Model Assessment and Selection under Temporal Distribution Shift

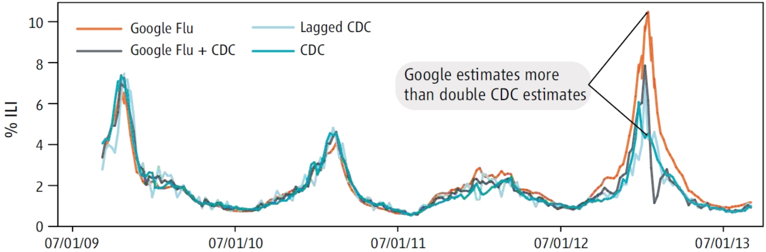
Elise Han, **Chengpiao Huang**, Kaizheng Wang

Columbia University

ICML 2024

Temporal Distribution Shift

Temporal distribution shift can affect model performance:



Lazer et al. (2014). The parable of Google Flu: traps in big data analysis. *Science.*

Problem Setup

Problem Setup

Past						Present
$\mathbb{P}_1$	$\longrightarrow$	$\cdots$	$\longrightarrow$	$\mathbb{P}_{t-1}$	$\longrightarrow$	$\mathbb{P}_t$

Problem Setup

	Past		Present
$\mathbb{P}_1 \longrightarrow$	$\cdots \longrightarrow$	$\mathbb{P}_{t-1} \longrightarrow$	$\mathbb{P}_t$
$\downarrow$		$\downarrow$	$\downarrow$
(x_1, y_1)	$\cdots$	(x_{t-1}, y_{t-1})	(x_t, y_t)

Problem Setup

Past **Present**

$$
\begin{array}{ccccccc}
\mathbb{P}_1 & \longrightarrow & \cdots & \longrightarrow & \mathbb{P}_{t-1} & \longrightarrow & \mathbb{P}_t \\
\downarrow & & & & \downarrow & & \downarrow \\
(x_1, y_1) & & \cdots & & (x_{t-1}, y_{t-1}) & & (x_t, y_t)
\end{array}
$$

Model assessment: Given a model f, estimate expected loss *at present*:

$$L_t(f) = \mathbb{E}_{(x_t, y_t) \sim \mathbb{P}_t} |f(x_t) - y_t|^2.$$

Problem Setup

	Past			**Present**
$\mathbb{P}_1$ $\longrightarrow$	$\cdots$ $\longrightarrow$	$\mathbb{P}_{t-1}$	$\longrightarrow$	$\mathbb{P}_t$
$\downarrow$		$\downarrow$		$\downarrow$
(x_1, y_1)	$\cdots$	(x_{t-1}, y_{t-1})		(x_t, y_t)

Model assessment: Given a model f, estimate expected loss *at present*:

$$L_t(f) = \mathbb{E}_{(x_t, y_t) \sim \mathbb{P}_t} |f(x_t) - y_t|^2.$$

Model selection: Given models $f_1, ..., f_m$, find $\underset{r \in [m]}{\operatorname{argmin}} L_t(f_r)$.

Problem Setup: General Case

Past **Present**

$$\begin{array}{ccccccc} \mathbb{P}_1 & \longrightarrow & \cdots & \longrightarrow & \mathbb{P}_{t-1} & \longrightarrow & \mathbb{P}_t \\ \downarrow & & & & \downarrow & & \downarrow \\ \left\{z_1^{(i)}\right\}_{i=1}^{B_1} & & \cdots & & \left\{z_{t-1}^{(i)}\right\}_{i=1}^{B_{t-1}} & & \left\{z_t^{(i)}\right\}_{i=1}^{B_t} \end{array}$$

Model assessment: Given a model f, estimate expected loss *at present*:

$$L_t(f) = \mathbb{E}_{z_t \sim \mathbb{P}_t} \ell(f, z_t).$$

Model selection: Given models $f_1, ..., f_m$, find $\underset{r \in [m]}{\operatorname{argmin}} L_t(f_r)$.

Problem Setup

Past **Present**

$$\begin{array}{ccccccc} \mathbb{P}_1 & \longrightarrow & \cdots & \longrightarrow & \mathbb{P}_{t-1} & \longrightarrow & {\color{red}\mathbb{P}_t} \\ \downarrow & & & & \downarrow & & \downarrow \\ (x_1, y_1) & & \cdots & & (x_{t-1}, y_{t-1}) & & (x_t, y_t) \end{array}$$

Model assessment: Given a model f, estimate expected loss *at present*:

$$L_{\color{red}t}(f) = \mathbb{E}_{(x_t, y_t) \sim {\color{red}\mathbb{P}_t}} |f(x_t) - y_t|^2.$$

Model selection: Given models $f_1, ..., f_m$, find $\operatorname*{argmin}_{r \in [m]} L_{\color{red}t}(f_r)$.

Model Assessment

Rolling window: Average data from the last k periods:

Model Assessment

Rolling window: Average data from the last k periods:

$$
\begin{array}{ccccccccc}
\mathbb{P}_1 & \longrightarrow & \cdots & \longrightarrow & \mathbb{P}_{t-k+1} & \longrightarrow & \cdots & \longrightarrow & \mathbb{P}_t \\
\downarrow & & & & \downarrow & & & & \downarrow \\
(x_1, y_1) & & \cdots & & (x_{t-k+1}, y_{t-k+1}) & & \cdots & & (x_t, y_t) \\
 & & & & & & \Downarrow & &
\end{array}
$$

$$
\widehat{L}_{t,k}(f) = \frac{1}{k} \sum_{i=t-k+1}^{t} |f(x_i) - y_i|^2
$$

Model Assessment

Rolling window: Average data from the last k periods:

$$
\begin{array}{ccccccccc}
\mathbb{P}_1 & \longrightarrow & \cdots & \longrightarrow & \mathbb{P}_{t-k+1} & \longrightarrow & \cdots & \longrightarrow & \mathbb{P}_t \\
\downarrow & & & & \downarrow & & & & \downarrow \\
(x_1, y_1) & & \cdots & & (x_{t-k+1}, y_{t-k+1}) & & \cdots & & (x_t, y_t) \\
& & & & & & \Downarrow & &
\end{array}
$$

$$
\widehat{L}_{t,k}(f) = \frac{1}{k} \sum_{i=t-k+1}^{t} |f(x_i) - y_i|^2
$$

How to choose k?

Window Selection

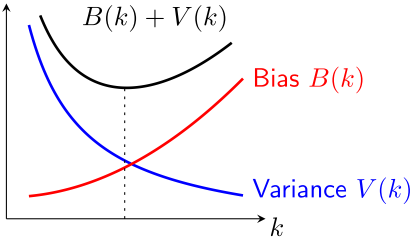
Bias-variance tradeoff: $|\widehat{L}_{t,k}(f) - L_t(f)| \le B(k) + V(k)$, where

$$V(k) \asymp \frac{\sigma_{t,k}}{\sqrt{k}} + \frac{1}{k} \quad \text{and} \quad B(k) = \max_{i \in [k]} \big| L_{t-i+1}(f) - L_t(f) \big|.$$

Window Selection

Bias-variance tradeoff: $|\widehat{L}_{t,k}(f) - L_t(f)| \le B(k) + V(k)$, where

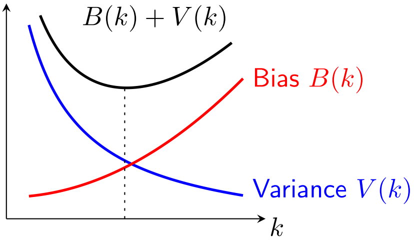
$$V(k) \asymp \frac{\sigma_{t,k}}{\sqrt{k}} + \frac{1}{k} \quad \text{and} \quad B(k) = \max_{i \in [k]} \big| L_{t-i+1}(f) - L_t(f) \big|.$$



Window Selection

Bias-variance tradeoff: $|\widehat{L}_{t,k}(f) - L_t(f)| \leq B(k) + V(k)$, where

$$V(k) \asymp \frac{\sigma_{t,k}}{\sqrt{k}} + \frac{1}{k} \quad \text{and} \quad B(k) = \max_{i \in [k]} \left| L_{t-i+1}(f) - L_t(f) \right|.$$



Challenge: $B(k)$ and $V(k)$ depend on unknown distribution shift

Adaptive Window Selection

Construct data-driven proxies for $V(k)$ and $B(k)$:

$$\widehat{V}(k) \asymp \frac{\widehat{\sigma}_{t,k}}{\sqrt{k}} + \frac{1}{k},$$

$$\widehat{B}(k) = \max_{i \in [k]} \left(\left| \widehat{L}_{t,k}(f) - \widehat{L}_{t,i}(f) \right| - \left[\widehat{V}(k) + \widehat{V}(i) \right] \right)_+ .$$

Adaptive Window Selection

Construct data-driven proxies for $V(k)$ and $B(k)$:

$$\widehat{V}(k) \asymp \frac{\widehat{\sigma}_{t,k}}{\sqrt{k}} + \frac{1}{k},$$

$$\widehat{B}(k) = \max_{i \in [k]} \left(\left| \widehat{L}_{t,k}(f) - \widehat{L}_{t,i}(f) \right| - \left[\widehat{V}(k) + \widehat{V}(i) \right] \right)_+ .$$

Theorem

Choose $\widehat{k} \in \operatorname{argmin} \left\{ \widehat{B}(k) + \widehat{V}(k) \right\}$*. With high probability,*

$$\left| \widehat{L}_{t,\widehat{k}}(f) - L_t(f) \right| \lesssim \min_{1 \le k \le t} \left\{ B(k) + V(k) \right\}.$$

Adaptive Window Selection

Construct data-driven proxies for $V(k)$ and $B(k)$:

$$\widehat{V}(k) \asymp \frac{\widehat{\sigma}_{t,k}}{\sqrt{k}} + \frac{1}{k},$$

$$\widehat{B}(k) = \max_{i \in [k]} \left(\left| \widehat{L}_{t,k}(f) - \widehat{L}_{t,i}(f) \right| - \left[\widehat{V}(k) + \widehat{V}(i) \right] \right)_{+}.$$

Theorem

Choose $\widehat{k} \in \operatorname{argmin} \left\{ \widehat{B}(k) + \widehat{V}(k) \right\}$. *With high probability,*

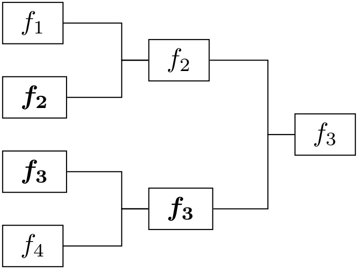
$$\left| \widehat{L}_{t,\widehat{k}}(f) - L_t(f) \right| \lesssim \min_{1 \leq k \leq t} \left\{ B(k) + V(k) \right\}.$$

Adaptivity to unknown distribution shift!

Model Selection

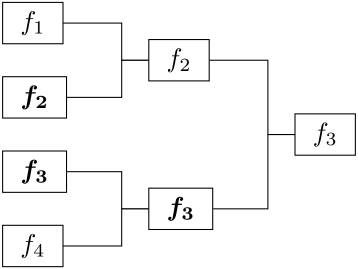
Model Selection

Single-elimination tournament based on pairwise comparisons:



Model Selection

Single-elimination tournament based on pairwise comparisons:

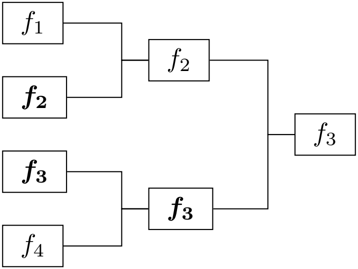


Pairwise comparison: Use rolling window to estimate *performance gap*

$$L_t(f_1) - L_t(f_2).$$

Model Selection

Single-elimination tournament based on pairwise comparisons:



Pairwise comparison: Use rolling window to estimate *performance gap*

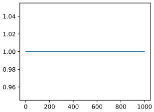
$$L_t(f_1) - L_t(f_2).$$

Theoretical guarantee: Near-optimal model selection.

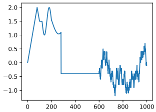
Numerical Experiments

Numerical Experiments

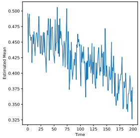
Model selection for prediction tasks:



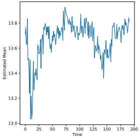
Synthetic-1



Synthetic-2



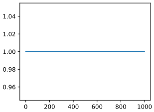
topic frequency on arXiv



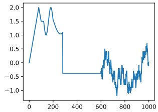
Dubai house prices

Numerical Experiments

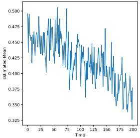
Model selection for prediction tasks:



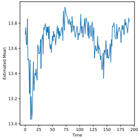
Synthetic-1



Synthetic-2



topic frequency on arXiv



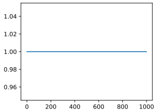
Dubai house prices

Candidate models $f_1, ..., f_m$:

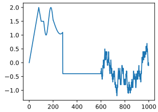
- Moving average, random forest, XGBoost
- Trained on different windows of past data

Numerical Experiments

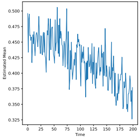
Model selection for prediction tasks:



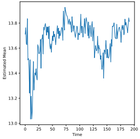
Synthetic-1



Synthetic-2



topic frequency on arXiv



Dubai house prices

Candidate models $f_1, ..., f_m$:

- Moving average, random forest, XGBoost
- Trained on different windows of past data

Benchmark algorithm $\mathcal{A}_k$: use a fixed window k to select models.

Numerical Experiments

Table: Mean excess risks of selection methods for different datasets

Data	Ours	$\mathcal{A}_1$	$\mathcal{A}_4$	$\mathcal{A}_{16}$	$\mathcal{A}_{64}$	$\mathcal{A}_{256}$
Synthetic-1	0.015	0.043	0.025	0.013	**0.010**	**0.010**
Synthetic-2	0.139	**0.157**	0.171	0.539	1.034	1.067
Arxiv (in 1E-3)	2.4	6.7	4.5	2.4	**1.7**	1.9
Housing	0.071	0.071	**0.069**	0.071	0.091	0.095

Numerical Experiments

Table: Mean excess risks of selection methods for different datasets

Data	Ours	$\mathcal{A}_1$	$\mathcal{A}_4$	$\mathcal{A}_{16}$	$\mathcal{A}_{64}$	$\mathcal{A}_{256}$
Synthetic-1	0.015	0.043	0.025	0.013	**0.010**	**0.010**
Synthetic-2	0.139	**0.157**	0.171	0.539	1.034	1.067
Arxiv (in 1E-3)	2.4	6.7	4.5	2.4	**1.7**	1.9
Housing	0.071	0.071	**0.069**	0.071	0.091	0.095

- Different distribution shift patterns lead to different optimal k.

Numerical Experiments

Table: Mean excess risks of selection methods for different datasets

Data	Ours	$\mathcal{A}_1$	$\mathcal{A}_4$	$\mathcal{A}_{16}$	$\mathcal{A}_{64}$	$\mathcal{A}_{256}$
Synthetic-1	0.015	0.043	0.025	0.013	**0.010**	**0.010**
Synthetic-2	0.139	**0.157**	0.171	0.539	1.034	1.067
Arxiv (in 1E-3)	2.4	6.7	4.5	2.4	**1.7**	1.9
Housing	0.071	0.071	**0.069**	0.071	0.091	0.095

- Different distribution shift patterns lead to different optimal k.
- Our algorithm is comparable to $\mathcal{A}_k$ **with the best k in hindsight**.