An adaptive transfer learning perspective on classification in non-stationary environments

*—— **Non-parametric Regression** for Online Label Shift*

Henry W J Reeve

Lecturer in Statistical Science, University of Bristol

Presented by **Yu-Yang Qian**

2024.9.12

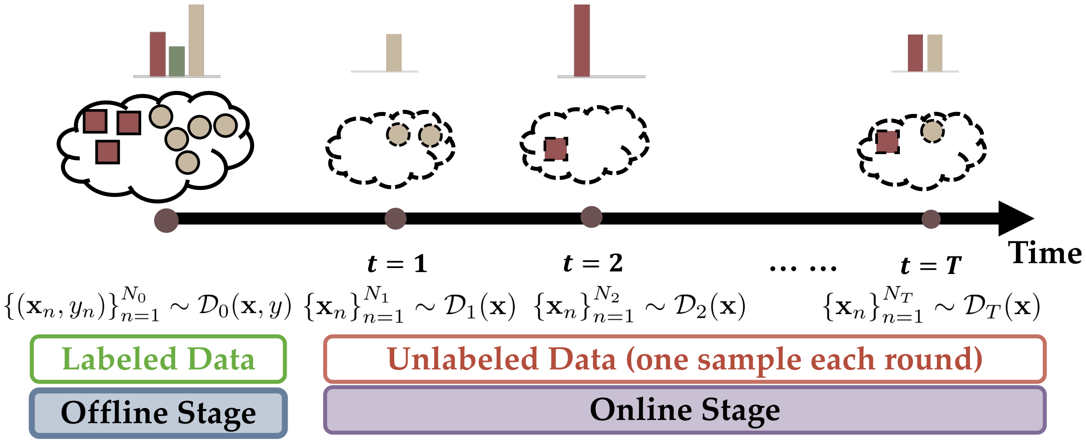
Outline

- Problem Formulation
- Background
- Method
  - Step 1: Reduce Regret Minimization to Parameter Estimation
  - Step 2: Parameter Estimation using ***Lepski method***
- Theory:
  - ***Dynamic Regret:*** Interval Regret + Number of Intervals
- Conclusion

Outline

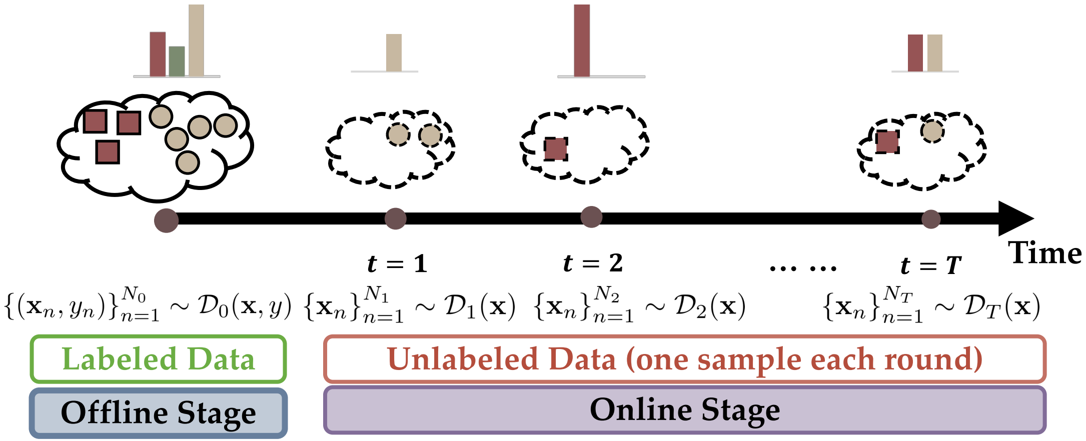
- Problem Formulation
- Background
- Method
  - Step 1: Reduce Regret Minimization to Parameter Estimation
  - Step 2: Parameter Estimation using ***Lepski method***
- Theory:
  - ***Dynamic Regret:*** Interval Regret + Number of Intervals
- Conclusion

Problem Formulation



- Two stages: offline & online
  - Offline stage: a lot of labeled data
  - Online stage: unlabeled data stream (one sample each round)

Problem Formulation



- Assumption 1: Online Label shift (OLS)

  - The conditional $\mathcal{D}_t(\mathbf{x} \mid y)$ is *identical* throughout the process.
  - $\mathcal{D}_0(y) > 0$ for any $y \in \mathcal{Y}$.

Problem Formulation

Assumption 1: Online Label shift (OLS)

- The conditional $\mathcal{D}_t(\mathbf{x} \mid y)$ is *identical* throughout the process.
- $\mathcal{D}_0(y) > 0$ for any $y \in \mathcal{Y}$.

Assumption 2: Temporal smoothness (informal)

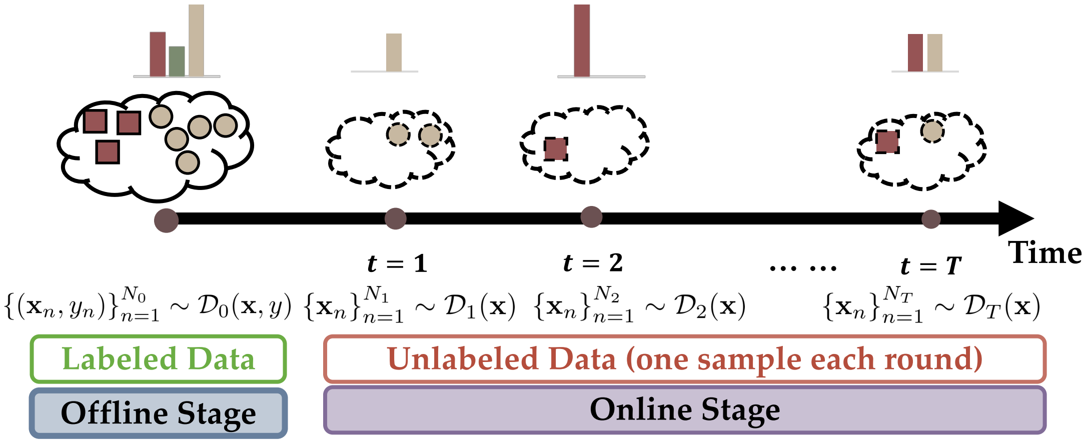
There exists $m \in [t]$, a smooth function g and a smoothness parameter λ such that for all $\ell \in \{t-m, \ldots, t\}$,

$$|\pi_t - \pi_{t-l}| \leq \lambda \qquad \text{where } \pi_t \triangleq \Pr(y_t = 1).$$

i.e., the label prior changes ***smoothly*** in the interval I = [t - m, t].

Certainly holds for $\lambda = V_{\mathcal{I}} = \sum_{\tau=t-m}^{t} |\pi_\tau - \pi_{\tau-1}|$

Problem Formulation



- Goal: learn a classifier $\widehat{h}_t : \mathcal{X} \mapsto \{0, 1\}$ to minimize dynamic regret:

$$\mathbf{Reg}_T^{\mathbf{d}}(\{R_t, h_t^\star\}_{t=1}^T) \triangleq \sum_{t=1}^T R_t(\widehat{h}_t) - \sum_{t=1}^T R_t(h_t^\star)$$

where $R_t(h) = \Pr(h(\mathbf{x}_t) \neq y_t)$.

Previous method for OLS

- ***ATLAS*** [Bai et al., 2022]: Unbiased estimator + Online Ensemble:
  - Construct unbiased gradient, then online gradient descent
- ***Restart-based*** [Qian et al., 2024]: Previous Online Algorithm + *Explicit Partition*:
  - Achieve a good regret in **stationary** interval, ***restart*** to find optimal partition
- ***FLH-FTL*** [Baby et al., 2023]: Adaptive Regret + *Implicit Partition* (only in analysis):
  - Achieve a good regret in **stationary** interval, ***implicitly*** find optimal partition

⇨ All of them achieving ***optimal*** dynamic regrets of $\mathcal{O}(T^{\frac{2}{3}} V_T^{\frac{1}{3}})$.

where $V_T = \sum_{t=2}^{T} |\pi_t - \pi_{t-1}|$ is the class-prior variation.

This paper: A totally statistical way

- All three method somehow need “models” for classification.
  - Need to impose prior on model structure
  - e.g., linear model, logistic regression, reweighting the initial model

- This paper takes a totally ***different*** view:
  - ***Non-parametric Regression*** to solve Online Label Shift
  - A totally ***statistical*** method

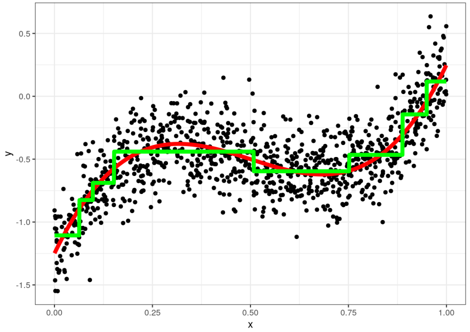
- In the end, we will see how this method ***coincident*** with adaptive reg.

Outline

- Problem Formulation
- Background
- Method
  - Step 1: Reduce Regret Minimization to Parameter Estimation
  - Step 2: Parameter Estimation using ***Lepski method***
- Theory:
  - ***Dynamic Regret:*** Interval Regret + Number of Intervals
- Conclusion

Some History Bits: Non-parametric Regression

- Non-parametric Regression: estimating ***certain parameters***



  - An offline setting: observe a batch of sampled data.

Underlying (unknown) oracle labels: $\theta_i = \mathbb{E}(Y \mid X = X_i)$. For simplicity, $X_i = i$.

We can only obeserve i.i.d. samples $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \ldots, T$,

$$Y_i \sim \mathcal{D}_i, \quad i = 1, \ldots, T$$

Goal: minimize ground-truth distributions' loss: $\min \sum_{i=1}^{T} \mathbb{E}_{Y \sim \mathcal{D}_i} (f(X_i) - Y)^2$

Some History Bits: Non-parametric Regression?

Underlying (unknown) oracle labels: $\theta_i = \mathbb{E}(Y \mid X = X_i)$. For simplicity, $X_i = i$.

We can only obeserve i.i.d. samples $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \ldots, T,$

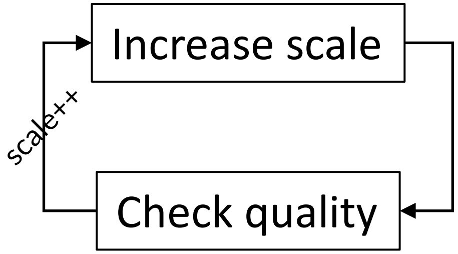
$$Y_i \sim \mathcal{D}_i, \quad i = 1, \ldots, T$$

Goal: minimize ground-truth distributions' loss: $\min \sum_{i=1}^{T} \mathbb{E}_{Y \sim \mathcal{D}_i} (\boxed{f(X_i)} - Y)^2$

- *f* can be KNN, spline, local smoothing, historical averaging...
- Still have 'parameter' in the model, why called non-parametric?
  - 'non-parametric' does not mean we use no parameters in prediction model
  - but mean we impose no assumption on data generation
  - OTOH, parametric regression explicitly assume generation function on Y_i

History of Lepski method

- Lepski method is first introduced in 1997 [Lepski and Spokoiny, 1997]
- Goal: handle ***non-parametric regression*** tasks.
- Key steps (select the optimal trade-off between bias and variance):
  - Increasing the estimator's ***scale*** (e.g., use more historical data for average, more samples for KNN)
  - The selection of the ***optimal scale*** is by checking the error bar (e.g., check its error on current data)
  - Once the error bar is ***reached***, stop and output



History of Lepski method

- Lepski method is first introduced in 1997 [Lepski and Spokoiny, 1997]
- Goal: handle ***non-parametric regression*** tasks.
- It is now be applied for many areas, especially ***transfer learning***
  - Mostly offline (one-stage) transfer learning
  - For example, [Cai and Wei, 2019]: weighted KNN (of increasing source data size) for transfer learning
- This paper seems to be the first to apply it to online learning problem.

Outline

- Problem Formulation
- Background
- Method
  - Step 1: Reduce Regret Minimization to Parameter Estimation
  - Step 2: Parameter Estimation using ***Lepski method***
- Theory:
  - ***Dynamic Regret:*** Interval Regret + Number of Intervals
- Conclusion

Method Overview

- Build entirely upon statistical method:
  - Reduce the Regret Minimization to Parameter Estimation
  - Estimate density ratio by ***non-parametric regression***
  - Prove that algorithm can perform well in any ***stationary intervals***
  - Find the optimal partition ***implicitly*** (only in analysis)

$$\mathbf{Reg}_T^{\mathbf{d}} = \sum_{i=1}^{J} \mathbf{Reg}_{\mathcal{I}_i}$$

intervals: $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_J\}$

Step 1: Reduce Regret Minimization to Parameter Estimation

- Reduce the risk to:

Lemma 1. Suppose $h : \mathcal{X} \to \{0, 1\}$ is a classifier. Then,

$$R_t(h) = \pi_t + \int h(1 - \pi_t - \eta) d(\nu_0 + \nu_1)$$

where $\pi_t = \Pr(y_t = 1), \nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0), \nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1), \eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x}) + \nu_1(\mathbf{x})}$

$$\Longrightarrow \quad h_t^\star(\mathbf{x}) = \mathbb{1}\{1 - \eta(\mathbf{x}) - \pi_t < 0\} = \mathbb{1}\{\eta(\mathbf{x}) + \pi_t > 1\}$$

To this end, we reduce instantaneous risk into a *parameter estimation* problem.

Step 1: Reduce Regret Minimization to Parameter Estimation

$$R_t(h) = R_t(h_t^\star) + \int_{\{h \neq h_t^\star\}} |1 - \pi_t - \eta| \, d(\nu_0 + \nu_1)$$

where $\pi_t = \Pr(y_t = 1), \nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0), \nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1), \eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x}) + \nu_1(\mathbf{x})}$

Proof of Lemma 1. By decomposing $R_t(h)$, we have

$$R_t(h) = \pi_t \int (1-h) d\nu_1 + (1-\pi_t) \int h d\nu_0 = \pi_t + \int h d \left\{(1-\pi_t)\nu_0 - \pi_t \nu_1\right\}$$

$$= \pi_t + 2 \int h \left\{(1-\pi_t)(1-\eta) - \pi_t \eta\right\} d\nu_{1/2} = \pi_t + 2 \int h(1 - \pi_t - \eta) d\nu_{1/2}$$

where $\left(\nu_{1/2} = \frac{\nu_1}{2(\nu_0 + \nu_1)}\right)$

Thus, with $h_t^\star(\mathbf{x}) = \mathbb{1}\left\{\eta(\mathbf{x}) + \pi_t > 1\right\}$ we have

$$R_t(h) - R_t(h_t^\star) = 2 \int (h - h_t^\star)(1 - \pi_t - \eta) d\nu_{1/2} = 2 \int_{\{h \neq h_t^\star\}} |1 - \pi_t - \eta| \, d\nu_{1/2} \qquad \square$$

Step 2.1: Estimating Conditional Distribution

$$\pi_t = \Pr(y_t = 1),$$
$$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0),$$
$$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1),$$
$$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x}) + \nu_1(\mathbf{x})},$$

$$h_t^\star(\mathbf{x}) = \mathbb{1}\left\{\eta(\mathbf{x}) + \pi_t > 1\right\}, \quad h_t(\mathbf{x}) = \mathbb{1}\left\{\widehat{\eta}(\mathbf{x}) + \widehat{\pi}_t > 1\right\},$$

- Next, we estimate conditional distribution $\eta(\mathbf{x})$:
  - 1. Roadmap: build a *confidence interval*;

$$\widehat{U}_b(q, \widetilde{\varepsilon}) \triangleq \frac{8\left(\sqrt{\widetilde{\varepsilon}\sigma^2(q) + \widetilde{\varepsilon}^2} + \{1 - 2q\}\widetilde{\varepsilon}\right)}{3(1 + 2\widetilde{\varepsilon})}$$

  - 2. Use *offline data* to estimate v_0 *and* v_1, therefore estimate η;

$$\widehat{\nu}_{y,\omega,\delta}(A) \triangleq \begin{cases} \widehat{\nu}_y(A) - \widehat{\mathrm{U}}\left(1 - \widehat{\nu}_y(A), n_y, \delta\right) & \text{if } \omega = -1 \\ \widehat{\nu}_y(A) + \widehat{\mathrm{U}}\left(\widehat{\nu}_y(A), n_y, \delta\right) & \text{if } \omega = 1 \end{cases} \qquad \widehat{\eta}(A) \triangleq \left(\left(\frac{\widehat{\nu}_{1,\omega,\delta}(A)}{\widehat{\nu}_{0,-\omega,\delta}(A) + \widehat{\nu}_{1,\omega,\delta}(A)}\right) \vee 0\right) \wedge 1$$

Step 2.1: Estimating Conditional Distribution

$\pi_t = \Pr(y_t = 1)$,
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0)$,
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1)$,
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x})+\nu_1(\mathbf{x})}$,

$$h_t^\star(\mathbf{x}) = \mathbb{1}\left\{\eta(\mathbf{x}) + \pi_t > 1\right\}, \quad h_t(\mathbf{x}) = \mathbb{1}\left\{\widehat{\eta}(\mathbf{x}) + \widehat{\pi}_t > 1\right\},$$

- Next, we estimate conditional distribution $\eta(\mathbf{x})$:
  - 3. Theoretical Guarantee: using *Dvoretzky–Kiefer–Wolfowitz–Massart* type concentration inequality:

Lemma 2. With probability at least $1-\delta$, the estimated conditioal distribution $\widehat{\eta}$ has the following guarantee:

$$|\widehat{\eta}(x) - \eta(x)| \leq 14 \left(\left\{ \theta^{\frac{1}{\beta}} \varepsilon_\delta^{\mathrm{il}}(\breve{n}) \right\}^{\frac{\beta}{2\beta+1}} \vee \sqrt{\varepsilon_\delta^{\mathrm{il}}(\breve{n})} \right) = \widetilde{\mathcal{O}}(\sqrt{\frac{1}{|S_0|}})$$

*Details are omitted, as Step 2.1 is a **standard** statistical learning.*

Step 2.2: Estimating Label Prior by ***Lepski Method***

$\pi_t = \Pr(y_t = 1),$
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0),$
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1),$
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x})+\nu_1(\mathbf{x})},$

$$h_t^\star(\mathbf{x}) = \mathbb{1}\left\{\eta(\mathbf{x}) + \pi_t > 1\right\}, \quad h_t(\mathbf{x}) = \mathbb{1}\left\{\widehat{\eta}(\mathbf{x}) + \widehat{\pi}_t > 1\right\},$$

- Then, we estimate the label prior $\widehat{\pi}_t$:
  - Remember in ATLAS [Bai et al., 2022], BBSE is used to estimate the label prior:

$$\widehat{\pi}_t = C_0^{-1} \cdot \widehat{\pi}_{\widehat{y}_t}$$

where $[C_0]_{i,j} \triangleq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_0(\mathbf{x}|y=j)}\left[\Pr\left(f_0(\mathbf{x}) = i\right)\right], \quad \left[\widehat{\pi}_{\widehat{y}_t}\right]_j = 1/n_t \cdot \sum_{\mathbf{x}\in S_t}\left[f_0(\mathbf{x})\right]_j$

$f_0 : \mathcal{X} \mapsto \Delta_2$ is the offline prediction model.

  - However, there is no ***offline model f_0*** here, we only have offline data.

Step 2.2: Estimating Label Prior by ***Lepski Method***

$\pi_t = \Pr(y_t = 1)$,
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0)$,
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1)$,
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x}) + \nu_1(\mathbf{x})}$,
$\mu_t = \mathcal{D}_t(\mathbf{x})$

$$h_t^\star(\mathbf{x}) = \mathbb{1}\left\{\eta(\mathbf{x}) + \pi_t > 1\right\}, \quad h_t(\mathbf{x}) = \mathbb{1}\left\{\widehat{\eta}(\mathbf{x}) + \widehat{\pi}_t > 1\right\},$$

- Then, we estimate the label prior $\widehat{\pi}_t$:
  - We define $\mu_t = \mathcal{D}_t(\mathbf{x})$, therefore

$$\mu_t = (1 - \pi_t)\nu_0 + \pi_t \nu_1$$

$$\Longrightarrow \quad \pi_t = \frac{\mu_t - \nu_0}{\nu_1 - \nu_0}$$

  - ν_0 and ν_1 has already been estimated before, therefore we ***focus on μ_t***

Step 2.2: Estimating Label Prior by *Lepski Method*

$$\Rightarrow \qquad \pi_t = \frac{\mu_t - \nu_0}{\nu_1 - \nu_0}$$

$\pi_t = \Pr(y_t = 1),$
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0),$
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1),$
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x}) + \nu_1(\mathbf{x})},$
$\mu_t = \mathcal{D}_t(\mathbf{x})$

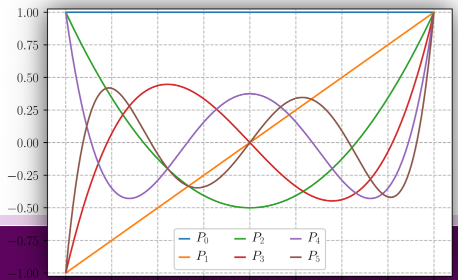
- ν_0 and ν_1 has already been estimated before, therefore we ***focus on μ_t***
  - We estimate μ_t through ***weighted combination*** of history:

$$\widehat{\mu}_t(\mathbf{x}) \triangleq \sum_{i=1}^{q} \underbrace{w(i,q)}_{\text{weight of history}} \cdot \underbrace{\widehat{D}_{t-i}(X = \mathbf{x})}_{\text{history density}}$$

q: history length   weight of history   history density

  - weight of history is defined by *Legendre polynomials* (i.e., predefined *orthogonal bases*):

$$\mathcal{L}_k(z) := \frac{\sqrt{2k+1}}{k!} \frac{d^k}{dz^k} \{z(z-1)\}^k$$



Step 2.2: Estimating Label Prior by *Lepski Method*

$\pi_t = \Pr(y_t = 1),$
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0),$
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1),$
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x}) + \nu_1(\mathbf{x})},$
$\mu_t = \mathcal{D}_t(\mathbf{x})$

$$\Longrightarrow \quad \pi_t = \frac{\mu_t - \nu_0}{\nu_1 - \nu_0}$$

- ν_0 and ν_1 has already been estimated before, therefore we ***focus on*** $\boldsymbol{\mu_t}$
  - We estimate μ_t through ***weighted combination*** of history:

$$\widehat{\mu}_t(\mathbf{x}) \triangleq \sum_{i=1}^{q} \underbrace{w(i, q)}_{\text{weight of history}} \cdot \underbrace{\widehat{D}_{t-i}(X = \mathbf{x})}_{\text{history density}}$$

q: history length   weight of history   history density

  - Key: How to determine the history length *q*?

$\Longrightarrow$ ***Lepski-based*** method

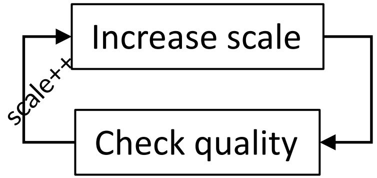
Step 2.2: Estimating Label Prior by ***Lepski Method***

$\pi_t = \Pr(y_t = 1)$,
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0)$,
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1)$,
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x})+\nu_1(\mathbf{x})}$,
$\mu_t = \mathcal{D}_t(\mathbf{x})$

$$\Rightarrow \quad \pi_t = \frac{\mu_t - \nu_0}{\nu_1 - \nu_0} \qquad \widehat{\mu}_t(\mathbf{x}) \triangleq \sum_{i=1}^{q} w(i,q) \cdot \widehat{D}_{t-i}(X = \mathbf{x})$$

- Key: How to determine the history length q?

$\Rightarrow$ ***Lepski-based*** method



- We define the weight's complexity as:

$$\mathbb{S}_{n,\delta}(q) := \|w(q)\|_2 \sqrt{2 \log\left(\frac{\pi^2 q^2}{\delta}\right)} \qquad \text{where } \|w(q)\|_2^2 \triangleq \textstyle\sum_{i=1}^{q} w(i,q)^2.$$

We ***choose*** $\widehat{q}$ as the maximal value of $q \in \left\{8\alpha^2(\alpha+1)^2, \ldots, n\right\}$ such that

$$\left|\widehat{\mu}_n^{q}(f) - \widehat{\mu}_n^{q_b}(f)\right| \leq 2\left\{\mathbb{S}_{n,\delta}(q) + \mathbb{S}_{n,\delta}(q_b)\right\}$$

for all $q_b \in \left\{8\alpha^2(\alpha+1)^2, \ldots, q-1\right\}$. (where α is a constant about temporal smoothness)

Step 2.2: Estimating Label Prior by ***Lepski Method***

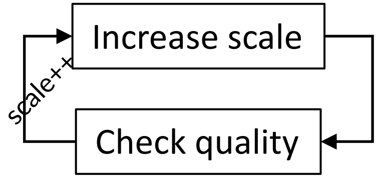
$\pi_t = \Pr(y_t = 1),$
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0),$
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1),$
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x})+\nu_1(\mathbf{x})},$
$\mu_t = \mathcal{D}_t(\mathbf{x})$

$$\Longrightarrow \quad \pi_t = \frac{\mu_t - \nu_0}{\nu_1 - \nu_0} \qquad \widehat{\mu}_t(\mathbf{x}) \triangleq \sum_{i=1}^{q} w(i,q) \cdot \widehat{D}_{t-i}(X = \mathbf{x})$$

We ***choose*** $\widehat{q}$ as the maximal value of $q \in \left\{8\alpha^2(\alpha+1)^2, \ldots, n\right\}$ such that

$$|\widehat{\mu}_n^q(f) - \widehat{\mu}_n^{q_b}(f)| \leq 2\left\{\mathbb{S}_{n,\delta}(q) + \mathbb{S}_{n,\delta}(q_b)\right\}$$

for all $q_b \in \left\{8\alpha^2(\alpha+1)^2, \ldots, q-1\right\}$. (where α is a constant about temporal smoothness)



Corollary 6. Suppose that Assumption 1 and 2 hold. For all $\delta \in (0,1)$ we have

$$\mathbb{P}\left\{ \frac{|\widehat{\pi}_t - \pi_t|\,|(\nu_1 - \nu_0)|}{\mathbb{J}_\delta(\mu_t) + 3\max_{y\in\{0,1\}}\left|(\widehat{\nu}_y^+ - \nu_y)\right|} > 1 \,\middle|\, \mathcal{D}_0, \mathcal{D}_1 \right\} \leq \frac{\delta}{3}$$

$\Longrightarrow$ $|\widehat{\pi}_t - \pi_t| \leq \widetilde{O}\left(\frac{1}{\sqrt{t}} + \lambda\right)$ for t in smooth intervals (Asp. 2)

Step 2.2: Estimating Label Prior by *Lepski Method*

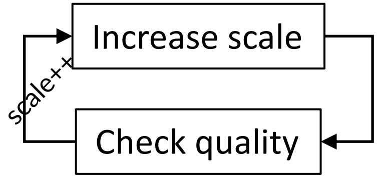
$\pi_t = \Pr(y_t = 1),$
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0),$
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1),$
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x})+\nu_1(\mathbf{x})},$
$\mu_t = \mathcal{D}_t(\mathbf{x})$

$$\Rightarrow \quad \pi_t = \frac{\mu_t - \nu_0}{\nu_1 - \nu_0} \qquad \widehat{\mu}_t(\mathbf{x}) \triangleq \sum_{i=1}^{q} w(i,q) \cdot \widehat{D}_{t-i}(X = \mathbf{x})$$

We ***choose*** $\widehat{q}$ as the maximal value of $q \in \left\{8\alpha^2(\alpha+1)^2, \ldots, n\right\}$ such that

$$|\widehat{\mu}_n^q(f) - \widehat{\mu}_n^{q_b}(f)| \le 2\left\{\mathbb{S}_{n,\delta}(q) + \mathbb{S}_{n,\delta}(q_b)\right\}$$

for all $q_b \in \left\{8\alpha^2(\alpha+1)^2, \ldots, q-1\right\}$. (where α is a constant about temporal smoothness)



Assumption 2. There exists $m \in [t]$, a smooth function g and a smoothness parameter λ such that for all $\ell \in \{t-m, \ldots, t\}$,

$$|\pi_t - \pi_{t-l}| \le \lambda \qquad \text{where } \pi_t \triangleq \Pr(y_t = 1).$$

$$\Rightarrow \quad |\widehat{\pi}_t - \pi_t| \le \widetilde{O}\left(\frac{1}{\sqrt{t}} + \lambda\right)$$ for t in smooth intervals (Asp. 2)

Step 2.2: Estimating Label Prior by ***Lepski Method***

$\pi_t = \Pr(y_t = 1),$
$\nu_0 = \mathcal{D}(\mathbf{x} \mid y = 0),$
$\nu_1 = \mathcal{D}(\mathbf{x} \mid y = 1),$
$\eta(\mathbf{x}) = \frac{\nu_1(\mathbf{x})}{\nu_0(\mathbf{x}) + \nu_1(\mathbf{x})},$
$\mu_t = \mathcal{D}_t(\mathbf{x})$

Corollary 6. Suppose that Assumption 1 and 2 hold. For all $\delta \in (0,1)$ we have

$$\mathbb{P}\left\{ \frac{|\widehat{\pi}_t - \pi_t| \, |(\nu_1 - \nu_0)|}{\mathbb{J}_\delta(\mu_t) + 3 \max_{y \in \{0,1\}} \left|\left(\widehat{\nu}_y^+ - \nu_y\right)\right|} > 1 \,\middle|\, \mathcal{D}_0, \mathcal{D}_1 \right\} \le \frac{\delta}{3}$$

$$\Longrightarrow \quad |\widehat{\pi}_t - \pi_t| \le \widetilde{O}\left(\frac{1}{\sqrt{t}} + \lambda\right) \text{ for } t \text{ in smooth intervals (Asp. 2)}$$

Proof Sketch. (1) Decompose history distribution using *a set of orthogonal bases* (Legendre polynomials) using *Tayler expansion*

(2) The historical length is tunned to be ***optimal*** so that it achieve the minimal error (estimated bias + variance)

If interested, you can check the paper's Section 9…

Outline

- Problem Formulation
- Background
- Method
  - Step 1: Reduce Regret Minimization to Parameter Estimation
  - Step 2: Parameter Estimation using ***Lepski method***
- Theory:
  - ***Dynamic Regret:*** Interval Regret + Number of Intervals
- Conclusion

Dynamic Regret

Remember: $h_t^\star(\mathbf{x}) = \mathbb{1}\left\{\eta(\mathbf{x}) + \pi_t > 1\right\}, \quad h_t(\mathbf{x}) = \mathbb{1}\left\{\widehat{\eta}(\mathbf{x}) + \widehat{\pi}_t > 1\right\},$

$$\Longrightarrow \quad R_t(h) - R_t(h_t^\star) \lesssim \mathcal{O}(|\eta(\mathbf{x}_t) - \widehat{\eta}(\mathbf{x}_t)| + |\pi_t - \widehat{\pi}_t|)$$

$$\leq \underbrace{\frac{1}{\sqrt{|S_0|}}}_{\text{Error of Step 2.1}} + \underbrace{\left(\frac{1}{\sqrt{t}} + \lambda\right)}_{\text{Error of Step 2.2}}$$

Therefore, for an interval $\mathcal{I}_i = [s, e]$ that satisfy Assumption 2 with λ:

$$\mathbf{Reg}_{\mathcal{I}_i} = \sum_{t=s}^{e} R_t(h) - \sum_{t=s}^{e} R_t(h_t^\star) = \frac{|\mathcal{I}_i|}{\sqrt{|S_0|}} + \underbrace{\sum_{t=s}^{e} \frac{1}{\sqrt{t-s}}}_{(\leq 2\sqrt{|\mathcal{I}_i|} - 1)} + \lambda|\mathcal{I}_i|$$

Dynamic Regret: sum of interval regret

$$\mathbf{Reg}_{\mathcal{I}_i} \leq \widetilde{\mathcal{O}}\left(\frac{|\mathcal{I}_i|}{\sqrt{|S_0|}} + \lambda|\mathcal{I}_i| + \sqrt{|\mathcal{I}_i|}\right)$$

Suppose there are J intervals: $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_J\}$

$$\begin{aligned}
\mathbf{Reg}_T^{\mathbf{d}} = \sum_{i=1}^{J} \mathbf{Reg}_{\mathcal{I}_i} \quad &= \frac{T}{\sqrt{|S_0|}} + \lambda T + \sum_{i=1}^{J} \sqrt{|\mathcal{I}_i|} \\
&\leq \frac{T}{\sqrt{|S_0|}} + \lambda T + \sqrt{J}\left(\sum_{i=1}^{J}\left(\sqrt{|\mathcal{I}_i|}\right)^2\right)^{1/2} \\
&= \frac{T}{\sqrt{|S_0|}} + \lambda T + \sqrt{JT}
\end{aligned}$$

Dynamic Regret: control number of intervals J

$$\mathbf{Reg}_T^{\mathbf{d}} = \sum_{i=1}^{J} \mathbf{Reg}_{\mathcal{I}_i} = \frac{T}{\sqrt{|S_0|}} + \lambda T + \sqrt{JT}$$

Consider the simple case of same λ in all intervals:

Assumption 2. There exists $m \in [t]$, a smooth function g and a smoothness parameter λ such that for all $\ell \in \{t-m, \ldots, t\}$,

$$|\pi_t - \pi_{t-l}| \leq \lambda \qquad \text{where } \pi_t \triangleq \Pr(y_t = 1).$$

Asp. 2 certainly holds for $\lambda = V_{\mathcal{I}_i} = \sum_{t=s}^{e} |\pi_t - \pi_{t-1}|$, therefore we choose $\lambda \leq V_{\mathcal{I}_i}$

$$\lambda J \leq \sum_{j=1}^{J} V_{\mathcal{I}_j} = V_T \quad \Longrightarrow \quad J \leq \frac{V_T}{\lambda}$$

Dynamic Regret: finish the proof

$$\mathbf{Reg}_T^{\mathbf{d}} = \sum_{i=1}^{J} \mathbf{Reg}_{\mathcal{I}_i} \leq \frac{T}{\sqrt{|S_0|}} + \lambda T + \sqrt{\frac{V_T T}{\lambda}}$$

Finally, tune λ to get the regret bound:

$$\begin{aligned}
\mathbf{Reg}_T^{\mathbf{d}} = \sum_{i=1}^{J} \mathbf{Reg}_{\mathcal{I}_i} &\leq \frac{T}{\sqrt{|S_0|}} + \lambda T + \frac{1}{2}\sqrt{\frac{V_T T}{\lambda}} + \frac{1}{2}\sqrt{\frac{V_T T}{\lambda}} \\
&\leq \mathcal{O}\left(\frac{T}{\sqrt{|S_0|}} + (V_T T^2)^{1/3}\right) \qquad \left(\lambda = V_T^{1/3} T^{-1/3}, \quad J = T^{1/3} V_T^{2/3}\right) \\
&\leq \mathcal{O}\left(\max\left(\frac{T}{\sqrt{|S_0|}} + (V_T)^{1/3} T^{2/3}, \sqrt{T}\right)\right) \qquad (J \geq 1) \qquad \square
\end{aligned}$$

Compared with ***FLH-FTL*** [Baby et al., 2023]:

- ☐ Get a good regret in any interval (by Lepski method)
- ☐ Implicit find the optimal partition (only in analysis)

- ☐ Get a good regret in any interval (by Strongly Adaptive Methods)
- ☐ Implicit find the optimal partition (only in analysis)

Getting the adaptive regret using non-parametric method!

If fact: if check partition in this paper:

$$\begin{cases} \text{number of intervals: } J = T^{1/3} V_T^{2/3} \\ \text{Regret in each interval: } \mathcal{O}(\sqrt{\mathcal{I}_i} + 1) \end{cases}$$

Matches with Baby's best partition!

Lemma 5 *(key partition) Initialize $\mathcal{Q} \leftarrow \Phi$. Starting from time 1, spawn a new bin $[i_s, i_t]$ whenever $\sum_{j=i_s+1}^{i_t+1} |u_j - u_{j-1}| > B/\sqrt{n_i}$, where $n_i = i_t - i_s + 2$. Add the spawned bin $[i_s, i_t]$ to $\mathcal{Q}$. Consider the following post processing routine.*

1. *Initialize $\mathcal{P} \leftarrow \Phi$.*
2. *For $i \in [|\mathcal{Q}|]$:*
   - *if $u_{i_t} = u_{i_t+1}$:*
     - *(a) Let p be the largest time point with $u_{p:i_t}$ being constant and let q be the smallest time point with $u_{i_t+1:q}$ being constant.*
     - *(b) Add bin $[i_s, p-1]$ to $\mathcal{P}$.*
     - *(c) If $(i+1)_t > q$ then add $[p, q]$ to $\mathcal{P}$ and set $(i+1)_s \leftarrow q+1$.*
     - *(d) Goto Step 2.*
   - *Add $[i_s, i_t]$ to $\mathcal{P}$. Goto Step 2.*

Let $M := |\mathcal{P}|$. We have $\boxed{M = O\left(1 \vee n^{1/3} C_n^{2/3} B^{-2/3}\right)}$. Further for any bin $[i_s, i_t] \in \mathcal{P}$, it holds that $\sum_{j=i_s+1}^{i_t} |u_j - u_{j-1}| \le B/\sqrt{n_i}$ where $n_i = i_t - i_s + 1$.

Compared with ***FLH-FTL*** [Baby et al., 2023]:

- **Pros**
  - No diameter assumption
  - Truly non-parametric, parameter-free

- **Cons**
  - Very costly: computational and storage is O(T), can hardly be runnable in practice
  - Need to derive specific method for specific problem structure (可能一旦不是OLS/密度估计问题，就很难再有良好保障)

Take home message

- A new view of handling non-stationary environments:
  - A ***non-parametric regression*** method (Lepski method) can also get an *strongly adaptive regret*
  - Use ***implicit partition*** (only in analysis) to get the optimal dynamic regret
- May have a chance to accelerate:
  - Lazy update the summation of history margin distribution
  - A segment tree to get the sum of history intervals
  - May achieve O(log T) complexity each round

Thanks!