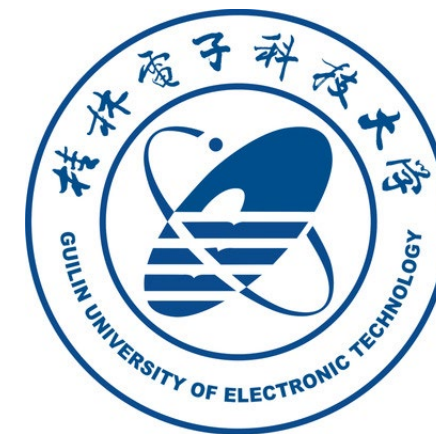


Online Ensemble: A Theoretical Framework for Non-stationary Online Learning

Peng Zhao

School of Artificial Intelligence
Nanjing University
2025.05.31 @桂林电子科技大学



桂林山水甲天下



Outline



- Background
- Online Ensemble
- Case Studies
- Conclusion

Outline



- Background
- Online Ensemble
- Case Studies
- Conclusion

Machine Learning

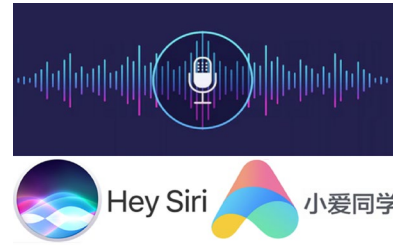
- Machine Learning has achieved great success in recent years.



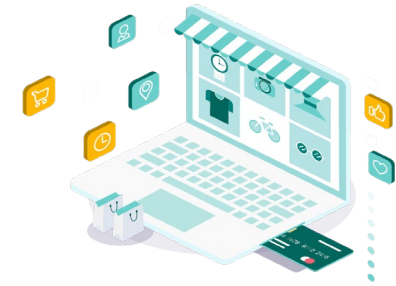
image recognition



search engine



voice assistant



recommendation



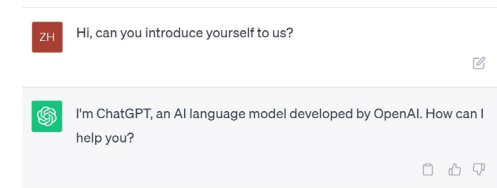
AlphaGo Games



automatic driving

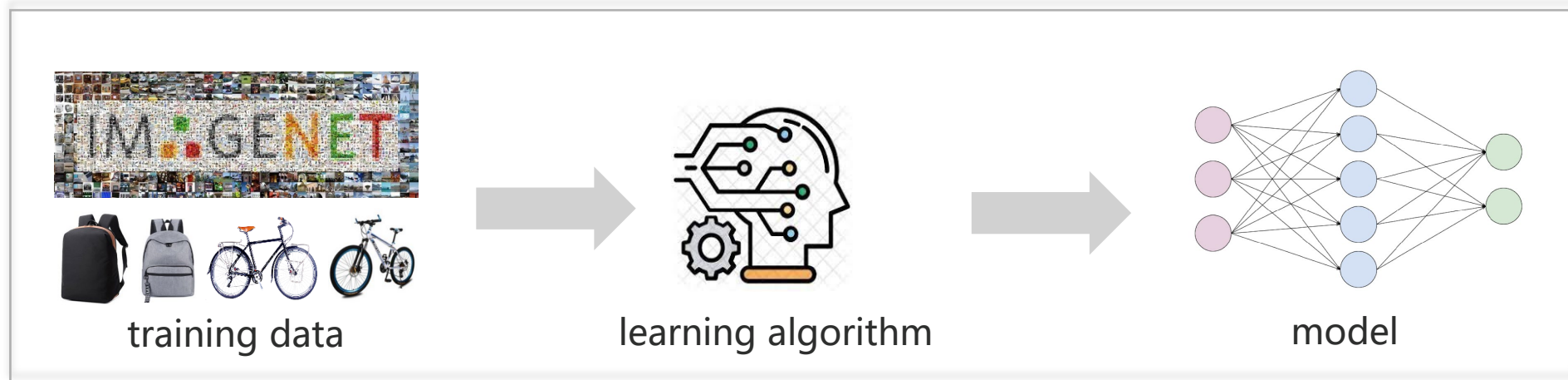


medical diagnosis

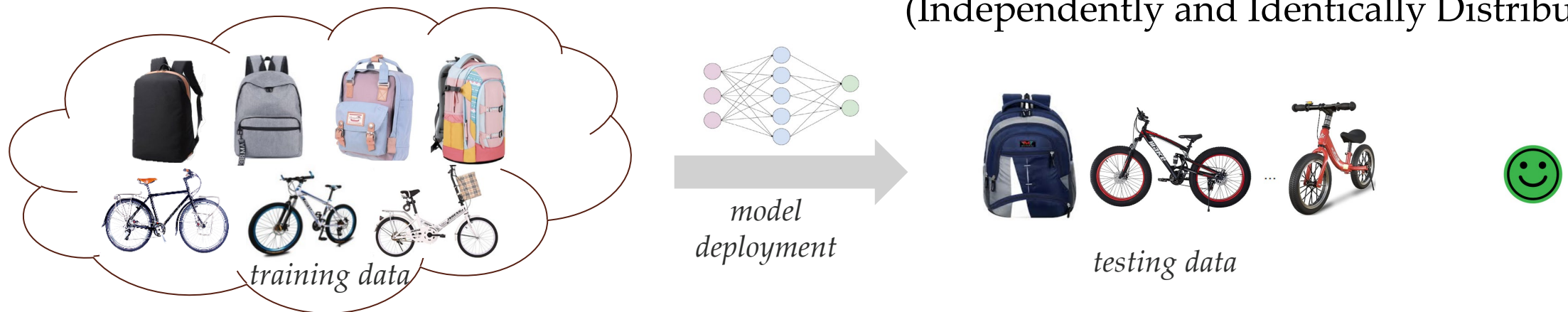


large language model

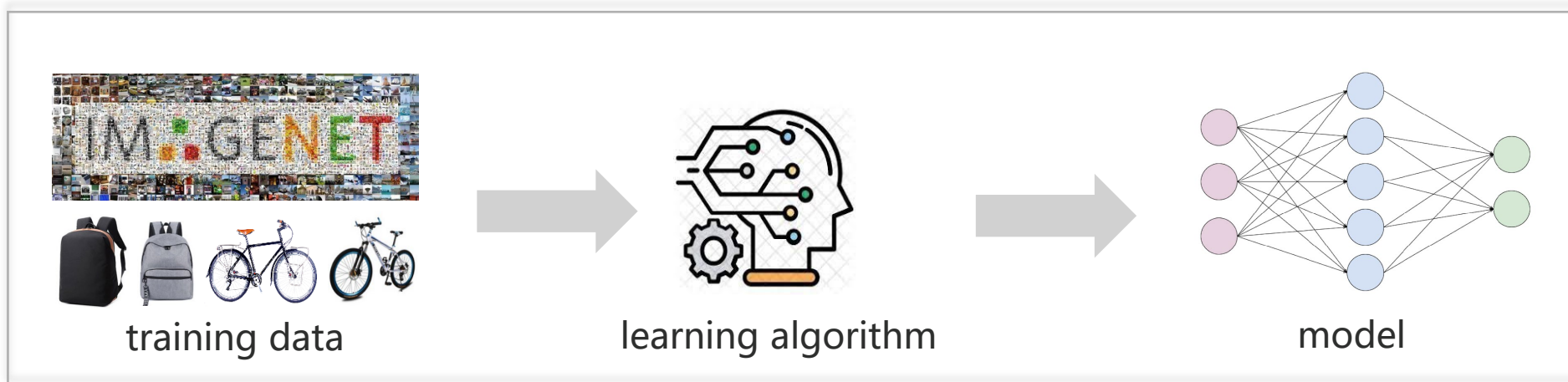
Machine Learning



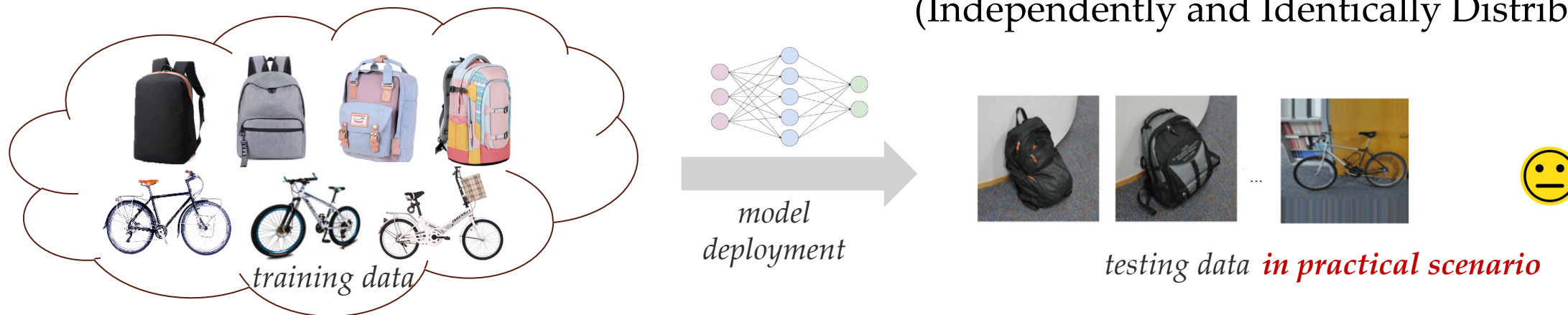
- The theoretical foundation for ML to work well: **I.I.D. assumption**
(Independently and Identically Distributed)



Machine Learning

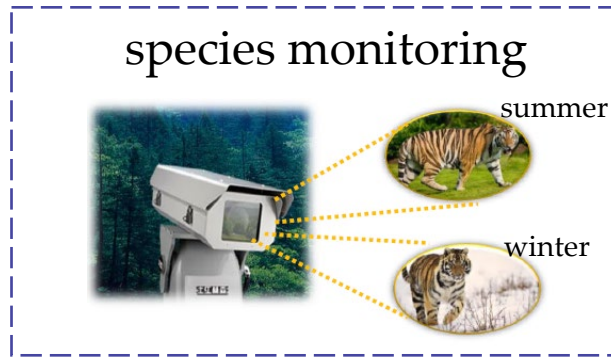


- The theoretical foundation for ML to work well: **I.I.D. assumption**
(Independently and Identically Distributed)



Open-environment Machine Learning

- **Distribution shift:** data are usually collected in open environments



- In many applications, data are coming in an online fashion, like a “*stream*”



provably robust methods for
non-stationary online learning

Community Discussions



“机器学习：发展与未来”

2016年中国计算机大会 特邀报告



Zhi-Hua Zhou

Nanjing University

IJCAI President

Fellow of AAAI/ACM/IEEE



机器学习：发展与未来

周志华

南京大学
计算机软件新技术国家重点实验室

<http://cg.nju.edu.cn/~zhouzh/>

传统机器学习任务

主要针对**封闭静态环境**（重



传统机器学习任务

主要针对**封闭静态环境**（重要因素大多是“定”的）



<http://cs.nju.edu.cn/zhouzh/>

Community Discussions

“Deep Learning for AI”

Communication of ACM

July, 2021. Vol 64. No 7.



2018 Turing Award Recipients

turing lecture

DOI:10.1145/3448250

How can neural networks learn the rich internal representations required for difficult tasks such as recognizing objects or understanding language?

BY YOSHUA BENGIO, YANN LECUN, AND GEOFFREY HINTON

Deep Learning for AI

TURING LECTURE

Yoshua Bengio, Yann LeCun, and Geoffrey Hinton are recipients of the 2018 ACM A.M. Turing Award for breakthroughs that have made deep neural networks a critical component of computing.

RESEARCH ON ARTIFICIAL neural networks was motivated by the observation that human intelligence emerges from highly parallel networks of relatively simple, non-linear neurons that learn by adjusting the strengths of their connections. This observation leads to a central computational question: How is it possible for networks of this general kind to learn the complicated internal representations that are required for difficult tasks such as recognizing

objects or understanding language? Deep learning seeks to answer this question by using many layers of activity vectors as representations and learning the connection strengths that give rise to these vectors by following the stochastic gradient of an objective function that measures how well the network is performing. It is very surprising that such a conceptually simple approach has proved to be so effective when applied to large training sets using huge amounts of computation and it appears that a key ingredient is depth: shallow networks simply do not work as well.

We reviewed the basic concepts and some of the breakthrough achievements of deep learning several years ago.⁸⁰ Here we briefly describe the origins of deep learning, describe a few of the more recent advances, and discuss some of the future challenges. These challenges include learning with little or no external supervision, coping with test examples that come from a different distribution than the training examples, and using the deep learning approach for tasks that humans solve by using a deliberate sequence of steps which we attend to consciously—tasks that Kahneman⁸⁶ calls *system 2* tasks as opposed to *system 1* tasks like object recognition or immediate natural language understanding, which generally feel effortless.

From Hand-Coded Symbolic Expressions to Learned Distributed Representations

There are two quite different paradigms for AI. Put simply, the logic-inspired paradigm views sequential reasoning as the essence of intelligence and aims to implement reasoning in computers using hand-designed rules of inference that operate on hand-designed symbolic expressions that formalize knowledge. The brain-inspired paradigm views learning representations from data as the essence of intelligence and aims to implement learning by hand-designing or evolving rules for modifying the connec-

What needs to be improved. From the early days, theoreticians of machine learning have focused on the iid assumption, which states that the test cases are expected to come from the same distribution as the training examples. Unfortunately, this is not a realistic assumption in the real world: just consider the non-stationarities due to actions of various agents changing the world, or the gradually expanding mental horizon of a learning agent which always has more to learn and discover. As a practical consequence, the performance of today’s best AI systems tends to take a hit when they go from the lab to the field.

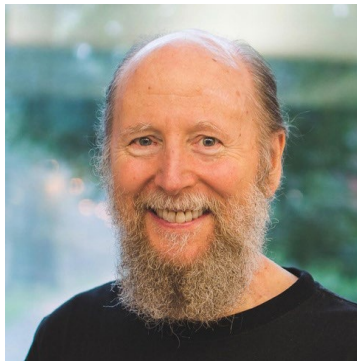
Our desire to achieve greater robustness when confronted with changes in distribution (called out-of-distribution generalization) is a special case of the more general objective of reducing sample complexity (the number of examples needed to generalize well) when faced with a new task—as in transfer learning and lifelong learning⁸¹—or simply with a change in distribution or

Community Discussions



“The Alberta Plan for AI Research”

Aug, 2022



Rich Sutton

2024 Turing Award Recipient

The Alberta Plan for AI Research

Richard S. Sutton, Michael Bowling, and Patrick M. Pilarski

University of Alberta
Alberta Machine Intelligence Institute
DeepMind Alberta

History suggests that the road to a firm research consensus is extraordinarily arduous.
— Thomas Kuhn, *The Structure of Scientific Revolutions*

Herein we describe our approach to artificial intelligence (AI) research, which we call *the Alberta Plan*. The Alberta Plan is pursued within our research groups in Alberta and by others who are like minded throughout the world. We welcome all who would join us in this pursuit.

The Alberta Plan is a long-term plan oriented toward basic understanding of computational intelligence. It is a plan for the next 5–10 years. It is not concerned with immediate applications of what we currently know how to do, but rather with filling in the gaps in our current understanding. As computational intelligence comes to be understood it will undoubtedly profoundly affect our economy, our society, and our individual lives. Although all the consequences are difficult to foresee, and every powerful technology contains the potential for abuse, we are convinced that the existence of more far-sighted and complex intelligence will overall be good for the world.

Following the Alberta Plan, we seek to understand and create long-lived computational agents that interact with a vastly more complex world and come to predict and control their sensory

“在有限的计算能力下，并在其他智能体可能存在的情况下，**通过持续感知和行动实现在线实现回报最大化**。这种描述看似自然，甚至显而易见，但却与当前的实践截然不同，因为**当下研究仍普遍依赖于离线学习、准备好的训练集、人工辅助和无限计算资源**。”

The Alberta Plan characterizes the problem of AI as the online maximization of reward via continual sensing and acting, with limited computation, and potentially in the presence of other agents. This characterization might seem natural, even obvious, but it is also contrary to current practice, which is often focused on offline learning, prepared training sets, human assistance, and unlimited computation. The Alberta Plan research vision is both classical and contrarian, and radical in the sense of going to the root.

Outline



- Background
- Online Ensemble
- Case Studies
- Conclusion

Non-stationary Online Learning

- How to achieve theoretical guarantee for non-stationary online learning?



provably robust methods for
non-stationary online learning

- We need:
 - ✓ A clear problem formulation (with assumptions)
 - ✓ A clear performance measure
 - ✓ A clear methodology to guide algorithm designs

Formulation: Online Learning

- View online learning as an **interaction** between *learner* and *environment*.

Online Convex Optimization

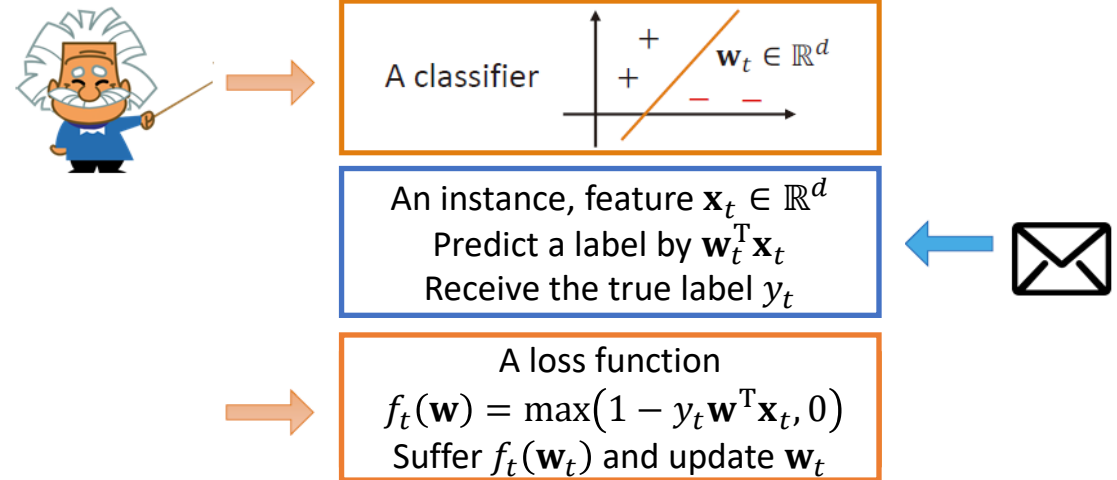
At each round $t = 1, 2, \dots, T$

1. learner first provides a model $\mathbf{w}_t \in \mathcal{W}$;
2. and simultaneously the environment picks a convex **online function** $f_t : \mathcal{W} \mapsto [0, 1]$;
3. the learner then suffers loss $f_t(\mathbf{w}_t)$ and observes some information of f_t .

Example: online function $f_t : \mathcal{W} \mapsto \mathbb{R}$ is composition of

- (i) loss $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$, and
- (ii) data item: $(\mathbf{x}_t, y_t) \in \mathcal{X} \times \mathcal{Y}$.

$$\Rightarrow f_t(\mathbf{w}) = \ell(\mathbf{w}^\top \mathbf{x}_t, y_t)$$



Spam Filtering
Regular vs Spam ?

Formulation: Online Learning

- View online learning as an **interaction** between *learner* and *environment*.

Online Convex Optimization

At each round $t = 1, 2, \dots, T$

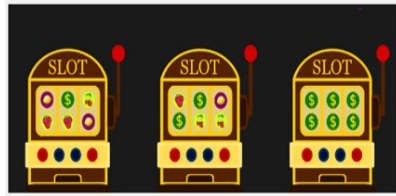
1. learner first provides a model $\mathbf{w}_t \in \mathcal{W}$;
2. and simultaneously the environment picks a convex online function $f_t : \mathcal{W} \mapsto [0, 1]$;
3. the learner then suffers loss $f_t(\mathbf{w}_t)$ and observes **some information of f_t** .

full information

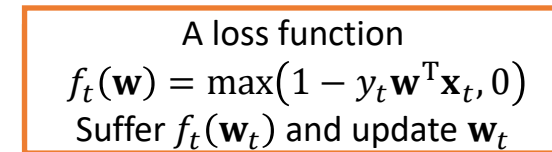
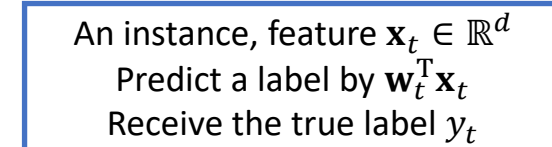
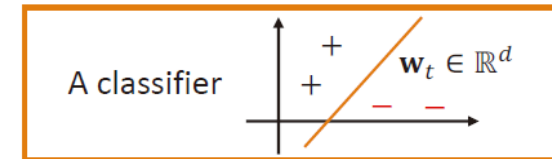


horse racing

partial information



multi-armed bandits



Spam Filtering
 Regular vs Spam ?

Performance Measure

Regret: online prediction as good as the best offline model

$$\text{Regret}_T \triangleq \sum_{t=1}^T f_t(\mathbf{w}_t) - \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^T f_t(\mathbf{w})$$

cumulative loss of the best offline model

Dynamic Regret

$$\text{D-Regret}(\mathbf{u}_1, \dots, \mathbf{u}_T) \triangleq \sum_{t=1}^T f_t(\mathbf{w}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t),$$

where $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{W}$ is a sequence of changing comparators that can be arbitrary chosen.

optimal model **changes** in non-stationary environments

Example: in online supervised learning

- $f_t(\mathbf{w}) \triangleq \ell(\mathbf{w}; \mathbf{x}_t, y_t)$, with $(\mathbf{x}_t, y_t) \sim \mathcal{D}_t$ (unknown)
- $F_t(\mathbf{w}) \triangleq \mathbb{E}_{(\mathbf{x}_t, y_t) \sim \mathcal{D}_t} [\ell(\mathbf{w}; \mathbf{x}_t, y_t)] = \mathbb{E}[f_t(\mathbf{w})]$

$$\text{D-Regret} = \sum_{t=1}^T f_t(\mathbf{w}_t) - \sum_{t=1}^T f_t(\mathbf{w}_t^{*(F)})$$

i.e., $\mathbf{u}_t = \mathbf{w}_t^{*(F)} \in \arg \min_{\mathbf{w} \in \mathcal{W}} F_t(\mathbf{w})$

Challenge

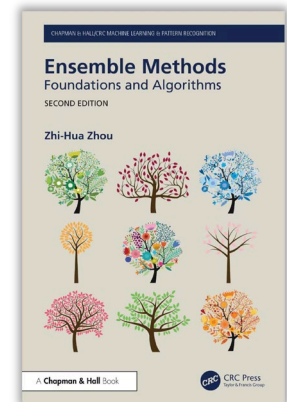
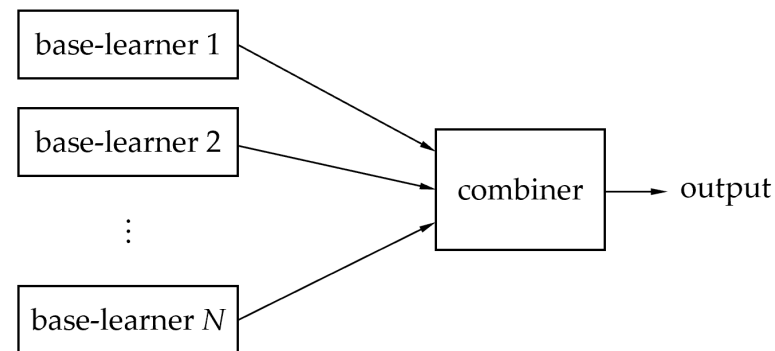


$$\text{D-Regret}(\mathbf{u}_1, \dots, \mathbf{u}_T) = \sum_{t=1}^T f_t(\mathbf{w}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t)$$

Key difficulty: the *uncertainty* due to unknown environmental changes.

Basic idea: Ensemble Method

- *Protocol*: combine multiple base learners to achieve robustness
- *Advantage*: achieve more robust results under uncertain or even changing environments

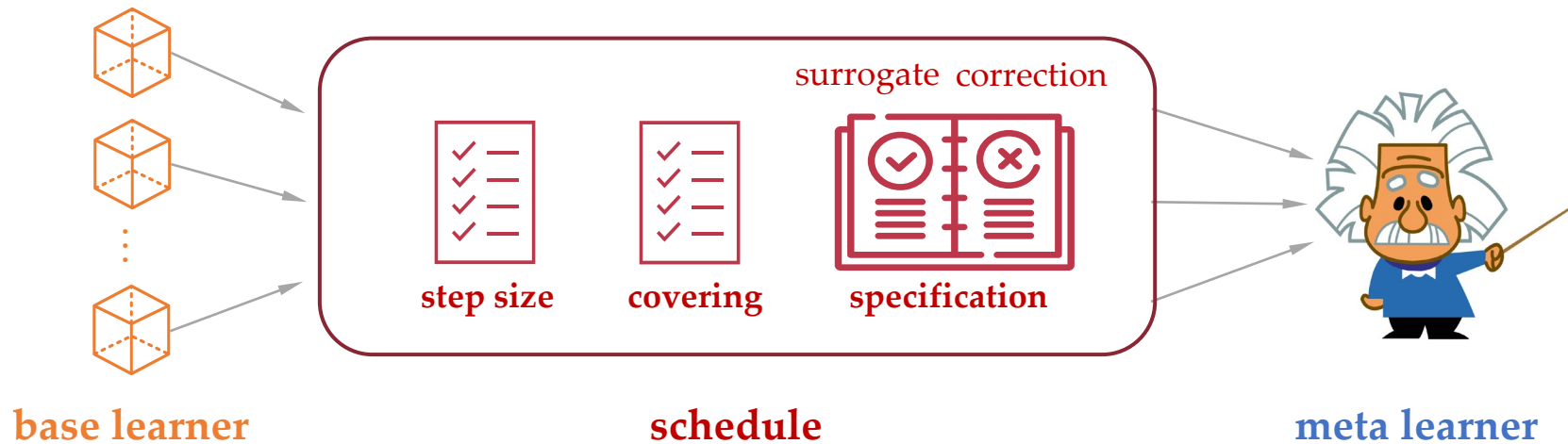


Zhi-Hua Zhou. Ensemble Methods: Foundations and Algorithms, 2nd edition, 2025.

在线集成 (Online Ensemble)

Key Components

- (1) **base learner**: an online learner to cope with a certain amount of non-stationarity
- (2) **schedule**: a set of parameters for initiating base learners that encourage diversity
- (3) **meta learner**: an expert-tracking learner that can combine base learners' decisions



Deploying Online Ensemble

According to the feedback information of online learning, we have

- ❑ Full-information online learning [NeurIPS'20; ICML'22; NerIPS'22; NerIPS'23; JMLR'24; ICML'25]
the learner can obtain the *gradient information* of the online function
- ❑ Partial-information online learning [AISTATST'20; JMLR'21; COLT'22; AISATST'23; ICML'24]
the learner can obtain the *function value* only, without the gradient information
- ❑ Decision-dependent online learning [ICML'22; JMLR'23; NeurIPS'23; AISATST'24; NeurIPS'24]
the historical decision may affect the current functions (including gradient and value)

**Based on the unified "online ensemble" framework,
we can obtain optimal (or best-known) dynamic regret**

A theoretical support for many practices



计算机研究与发展

Journal of Computer Research and Development

ISSN 1000-1239/CN 11-1777/TP

42(增刊): 222~227, 2005

集成学习算法在增量学习中的应用研究

文益民^{1,2} 杨 旸¹ 吕宝粮¹

¹(上海交通大学计算机科学与工程系 上海 200030)

²(湖南工业职业技术学院 长沙 410007)

(ymwen@cs.sjtu.edu.cn)

Research on the Application of Ensemble Learning Algorithms to Incremental Learning

Wen Yimin^{1,2}, Yang Yang¹, and Lü Baoliang¹

¹(Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200030)

²(Hunan Industry Polytechnic, Changsha 410007)

文益民、杨旸、吕宝粮. [集成学习算法在增量学习中的应用研究](#). 计算机研究与发展. 2005.

A theoretical support for many practices



A Survey on Ensemble Learning for Data Stream Classification

HEITOR MURILO GOMES, JEAN PAUL BARDDAL, and FABRÍCIO ENEMBRECK,
Pontícia Universidade Católica do Paraná
ALBERT BIFET, Institut Mines-Télécom, Télécom ParisTech, Université Paris-Saclay

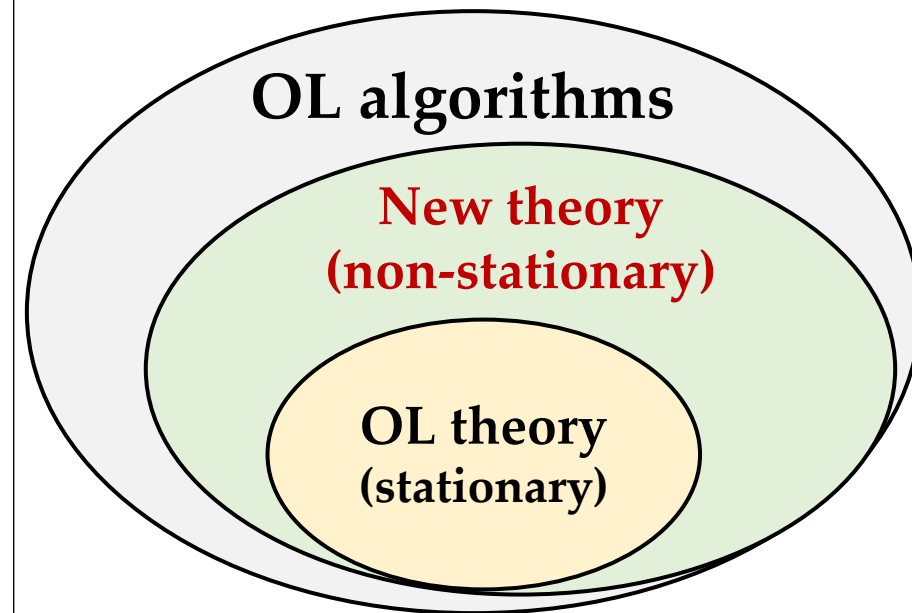
Ensemble-based methods are among the most widely used techniques for data stream classification. Their popularity is attributable to their good performance in comparison to strong single learners while being relatively easy to deploy in real-world applications. Ensemble algorithms are especially useful for data stream learning as they can be integrated with drift detection algorithms and incorporate dynamic updates, such as selective removal or addition of classifiers. This work proposes a taxonomy for data stream ensemble learning as derived from reviewing over 60 algorithms. Important aspects such as combination, diversity, and dynamic updates, are thoroughly discussed. Additional contributions include a listing of popular open-source tools and a discussion about current data stream research challenges and how they relate to ensemble learning (big data streams, concept evolution, feature drifts, temporal dependencies, and others).

CCS Concepts: • **Computing methodologies** → **Ensemble methods**; *Online learning settings*; Supervised learning by classification;

Additional Key Words and Phrases: Ensemble learning, supervised learning, data stream classification

ACM Reference Format:

Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. 2017. A survey on ensemble learning for data stream classification. *ACM Comput. Surv.* 50, 2, Article 23 (March 2017), 36 pages.
DOI: <http://dx.doi.org/10.1145/3054925>



Gomes, Heitor Murilo, et al. "A survey on ensemble learning for data stream classification."
ACM Computing Surveys (CSUR) 50.2 (2017): 1-36.

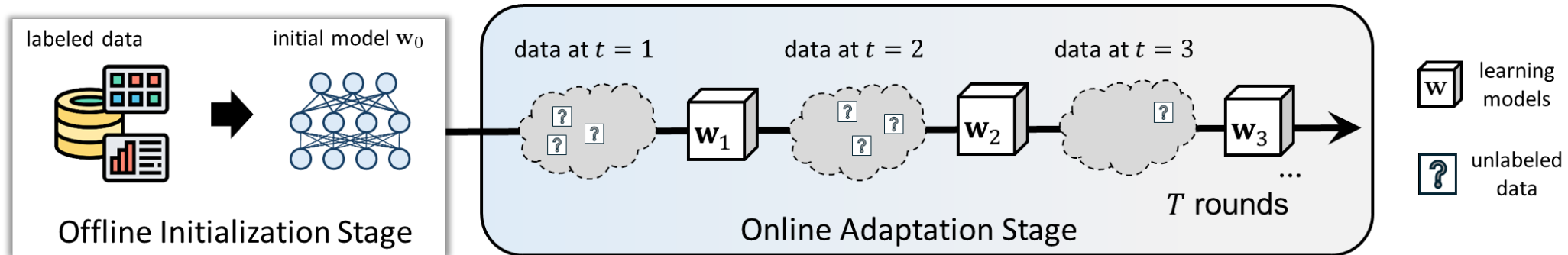
Outline



- Background
- Online Ensemble
- Case Studies
 - Online Label Shift
 - Online Covariate Shift
- Conclusion

Two Case Studies

We will showcase that properly deploying online ensemble can effectively resolve several important online learning problems.

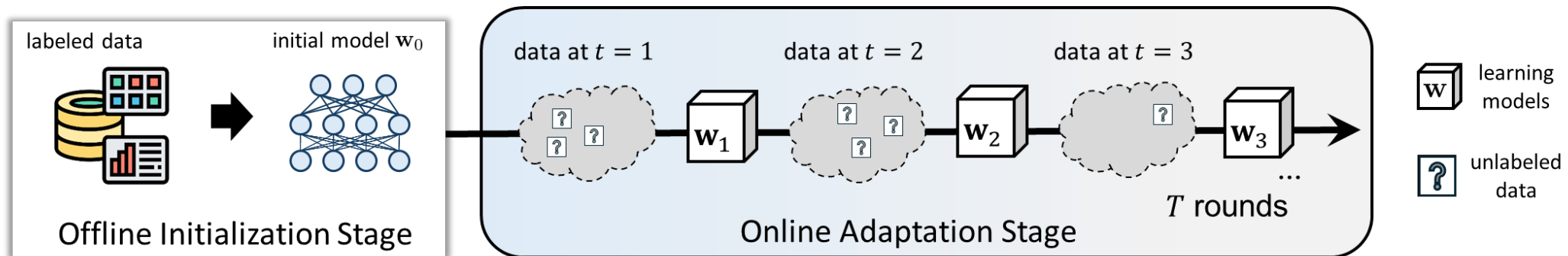


❑ **Label Shift:** $p_{\text{train}}(y) \neq p_{\text{test}}(y)$; but $P_{\text{train}}(x|y) = P_{\text{test}}(x|y)$

❑ **Covariate Shift:** $p_{\text{train}}(\mathbf{x}) \neq p_{\text{test}}(\mathbf{x})$; but $P_{\text{train}}(y|\mathbf{x}) = P_{\text{test}}(y|\mathbf{x})$

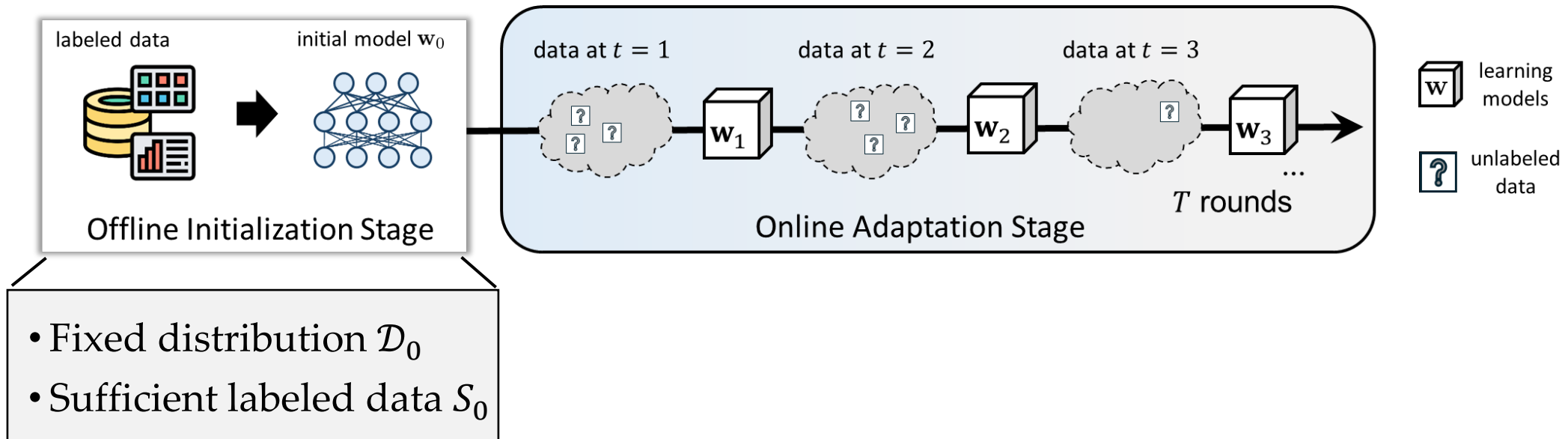
Problem Setup: Two-stage model

- Online distribution shift adaptation: a **two-stage modeling**



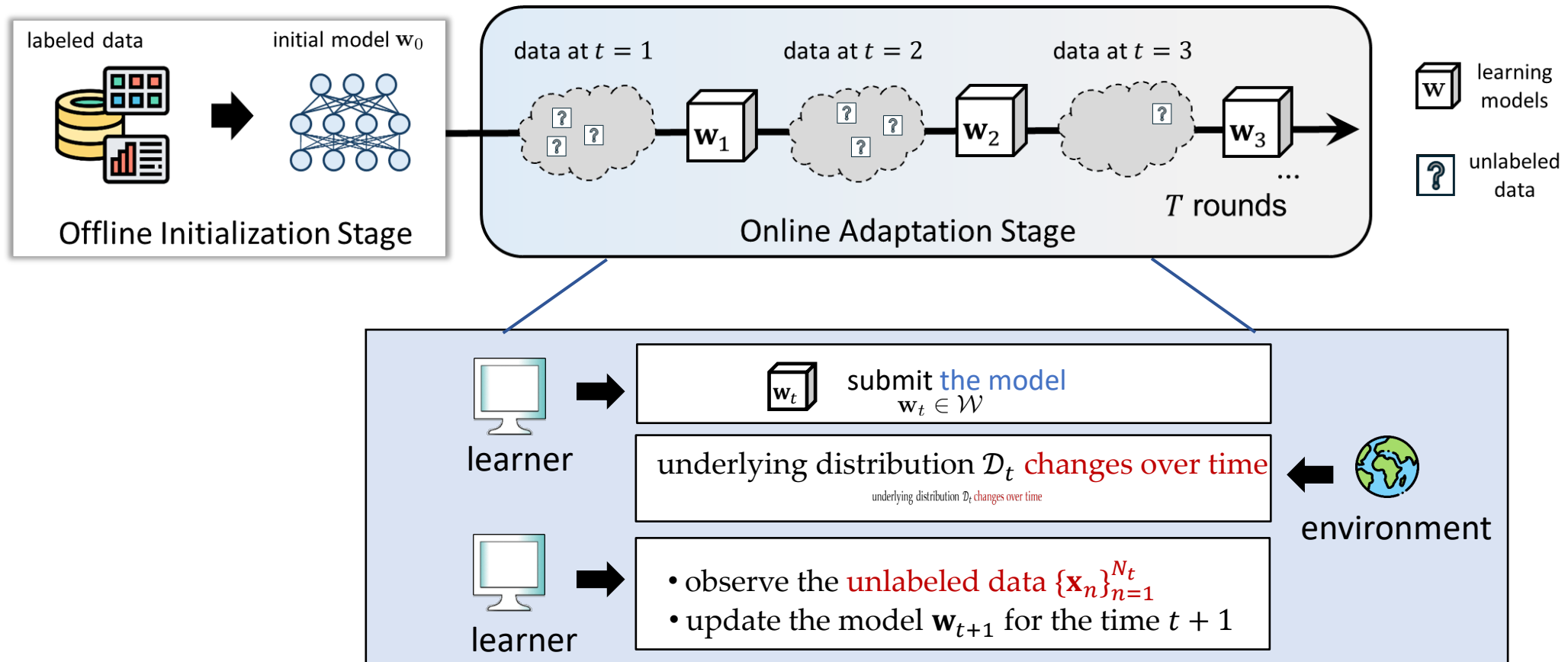
- ① **Initialization:** train an initial model with offline training data
- ② **Online adaptation:** adapt to **sequential distribution shift** with unlabeled test data stream.

Problem Setup: Initialization Stage



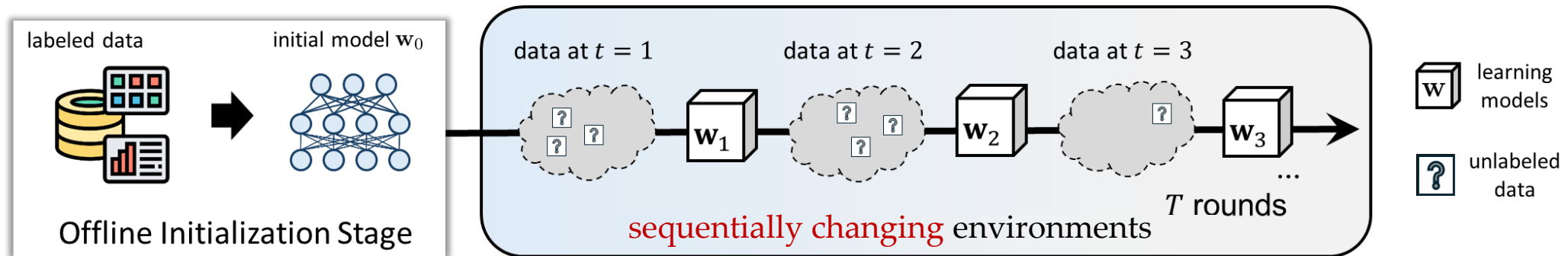
① *Initialization*: train an initial model with offline training data

Problem Setup: Adaptation Stage

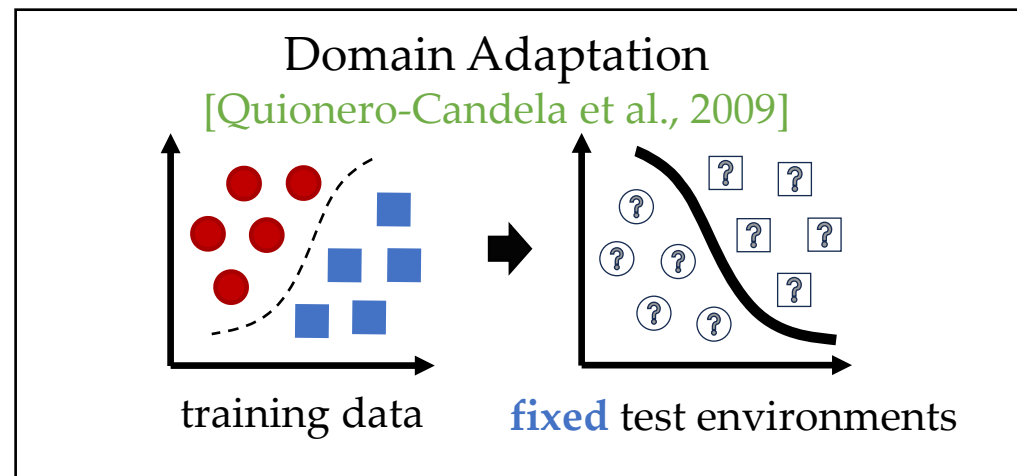


② *Online adaptation:* adapt to **sequential distribution shift** with unlabeled test data stream.

Connection to offline setting



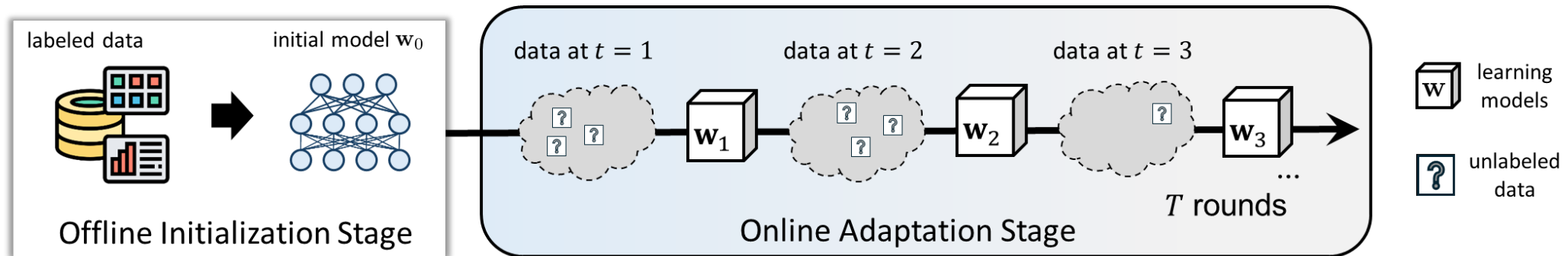
an "online" version



previous works focus on the
"one-step" adaptation

whereas now we consider a
continuous one

Performance Measure

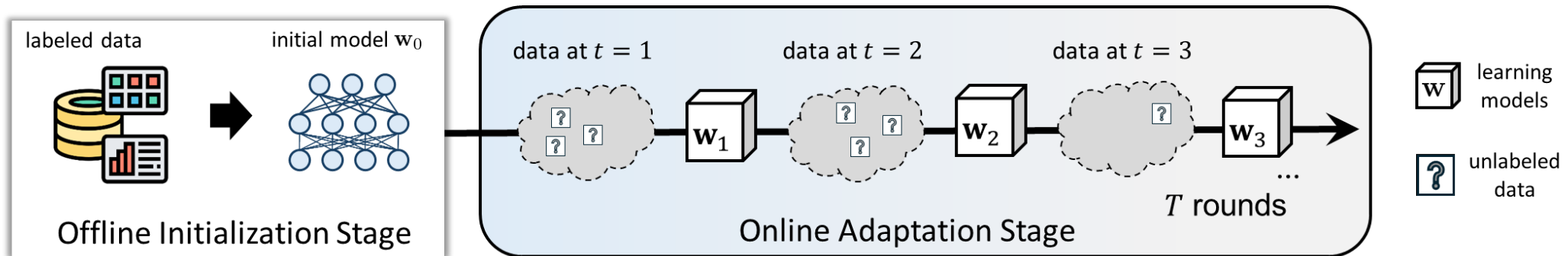


First, to measure the **cumulative risk** of the online models $\{\mathbf{w}_t\}_{t=1}^T$,

$$\sum_{t=1}^T R_t(\mathbf{w}_t) \triangleq \sum_{t=1}^T \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_t} [\ell(\mathbf{w}_t^\top \mathbf{x}, y)],$$

where $\ell(\cdot, \cdot)$ is a certain loss function and risk $R_t(\mathbf{w}_t)$ measures the **averaged error** of the model \mathbf{w}_t over the distribution \mathcal{D}_t

Performance Measure



Goal: minimize the dynamic regret defined over the expected risk,

$$\text{D-Regret}_T = \sum_{t=1}^T R_t(\mathbf{w}_t) - \sum_{t=1}^T R_t(\mathbf{w}_t^*)$$

where $\mathbf{w}_t^* \in \arg \min_{\mathbf{w} \in \mathcal{W}} R_t(\mathbf{w})$ is the Bayes optimal classifier within the model class \mathcal{W} .

Case 1: Online Label Shift

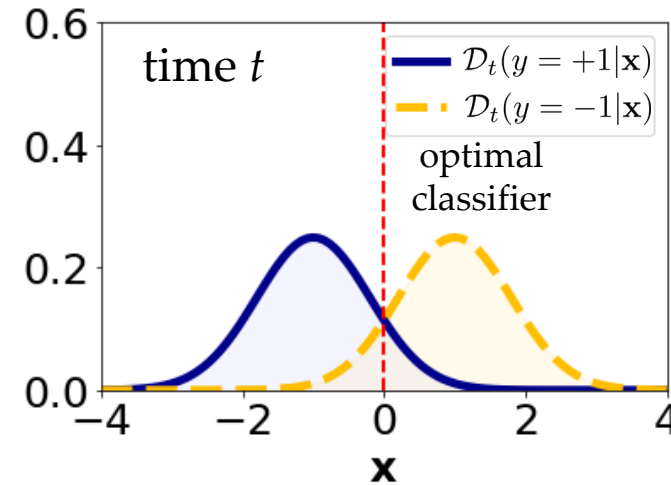
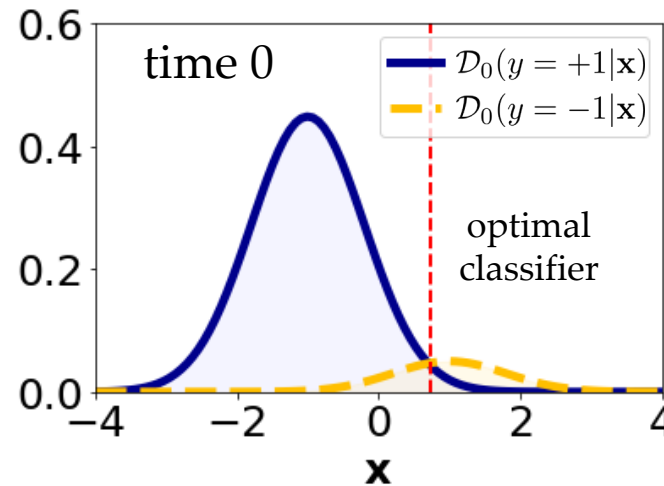
- Label Shift Condition

- Label-conditional input density is unchanged

$$\mathcal{D}_0(\mathbf{x} | y) = \mathcal{D}_t(\mathbf{x} | y)$$

- Change happens on the label distribution

$$\mathcal{D}_t(y) \neq \mathcal{D}_0(y)$$



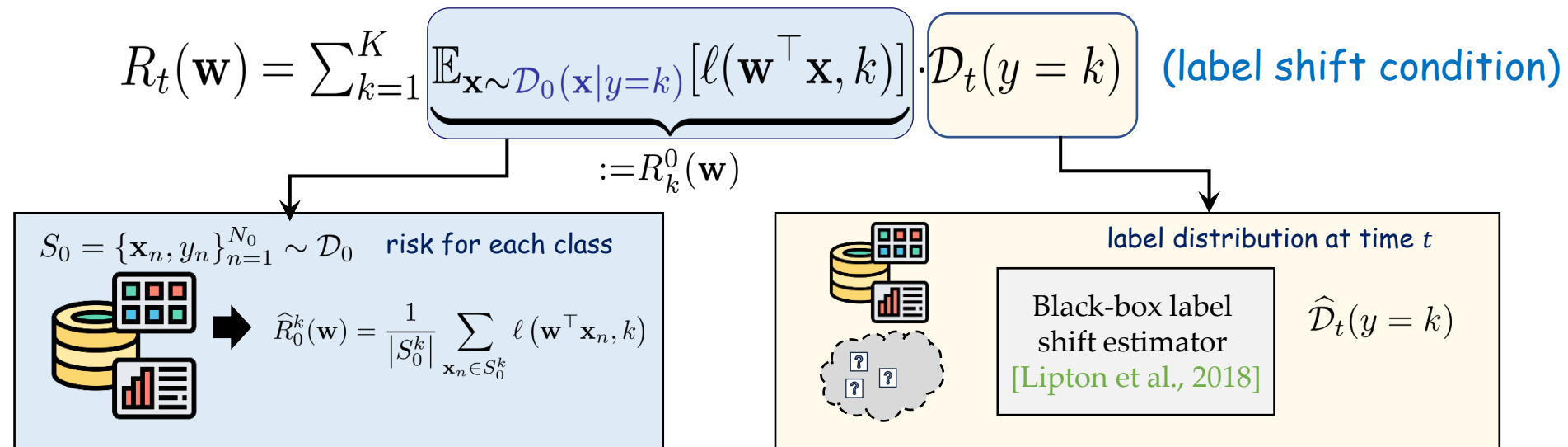
Label shift changes the optimal boundary!

Deploying Online Ensemble to OLS

• Challenge 1: Lack of Supervision

There is **NO label information** available in the online adaptation stage.

➔ **Our Method:** establish an *unbiased* risk estimator $\hat{R}_t(\mathbf{w})$ to evaluate the quality of the model with S_0 and S_t



Accessible with the offline labeled data S_0 .

Estimated with the S_0 and unlabeled data S_t at time t .

Deploying Online Ensemble to OLS

- Challenge 1: Lack of Supervision

There is **NO label information** available in the online adaptation stage.

➡ **Our Method:** establish an *unbiased* risk estimator $\hat{R}_t(\mathbf{w})$ to evaluate the quality of the model with S_0 and S_t

$$R_t(\mathbf{w}) = \sum_{k=1}^K \underbrace{\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_0(\mathbf{x}|y=k)} [\ell(\mathbf{w}^\top \mathbf{x}, k)]}_{:= R_k^0(\mathbf{w})} \cdot \mathcal{D}_t(y = k) \quad (\text{label shift condition})$$

Estimator $\hat{R}_t(\mathbf{w}) = \sum_{k=1}^K \hat{R}_0^k(\mathbf{w}) \cdot \hat{D}_t(y = k)$ is unbiased.

$$\mathbb{E}[\hat{R}_t(\mathbf{w})] = R_t(\mathbf{w}),$$

which guarantees that we can use $\hat{R}_t(\mathbf{w})$ to evaluate the model

Deploying Online Ensemble to OLS

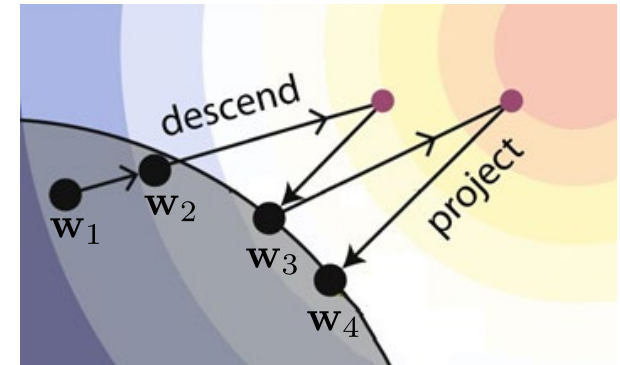
- Based on the feedback model $\hat{R}_t(\mathbf{w})$, we can update the model

$$\text{OGD update: } \mathbf{w}_{t+1} = \Pi_{\mathcal{W}} \left[\mathbf{w}_t - \eta \nabla \hat{R}_t(\mathbf{w}_t) \right]$$

Online Gradient Descent (OGD)

$$\mathbf{w}_{t+1} = \arg \min_{\mathbf{w} \in \mathcal{W}} \underbrace{\langle \nabla \hat{R}_t(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t^* \rangle}_{\text{"loss" on new data}} + \underbrace{\frac{1}{\eta} \|\mathbf{w} - \mathbf{w}_t\|_2^2}_{\text{distance to previous model}}$$

The step size η controls the “amount” of previous information used



<https://www.nature.com/articles/s41534-017-0043-1>

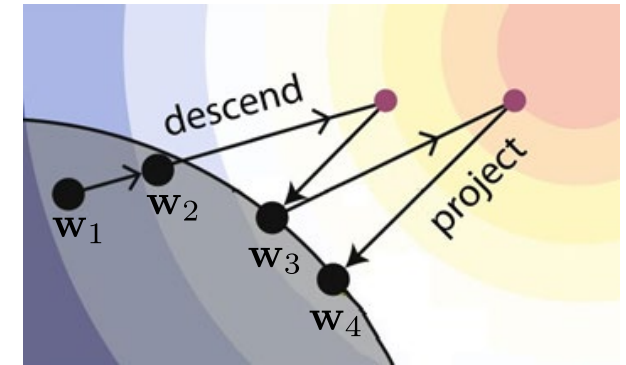
Deploying Online Ensemble to OLS

- Based on the feedback model $\hat{R}_t(\mathbf{w})$, we can update the model

Online Gradient Descent (OGD)

$$\mathbf{w}_{t+1} = \arg \min_{\mathbf{w} \in \mathcal{W}} \underbrace{\langle \nabla \hat{R}_t(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t^* \rangle}_{\text{"loss" on new data}} + \underbrace{\frac{1}{\eta} \|\mathbf{w} - \mathbf{w}_t\|_2^2}_{\text{distance to previous model}}$$

The step size η controls the “amount” of previous information used



<https://www.nature.com/articles/s41534-017-0043-1>

- Observation: choice of the **step size η** matters a lot!

- ✓ Slow change: **small size** to keep close to previous model
- ✓ Fast change: **large step size** to focus on the new data more

Unknown shift intensity of environments!
*Update model **fast or slowly?***

Deploying Online Ensemble to OLS

- Based on the feedback model $\hat{R}_t(\mathbf{w})$, we can update the model

$$\text{OGD update: } \mathbf{w}_t = \Pi_{\mathcal{W}} \left[\mathbf{w}_{t-1} - \eta \nabla \hat{R}_{t-1}(\mathbf{w}_{t-1}) \right]$$

Theorem 1. UOGD update in above with step size η enjoys

$$\mathbb{E} [\text{D-Regret}_T] = \mathbb{E} \left[\sum_{t=1}^T R_t(\mathbf{w}_t) - \sum_{t=1}^T R_t(\mathbf{w}_t^*) \right] \leq \mathcal{O} \left(\eta T + \frac{1}{\eta} + \sqrt{\frac{V_T T}{\eta}} \right),$$

where $V_T = \sum_{t=2}^T \|\boldsymbol{\mu}_{y_t} - \boldsymbol{\mu}_{y_{t-1}}\|_1$ measures the *intensity of the label distribution shift*.

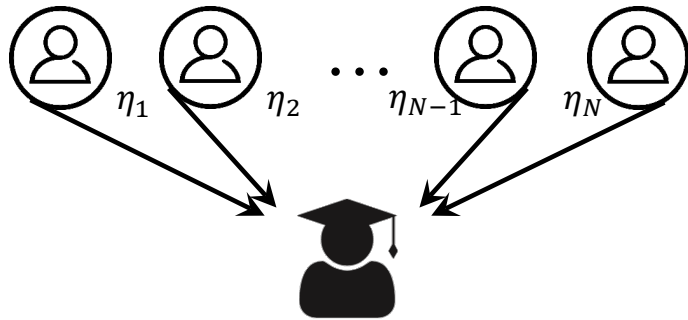
\Rightarrow minimax optimal $\mathcal{O} \left(V_T^{\frac{1}{3}} T^{\frac{2}{3}} \right)$ when setting an optimal step size $\eta^* = \Theta(T^{-\frac{1}{3}} V_T^{\frac{1}{3}})$

Deploying Online Ensemble to OLS

- Challenge 2: Unknown shift intensity

Unknown shift intensity of environments!
Update model *fast or slowly?*

online ensemble framework



$$\Rightarrow \mathbf{w}_t = \sum_{i=1}^N p_{t,i} \cdot \mathbf{w}_{t,i}$$

base-algorithm 

Multiple OGDs learning with **different step sizes**

$$\mathbf{w}_{t+1,i} = \Pi_{\mathbf{w} \in \mathcal{W}} \left[\mathbf{w}_{t,i} - \eta_i \nabla \hat{R}_t(\mathbf{w}_{t,i}) \right].$$

meta-algorithm 

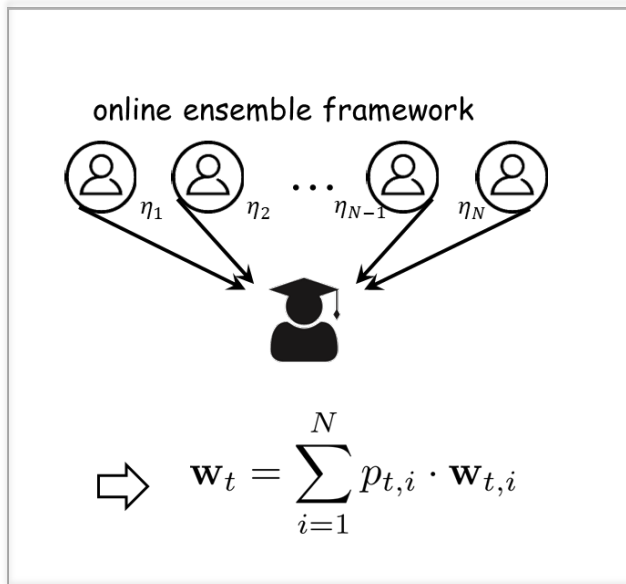
Combining base-algorithms by an adaptive weight \mathbf{p}_t

$$p_{t,i} \propto \exp \left(-\epsilon \sum_{s=1}^{t-1} \hat{R}_s(\mathbf{w}_{s,i}) \right)$$

Deploying Online Ensemble to OLS

- Challenge 2: Unknown shift intensity

Unknown shift intensity of environments!
Update model *fast or slowly?*



base-algorithm 

Multiple OGDs learning with *different step sizes*

$$\mathbf{w}_{t+1,i} = \Pi_{\mathbf{w} \in \mathcal{W}} \left[\mathbf{w}_{t,i} - \eta_i \nabla \hat{R}_t(\mathbf{w}_{t,i}) \right].$$

meta-algorithm 

Combining base-algorithms by an adaptive weight \mathbf{p}_t

$$p_{t,i} \propto \exp \left(-\epsilon \sum_{s=1}^{t-1} \hat{R}_s(\mathbf{w}_{s,i}) \right)$$

- By a careful setting of the candidate step sizes, we can ensure that there exists a base-learner that is trained with *a near-optimal step size*
- Our meta-algorithm can identify *the best base-learner* with a low cost.

Theoretical Guarantee

- Our algorithm for online label shift enjoys an *optimal* dynamic regret.

Theorem 2. Set the step size pool as

$$\mathcal{H} = \left\{ \eta_i = \frac{\Gamma\sigma}{2G\sqrt{KT}} \cdot 2^{i-1} \mid i \in [N] \right\}, \quad \Theta(\log T) \text{ base-learners}$$

where $N = 1 + \lceil \frac{1}{2} \log_2(1 + 2T) \rceil$ is the number of base-learners. ATLAS ensures that

$$\mathbb{E} [\text{D-Regret}_T] \leq \mathcal{O} \left(V_T^{\frac{1}{3}} T^{\frac{2}{3}} \right),$$

for non-degenerated cases of $V_T \geq \Theta(T^{-\frac{1}{2}})$.

Experiments

- Illustration 1: meta-algorithm can adaptively track the suitable step sizes.

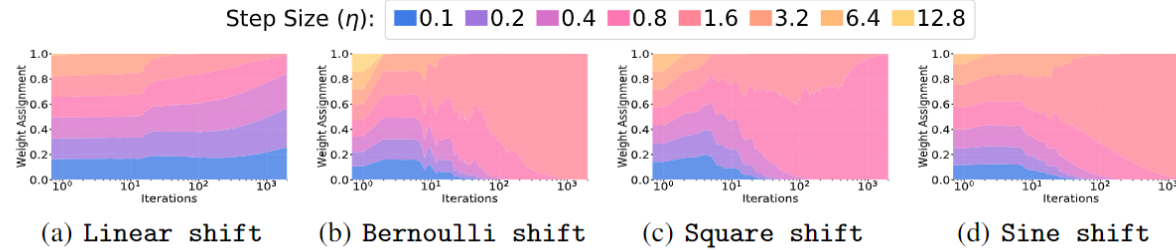


Figure 1: Weight assigned of the ATLAS algorithm for each step size along the learning process. Different colors are used to indicate different step sizes.

- Illustration 2: ATLAS-ADA with *hint functions* improve over vanilla ATLAS.

Table 2: Average error (%) for ATLAS-ADA with four hint functions under different sample sizes. The best one is emphasized in bold. Besides, • indicates a better result than vanilla ATLAS without hint (*None*).

Shift Type	Sample Size: 1					Sample Size: 10					Sample Size: 100				
	<i>None</i>	<i>Win</i>	<i>Peri</i>	<i>Fwd</i>	<i>OKM</i>	<i>None</i>	<i>Win</i>	<i>Peri</i>	<i>Fwd</i>	<i>OKM</i>	<i>None</i>	<i>Win</i>	<i>Peri</i>	<i>Fwd</i>	<i>OKM</i>
Lin	6.28	• 5.89	• 5.99	• 6.01	• 5.35	5.61	• 5.47	• 5.43	• 5.53	• 5.42	5.44	5.44	• 5.38	• 5.40	5.45
	± 0.21	± 0.26	± 0.29	± 0.31	± 0.31	± 0.04	± 0.04	± 0.03	± 0.05	± 0.05	± 0.02	± 0.03	± 0.02	± 0.02	± 0.03
Squ	6.03	• 5.83	• 5.27	5.88	• 5.07	4.59	4.69	• 3.85	• 3.72	• 3.91	4.27	4.68	• 3.39	• 3.33	• 3.46
	± 0.23	± 0.24	± 0.20	± 0.23	± 0.35	± 0.02	± 0.02	± 0.04	± 0.02	± 0.03	± 0.02	± 0.02	± 0.03	± 0.03	± 0.04
Sin	6.90	• 6.58	• 6.59	• 6.43	• 5.25	6.12	• 5.99	• 5.83	• 5.78	• 5.86	5.75	5.78	• 5.53	• 5.48	• 5.58
	± 0.22	± 0.22	± 0.25	± 0.26	± 0.22	± 0.07	± 0.06	± 0.05	± 0.05	± 0.04	± 0.01	± 0.01	± 0.00	± 0.01	± 0.00
Ber	5.55	• 5.42	• 5.43	5.63	• 4.69	4.39	4.45	4.43	• 3.66	• 3.73	4.04	4.29	4.26	• 3.19	• 3.45
	± 0.09	± 0.11	± 0.09	± 0.16	± 0.17	± 0.10	± 0.08	± 0.10	± 0.10	± 0.06	± 0.07	± 0.06	± 0.06	± 0.07	± 0.11

Experiments



■ contenders

■ our algorithms

Table 3: Average error (%) of different algorithms on various real-world datasets. We report the mean and standard deviation over five runs. The best algorithms are emphasized in bold. “•” indicates the algorithms that are significantly inferior to ATLAS-ADA by the paired t -test at a 5% significance level. Here AT-ADA represents ATLAS-ADA (with OKM). The online sample size is set as $N_t = 10$.

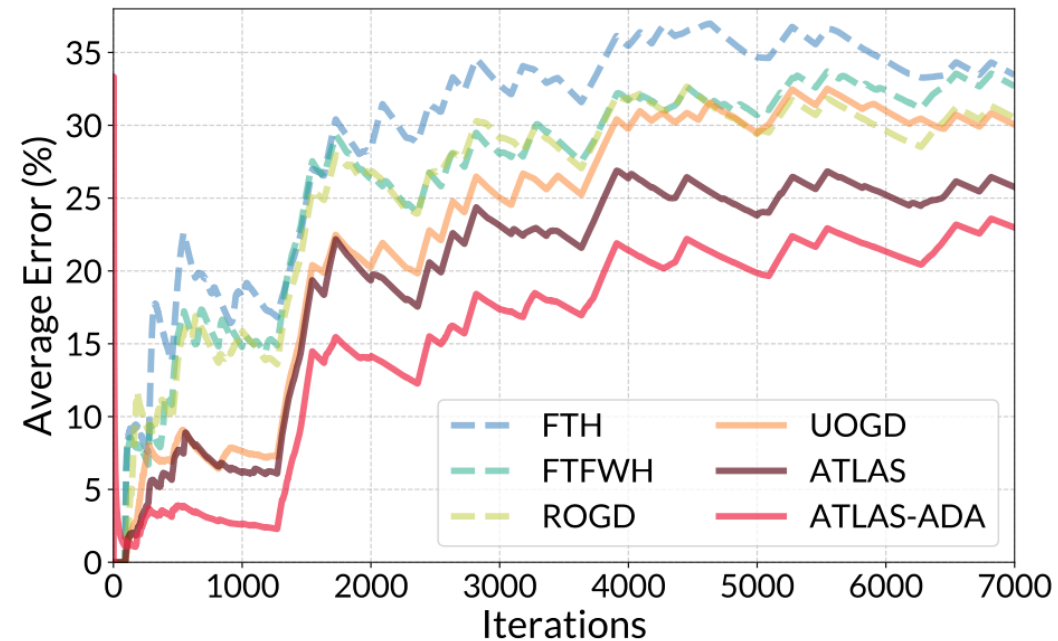
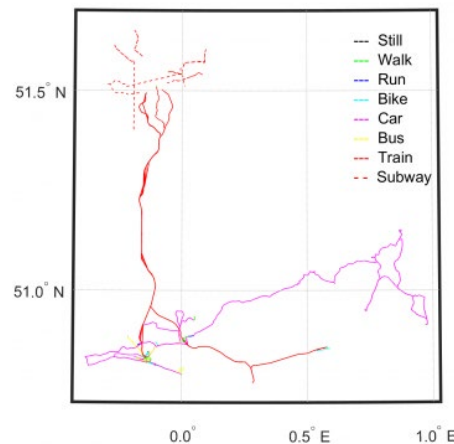
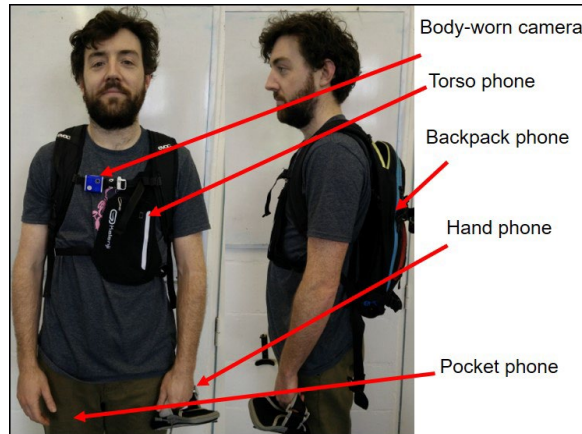
	Lin							Squ						
	FIX	FTH	FTFWH	ROGD	UOGD	ATLAS	AT-ADA	FIX	FTH	FTFWH	ROGD	UOGD	ATLAS	AT-ADA
ArXiv	• 30.28 ±0.07	• 28.18 ±0.28	• 25.74 ±0.21	• 23.09 ±0.20	21.04 ±0.11	• 22.10 ±0.09	21.28 ±0.09	• 30.35 ±0.06	• 26.72 ±0.39	• 28.05 ±0.20	• 24.44 ±0.17	• 21.96 ±0.07	• 21.36 ±0.06	20.80 ±0.06
EuroSAT	• 14.06 ±0.09	• 11.16 ±0.11	• 9.78 ±0.12	• 12.56 ±3.16	7.04 ±0.11	• 7.19 ±0.10	7.13 ±0.11	• 14.15 ±0.11	• 10.22 ±0.08	• 10.26 ±0.06	• 8.91 ±0.05	• 7.30 ±0.07	• 6.97 ±0.08	6.81 ±0.06
MNIST	• 1.79 ±0.02	• 1.38 ±0.03	• 1.20 ±0.02	• 1.25 ±0.02	1.06 ±0.02	1.06 ±0.02	1.06 ±0.02	• 1.79 ±0.04	• 1.26 ±0.03	• 1.28 ±0.04	• 1.32 ±0.04	• 1.13 ±0.03	• 1.04 ±0.02	1.01 ±0.04
Fashion	• 11.86 ±0.04	• 8.47 ±0.07	7.84 ±0.06	8.18 ±0.07	7.95 ±0.08	• 8.36 ±0.07	8.04 ±0.08	• 11.92 ±0.09	• 8.24 ±0.09	• 8.35 ±0.07	• 8.63 ±0.07	• 8.42 ±0.04	• 8.05 ±0.07	7.73 ±0.05
CIFAR10	• 20.77 ±0.12	• 17.36 ±0.14	15.77 ±0.12	• 18.45 ±0.47	15.54 ±0.15	• 15.77 ±0.11	15.62 ±0.14	• 20.77 ±0.08	• 16.67 ±0.12	• 16.72 ±0.12	• 17.40 ±0.11	• 16.29 ±0.09	• 15.18 ±0.07	14.84 ±0.05
CINIC10	• 33.98 ±0.22	• 28.85 ±0.10	• 26.87 ±0.13	• 32.54 ±2.59	26.21 ±0.15	• 26.66 ±0.19	26.38 ±0.16	• 33.99 ±0.16	• 27.99 ±0.09	• 28.08 ±0.08	• 28.58 ±0.09	• 27.00 ±0.14	• 25.94 ±0.13	25.56 ±0.12

Our method can automatically adapt **to continuous label shift**

- **Nearly stationary case (Lin):** comparable with method using all previous data
- **Highly Non-stationary case (Squ):** our algorithm achieves *overperforms all contenders*

Experiments

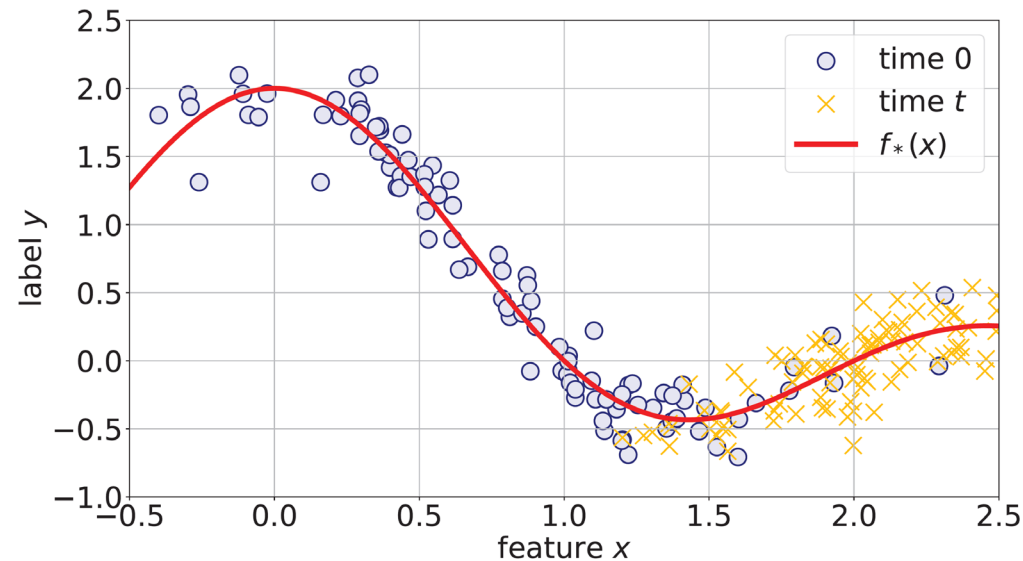
- Sussex-Huawei Locomotion Dataset



Our methods outperform other algorithms.

Case 2: Online Covariate Shift

- Covariate Shift Condition
 - The input-conditional output density is unchanged:
$$\mathcal{D}_0(y | \mathbf{x}) = \mathcal{D}_t(y | \mathbf{x})$$
 - Change happens on **the input distribution** $\mathcal{D}_t(\mathbf{x}) \neq \mathcal{D}_0(\mathbf{x})$



Importance Weighting

- Learn with unlabeled data by **importance weighting**

$$R_t(\mathbf{w}) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_0} [\ell(\mathbf{w}^\top \mathbf{x}, y) \cdot r_t(\mathbf{x})]$$

\mathcal{D}_0 is the offline data distribution $r_t(\mathbf{x}) = \frac{\mathcal{D}_t(\mathbf{x})}{\mathcal{D}_0(\mathbf{x})}$ is the **importance weight**

- Can we directly apply the method for online label shift?

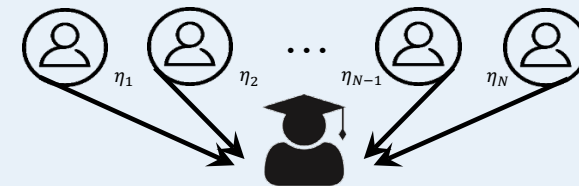
Step 1: establish **an unbiased risk estimator**

$$\hat{R}_t(\mathbf{w}) = \sum_{n=1}^{N_0} [\ell(\mathbf{w}^\top \mathbf{x}_n, y_n) \cdot \hat{r}_t(\mathbf{x}_n)]$$

where \hat{r} is an importance estimator



Step 2: learn with $\hat{R}_t(\mathbf{w})$ by **online ensemble**



No, the importance estimator \hat{r}_t is **hard to be** unbiased in general

Importance Weighting

- Just train the model by **importance-weighted** empirical risk minimization (IWERM):

$$\mathbf{w}_t = \arg \min_{\mathbf{w} \in \mathcal{W}} \hat{R}_t(\mathbf{w}),$$

where $\hat{R}_t(\mathbf{w}) = \sum_{n=1}^{N_0} [\ell(\mathbf{w}^\top \mathbf{x}_n, y_n) \cdot \hat{r}_t(\mathbf{x}_n)]$ is the empirical risk.

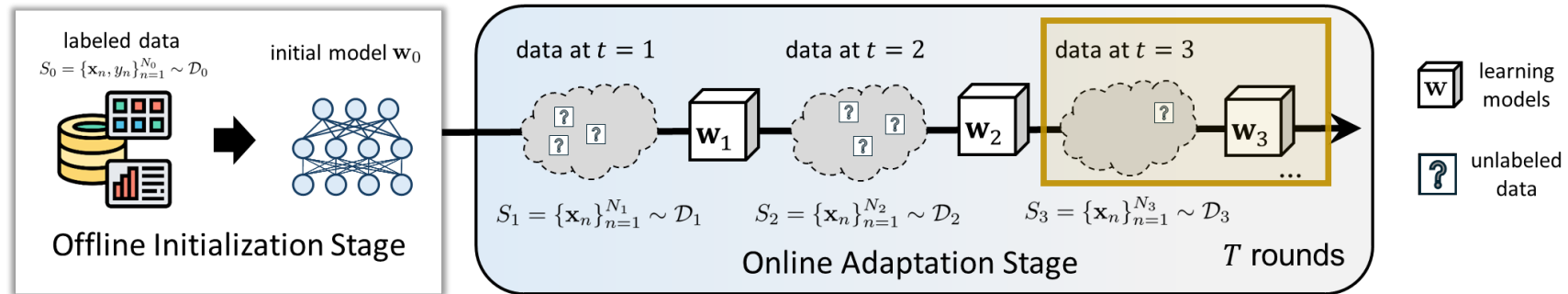
- The quality of the importance estimator \hat{r} matters

Proposition 2. IW works effectively once the time-varying density ratio can be accurately estimated!

Goal $\frac{1}{T} \left(\sum_{t=1}^T R_t(\mathbf{w}_t) - \sum_{t=1}^T R_t(\mathbf{w}_t^*) \right) \leq 5B_x B_\ell \sqrt{\frac{2 \ln((8T)/\delta)}{N_0}} + \frac{2B_\ell}{T} \sum_{t=1}^T \mathbb{E}_{\mathbf{x} \sim S_0} [|\hat{r}_t^*(\mathbf{x}) - r_t(\mathbf{x})|].$

importance weight estimation error

Online Density Ratio Estimation



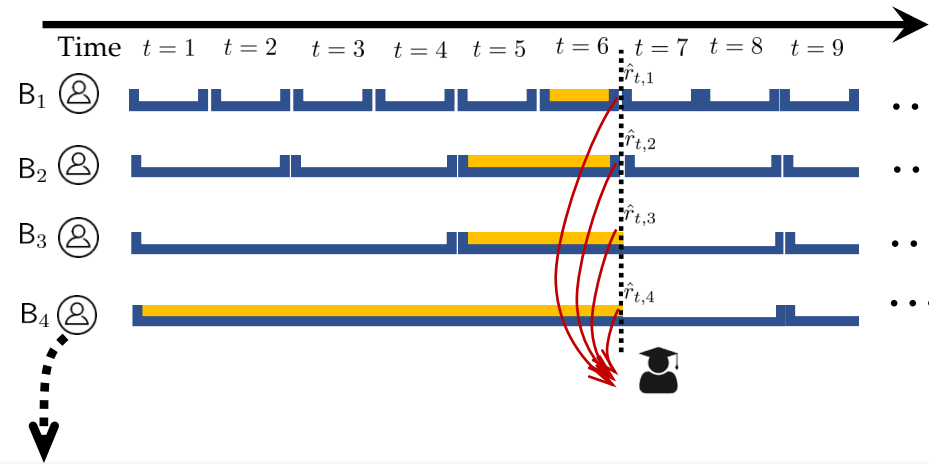
- How to estimate the time-varying density ratio?

$$r_t(\mathbf{x}) = \frac{\mathcal{D}_t(\mathbf{x})}{\mathcal{D}_0(\mathbf{x})} \Rightarrow \mathcal{D}_t(\mathbf{x}) \text{ is accessible with unlabeled online data}$$

- learn with **single-round** data $\Rightarrow |S_t|$ could be very small: **high variance!**
- learn with **all previous data** $\{S_\tau\}_{\tau=1}^t \Rightarrow$ Distribution shift: **high bias!**

We should learn with "**right amount**" of historical data!

Online Density Ratio Estimation



Base-algorithm B_i (person icon)

Update the parameter with data on **different time intervals**

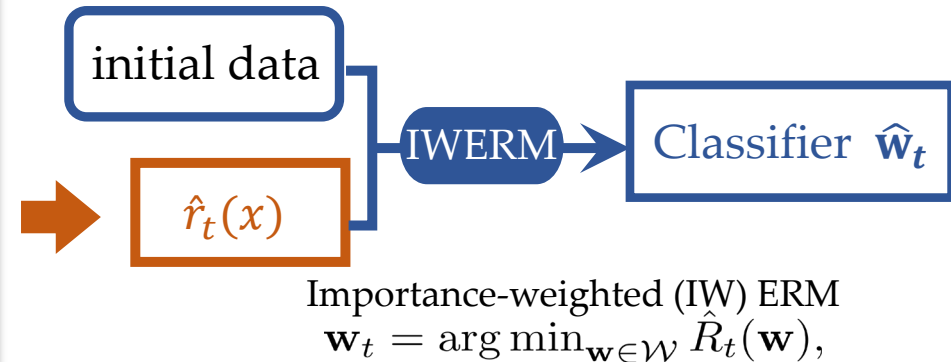
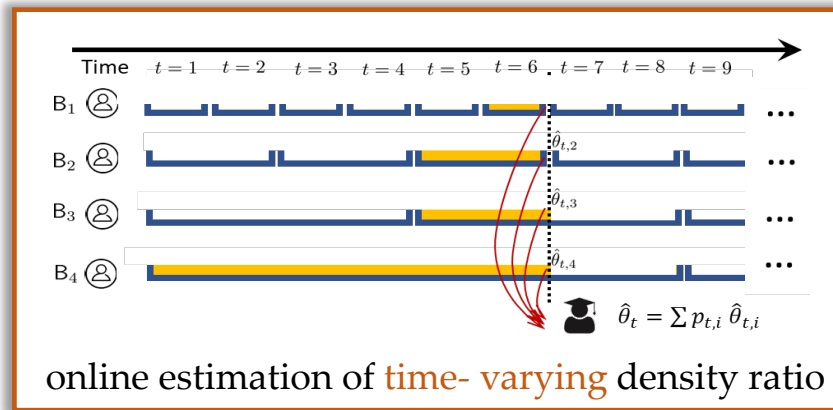
$$\hat{\theta}_{t+1,i} = \Pi_{\Theta}^{A_{t,i}} \left[\hat{\theta}_{t,i} - \gamma A_{t,i}^{-1} \nabla \hat{L}_t(\hat{\theta}_{t,i}) \right],$$

where $\hat{L}_t(\theta)$ is the loss function only established on S_t and S_0 .

Meta-algorithm (graduation cap icon)

Aggregate the base-algorithms model by wei $\hat{\theta}_t = \sum_{i \in K} p_{t,i} \hat{\theta}_{t,i}$.

Overall Algorithms



Theorem. Use the logistic regression-based density ratio estimation model.

With probability at least $1 - \delta$, running **IWERM** with the **estimated density ratio** function $\hat{r}_t(\mathbf{x})$ yields

$$\text{Goal } \frac{1}{T} \left(\sum_{t=1}^T R_t(\mathbf{w}_t) - \sum_{t=1}^T R_t(\mathbf{w}_t^*) \right) \leq \tilde{\mathcal{O}} \left(N_0^{-\frac{1}{2}} + \max \left\{ T^{-\frac{1}{3}} V_T^{\frac{1}{3}}, T^{-\frac{1}{2}} \right\} \right)$$

similar bound as the label shift case

where $V_T = \sum_{t=2}^T \|\mathcal{D}_t(\mathbf{x}) - \mathcal{D}_{t-1}(\mathbf{x})\|_1$ is the variation of input densities

Experiments

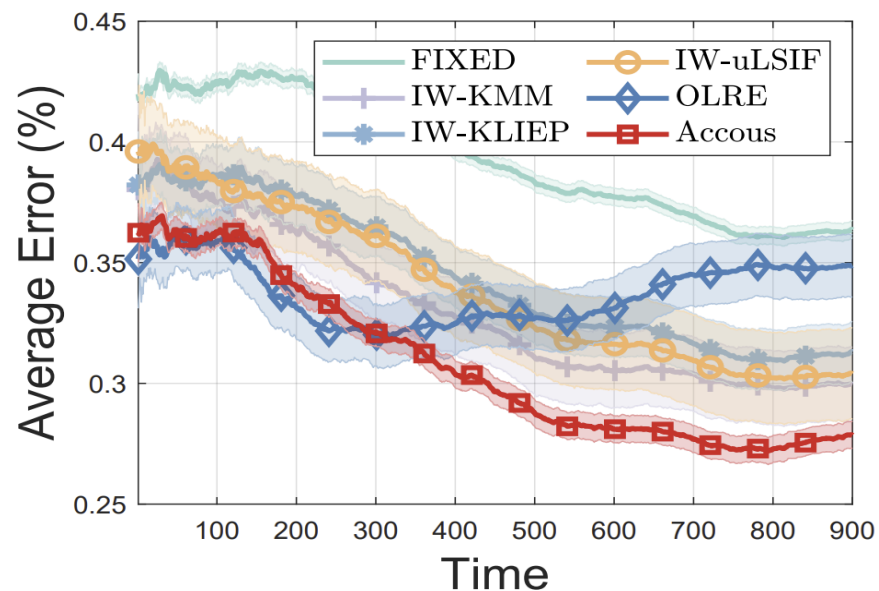
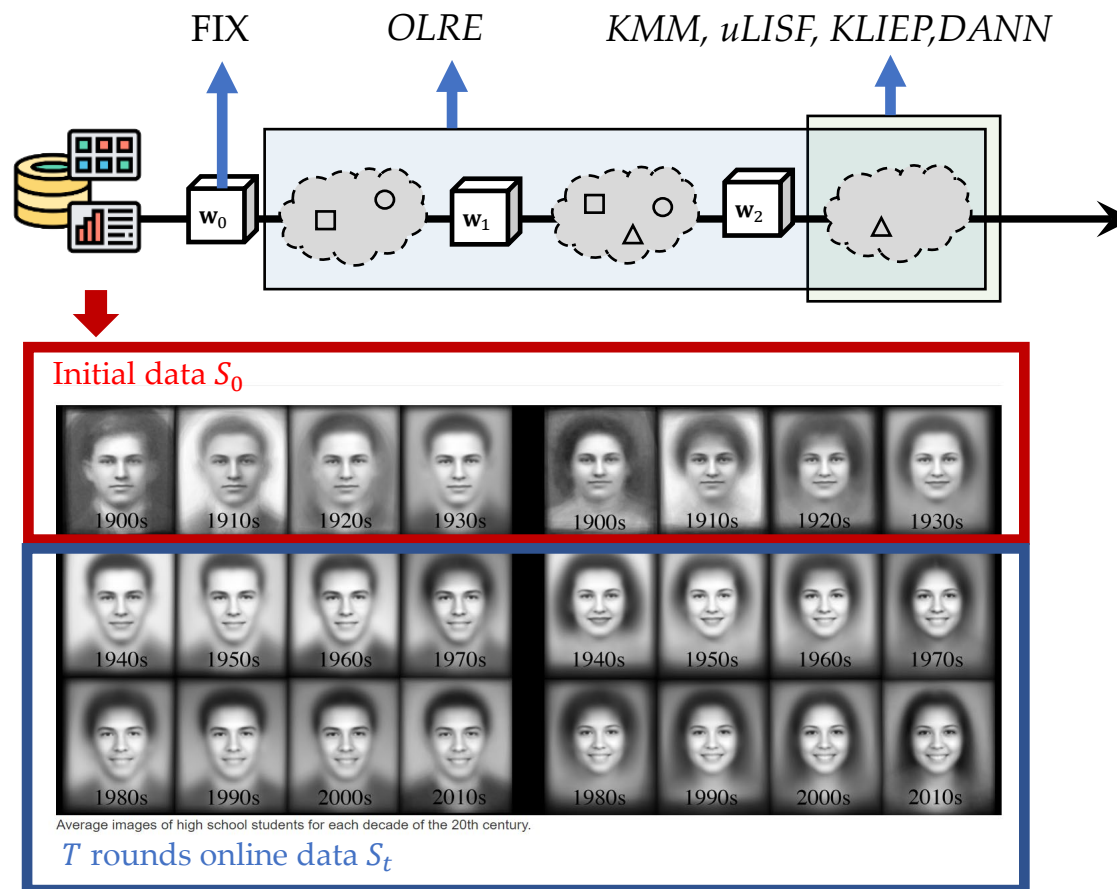
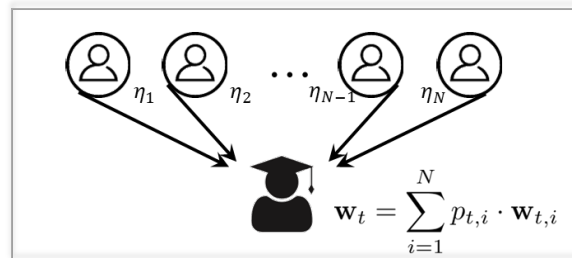


Figure 4.5: Average error on Yearbook dataset with real-life covariate shift



Online Label Shift vs Covariate Shift

- Same algorithmic principle: the **online ensemble** framework
- Different instantiations
 - online label shift: ensembling base learners with *different step sizes*

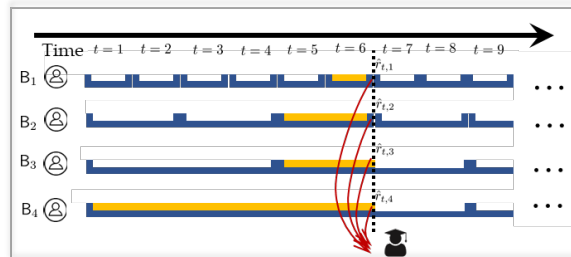


base-algorithm (Ⓜ)

Multiple OGDs learning with **different step sizes**

$$\mathbf{w}_{t+1,i} = \Pi_{\mathcal{W}} \left[\mathbf{w}_{t,i} - \eta_i \nabla \hat{R}_t(\mathbf{w}_{t,i}) \right].$$

- online covariate shift: ensembling base learners with *different time intervals*



base-algorithm (Ⓜ)

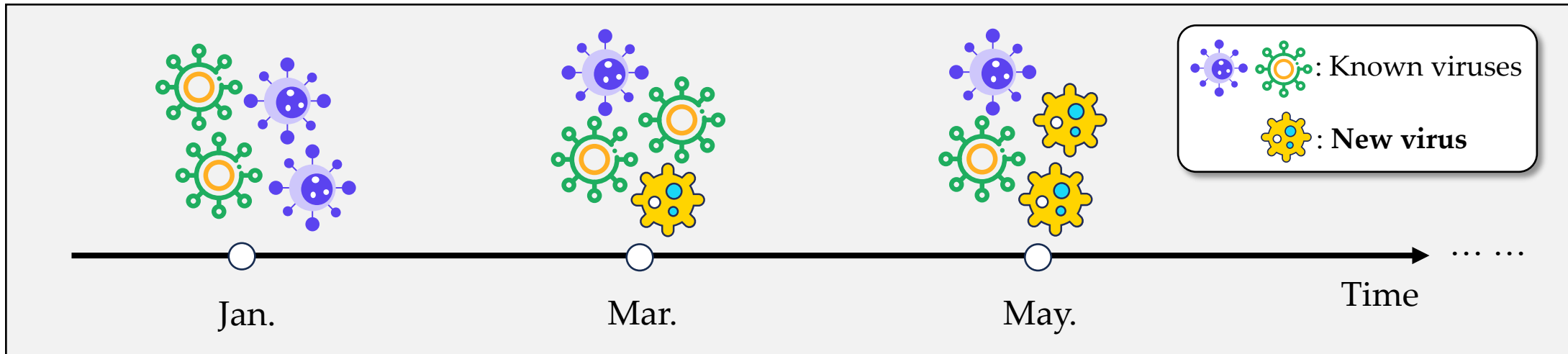
Update parameter with data on **different time intervals**

$$\hat{\theta}_{t+1,i} = \Pi_{\Theta}^{A_{t,i}} \left[\hat{\theta}_{t,i} - \gamma A_{t,i}^{-1} \nabla \hat{L}_t(\hat{\theta}_{t,i}) \right],$$

where $\hat{L}_t(\theta)$ is the loss function only established on S_t and S_0 .

More extensions: New Class

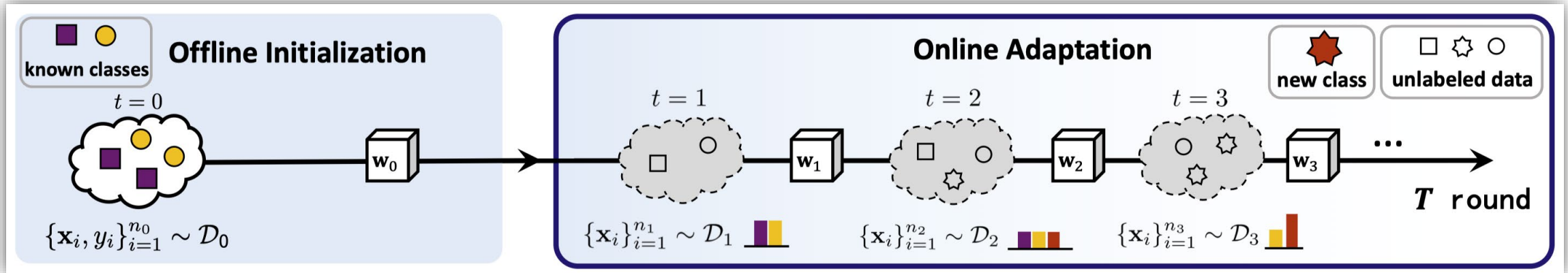
- We investigate New class in Online Label Shift (N-OLS):



Example: a disease diagnosis task

➤ *label distribution* changes; ➤ *new class* data appear;
These above two challenges may take place *simultaneously*.

Problem Formulation



Two challenges take place *simultaneously*, with only *unlabeled* data.

- *label distribution* changes;
- *new class* data appear.

Unbiased Risk Estimator

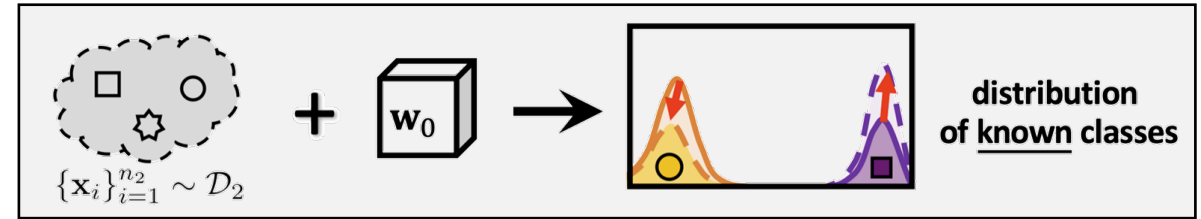
$$\hat{R}_t(\mathbf{w}) = \hat{\theta}_t \sum_{j=1}^K [\hat{\mu}_{y_t}]_j R_0^j(\mathbf{w}) + \mathbb{E}_{S_t}[\ell(\mathbf{w}; \text{nc})] - \hat{\theta}_t \sum_{j=1}^K [\hat{\mu}_{y_t}]_j \mathbb{E}_{S_0^j}[\ell(\mathbf{w}; \text{nc})].$$

- Estimating Proportion for Known Classes Data $\hat{\mu}_{y_t}$:

$$\hat{\mu}_{y_t} = C_0^{-1} \cdot \hat{\mu}_{\hat{y}_t}$$

(C_0 is misclassification matrix of initial model \mathbf{w}_0)

($\hat{\mu}_{\hat{y}_t}$ is predictions of \mathbf{w}_0 on unlabeled data S_t)

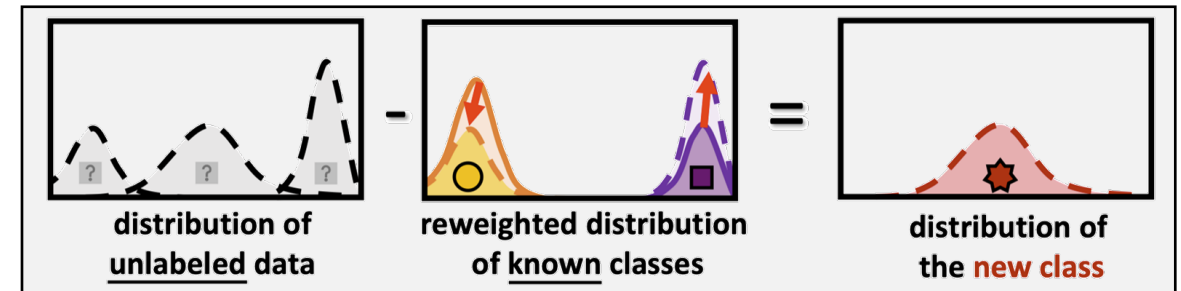


Black Box Shift Estimator (BBSE)

- Estimating Proportion for New Classes Data $\hat{\theta}_t$:

$$\hat{c} = \arg \min_{c \in [0,1]} \frac{q_u(c)}{q_p(c)} + \frac{1+\gamma}{q_p(c)} \left(\sqrt{\frac{\log(4/\delta)}{2S_{\text{win}}}} + \sqrt{\frac{\log(4/\delta)}{2S_0}} \right)$$

$$\Rightarrow \hat{\theta}_t = q_u(\hat{c}) / q_p(\hat{c})$$



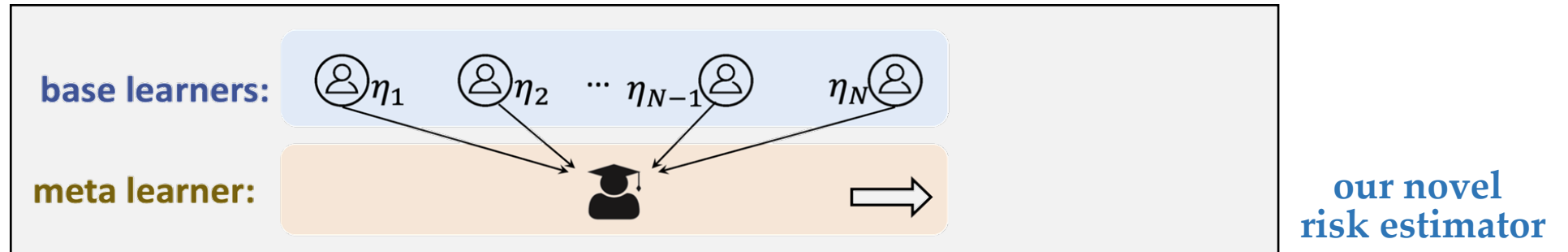
Sliding-window Mixture Proportion Estimation

Leverage *online unlabeled* and *offline labeled* data to estimate distribution.

Deploying Online Ensemble

Next: explore *online ensemble* to **robustly** update the model;

- Maintain multiple learning rates;
- Combine multiple models with various learning rates in a weighted fashion.



- Each **base learner** excel in handling different shift intensities: $\mathbf{w}_{t+1}^i = \Pi_{\mathcal{W}} \left[\mathbf{w}_t^i - \eta_i \nabla \hat{R}_t(\mathbf{w}_t^i) \right]$
- A **meta learner** is employed to obtain the final model: $\mathbf{w}_t = \sum_{i=1}^N p_t^i \cdot \mathbf{w}_t^i$

Very few (logarithmic order) base learners is needed to achieve robustness.

Theoretical Guarantee

The performance measure is the *dynamic regret*:

$$\text{D-Regret}_T \triangleq \sum_{t=1}^T R_t(\mathbf{w}_t) - \sum_{t=1}^T R_t(\mathbf{w}_t^*)$$

i.e., difference between *cumulative expected risk* of predictive models $\{\mathbf{w}_t\}_{t=1}^T$ and $\{\mathbf{w}_t^*\}_{t=1}^T$.

Theorem 1. Suppose the loss function is convex w.r.t. any model $\mathbf{w} \in \mathcal{W}$, and confusion matrix C_0 is invertible. Set the step size pool as $\mathcal{H} = \{\eta_i = \frac{\sigma\Gamma}{2G\sqrt{(K+1)T}} \cdot 2^{i-1} \mid i \in [N]\}$, where $N = 1 + \lceil \frac{1}{2} \log_2(1 + 2T) \rceil$ is the number of base-learners. Our method ensures that

$$\mathbb{E} [\text{Reg}_T^{\mathbf{d}}] \leq \mathcal{O} \left(V_T^{1/3} T^{2/3} \right)$$

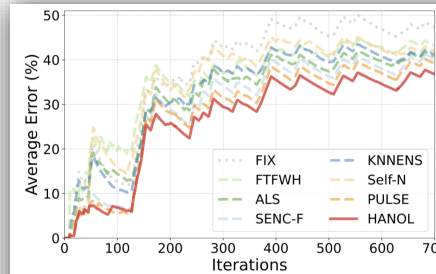
where $V_T = \sum_{t=2}^T \|\mathcal{D}_t(y) - \mathcal{D}_{t-1}(y)\|_1$ measures the intensity of label shift.

Experiments

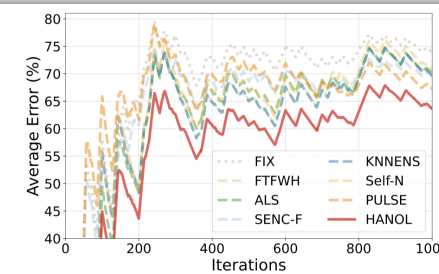
Two real-world applications: human locomotion & satellite image

TABLE II: Average error (%) of different algorithms on the real-world applications of SHL [14] and fMoW [15] datasets. The performance metrics reported include both the mean accuracy and the standard deviation of different algorithms over five separate runs.

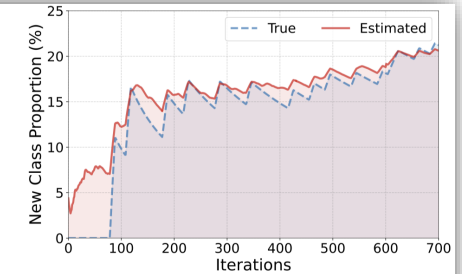
	FIX	FTFWH	ASL	SENC-F	KNNENS	Self-N	PULSE	HANOL
SHL	47.32 ±1.05	43.21 ±1.67	40.78 ±1.42	40.22 ±1.55	41.23 ±1.81	41.25 ±1.12	38.19 ±1.61	36.81 ±1.32
fMoW	73.15 ±3.31	69.38 ±2.64	69.54 ±2.13	68.87 ±3.34	69.23 ±1.81	70.37 ±2.84	66.32 ±2.71	63.16 ±3.01



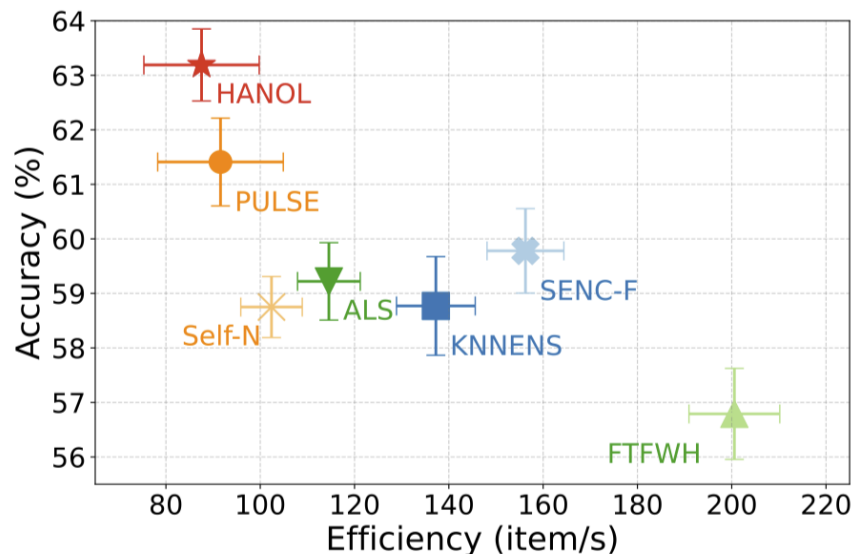
(a) accuracy curve on SHL



(b) accuracy curve on fMoW



(c) new class estimation on SHL



Validate the *efficiency*:





Albeit with a *slight compromise on efficiency* (owing to ensemble), our method attains the *best* performance.

Conclusion

- **Online Ensemble**: an effective theoretical framework (base learners; meta learners; schedule) to handle *uncertainty* in online environments
- **Non-stationary online learning**: online ensemble for dynamic regret
 - build on online convex optimization, optimal dynamic regret guarantees
 - Online Label Shift, Online Covariate Shift, New Classes
 - other results: online RL, online control, game theory, etc.
- Beyond non-stationarity: a general framework to handle uncertainty.
- Many todo: efficiency? continual learning? Unlearning? ...

Thanks!

References

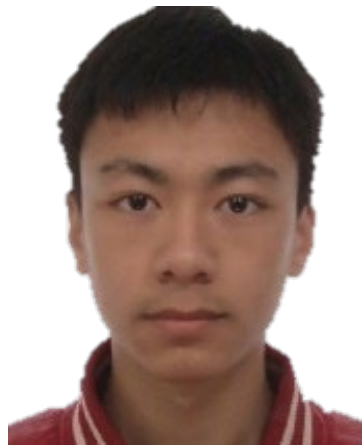
-  Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and Non-stationarity: Problem-dependent Dynamic Regret for Online Convex Optimization. *Journal of Machine Learning Research (JMLR)*, 2024.
(online ensemble)
-  Yong Bai, Yu-Jie Zhang, Peng Zhao, Masashi Sugiyama, and Zhi-Hua Zhou. Adapting to Online Label Shift with Provable Guarantees. **NeurIPS 2022**.
-  Yu-Jie Zhang, Zhen-Yu Zhang, Peng Zhao, and Masashi Sugiyama. Adapting to Continuous Covariate Shift via Online Density Ratio Estimation. **NeurIPS 2023**.
-  Yu-Yang Qian, Yong Bai, Zhen-Yu Zhang, Peng Zhao, and Zhi-Hua Zhou. Handling New Class in Online Label Shift. **ICDM 2023**.

Thanks!

Joint works with



Yong Bai
(NJU → Kuaishou)



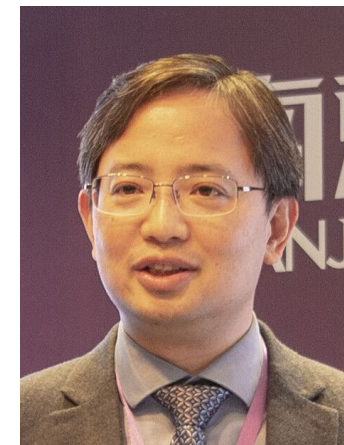
Yu-Jie Zhang
(NJU → U. Tokyo)



Yu-Yang Qian
(NJU)



Masashi Sugiyama
(U. Tokyo)



Zhi-Hua Zhou
(NJU)

端午安康



Thanks!