(P_T^k)^{\frac{1}{2k+3}}\right)$
- ➢ for ***exp-concave*** function: $\mathbf{Reg}_T \le \widetilde{\mathcal{O}}\left(T^{\frac{1}{2k+3}}(P_T^k)^{\frac{2}{2k+3}}\right)$

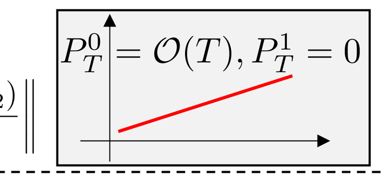
where $P_T^k \triangleq T^k\|\boldsymbol{D}^{k+1}\mathring{\mathbf{u}}_{[1,T]}\|_1$ is the k-th order path length ($k \ge 0$).

* Our method is flexible to accommodate ***higher-order (k-th order)*** path length:

Zeroth-order: $P_T^0 \triangleq \sum_{t=2}^T \|\mathbf{u}_t - \mathbf{u}_{t-1}\|$



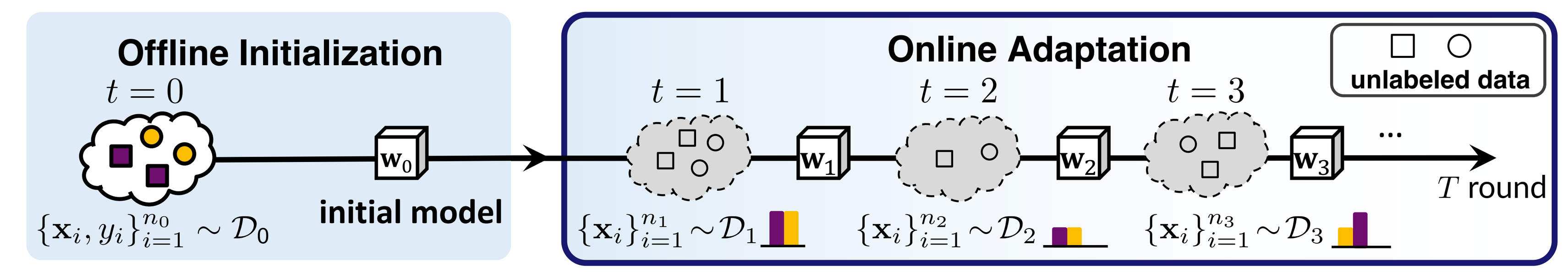
First-order: $P_T^1 \triangleq \sum_{t=3}^T \left\|\frac{(\mathbf{u}_t - \mathbf{u}_{t-1}) - (\mathbf{u}_{t-1} - \mathbf{u}_{t-2})}{2}\right\|$



……

Application: Online Label Shift [3]

☐ Setting: label distribution $\mathcal{D}_t(y)$ changes over time, and conditional $\mathcal{D}_t(\mathbf{x} \mid y)$ remains unchanged.



☐ Apply our detection-restart framework to solve it:

(i) Get unbiased estimation using BBSE (previous label shift estimator);

(ii) Maintain wavelet coefficients of the estimated label distribution;

(iii) Setting $\mathcal{A}$ as Reweighting or OGD, ***restart*** $\mathcal{A}$ if detecting changes.

Reweighting Update as $\mathcal{A}$: $[h_t(\mathbf{x})]_j = \frac{1}{Z(\mathbf{x})} \frac{[\widehat{\mu}_t]_j}{\mathcal{D}_0(y=j)} [h_0(\mathbf{x})]_j, \forall j \in [K]$

OGD Update as $\mathcal{A}$: $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta_t \nabla \widehat{R}_t(\mathbf{w}_t)]$, with $\eta_t = 1/\sqrt{t-s}$

Both achieving ***optimal*** dynamic regrets for online label shift.

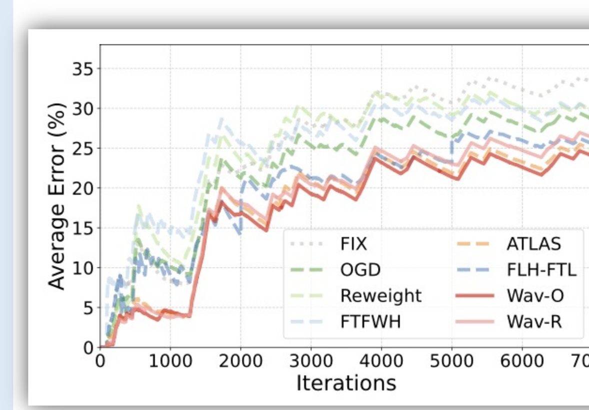
Experiments

		Linear	Square
FIX		7.87±0.03	7.98±0.04
OGD		5.35±0.02	6.10±0.03
Reweight		6.08±0.01	6.45±0.02
FTFWH		5.27±0.02	6.52±0.01
ATLAS		5.44±0.02	5.65±0.03
Effi-Dyn		5.30±0.02	5.83±0.01
FLH-FTL		5.28±0.03	5.64±0.02
Wav-O	$k=0$	5.30±0.01	5.67±0.03
	$k=1$	**5.25±0.01**	5.61±0.02
	$k=2$	5.47±0.03	5.58±0.03

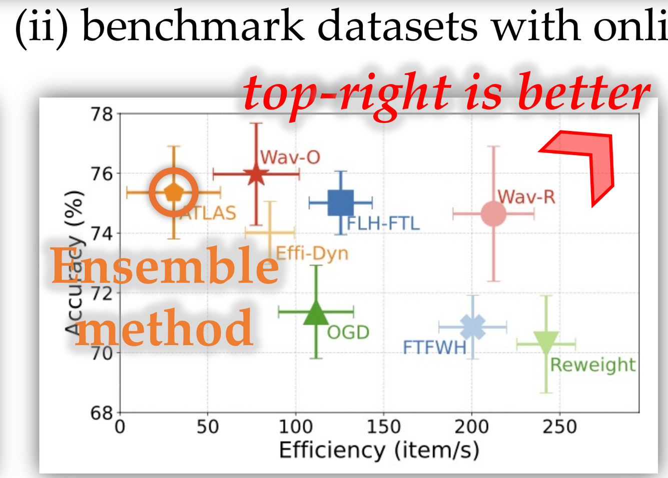
(i) synthetic OCO

	Linear Shift							Square Shift						
	FIX	OGD	Reweight	FTFWH	ATLAS	Wav-O	Wav-R	FIX	OGD	Reweight	FTFWH	ATLAS	Wav-O	Wav-R
CIFAR10	∘20.89 ±0.13	15.92 ±0.13	∘18.25 ±0.45	∘ 17.32 ±0.15	15.75 ±0.12	**15.52** ±0.15	15.68 ±0.14	∘20.79 ±0.04	∘16.31 ±0.13	∘17.38 ±0.15	∘16.70 ±0.14	∘15.21 ±0.08	**14.72** ±0.11	15.55 ±0.13
CINIC10	∘34.56 ±0.24	∘27.31 ±0.21	∘32.42 ±2.55	∘ 28.55 ±0.12	∘26.44 ±0.21	**26.11** ±0.13	28.21 ±0.11	∘34.01 ±0.12	∘28.96 ±0.10	∘28.62 ±0.13	∘28.01 ±0.05	∘27.01 ±0.11	**26.12** ±0.05	26.65 ±0.13
EuroSAT	∘15.42 ±0.12	9.13 ±0.15	∘12.34 ±3.17	∘ 11.35 ±0.12	7.21 ±0.13	**7.15** ±0.12	7.25 ±0.12	∘14.19 ±0.15	7.33 ±0.11	∘ 8.88 ±0.09	∘10.19 ±0.12	**6.99** ±0.09	7.43 ±0.05	7.82 ±0.03
Fashion	∘11.35 ±0.05	7.98 ±0.06	8.15 ±0.05	**7.84** ±0.08	8.39 ±0.09	8.34 ±0.13	8.37 ±0.08	∘11.94 ±0.13	∘ 8.46 ±0.07	∘ 8.67 ±0.11	∘ 8.28 ±0.13	∘ 8.13 ±0.12	7.88 ±0.07	**7.32** ±0.07
MNIST	∘ 1.72 ±0.03	1.13 ±0.03	∘ 1.32 ±0.04	∘ 1.25 ±0.03	**1.07** ±0.05	1.09 ±0.02	1.13 ±0.03	∘ 1.83 ±0.08	∘ 1.17 ±0.07	∘ 1.36 ±0.08	∘ 1.17 ±0.07	1.05 ±0.02	1.12 ±0.04	**1.03** ±0.03

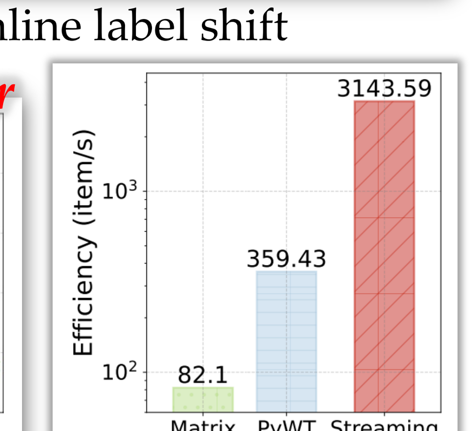
(ii) benchmark datasets with online label shift



(iii) *real-world* locomotion data



(iv) efficiency comparison



(v) *streaming wavelet operator*

[1] Zhao et al., Adaptivity and Non-stationarity: Problem-dependent Dynamic Regret for Online Convex Optimization, JMLR 2024.

[2] Baby and Wang, Online Forecasting of Total-Variation-bounded Sequences, NeurIPS 2019.

[3] Bai et al., Adapting to online label shift with provable guarantees, NeurIPS 2022.