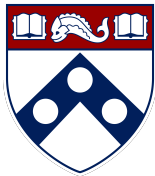


Optimal training-conditional regret for online conformal prediction



Jiadong Liang

Wharton Statistics & Data Science



Zhimei Ren
UPenn



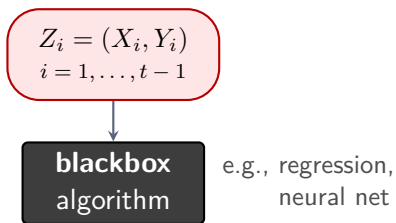
Yuxin Chen
UPenn

“Optimal training-conditional regret for online conformal prediction,” J. Liang, Z. Ren, Y. Chen, arXiv:2602.16537, 2026

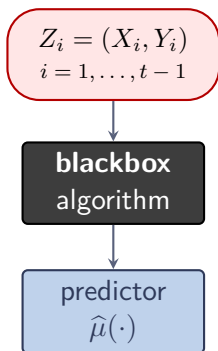
Standard online learning

$$\begin{aligned} Z_i &= (X_i, Y_i) \\ i &= 1, \dots, t-1 \end{aligned}$$

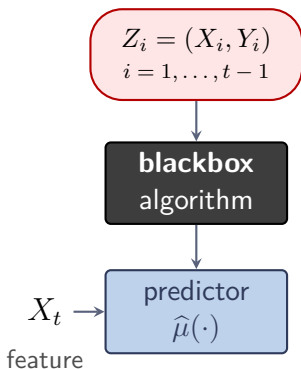
Standard online learning



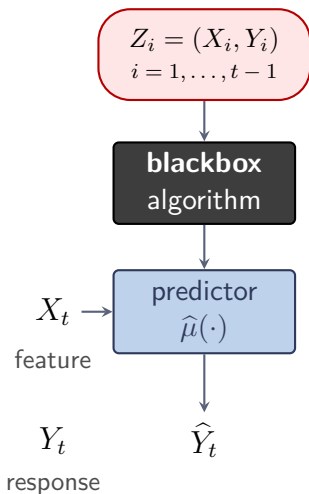
Standard online learning



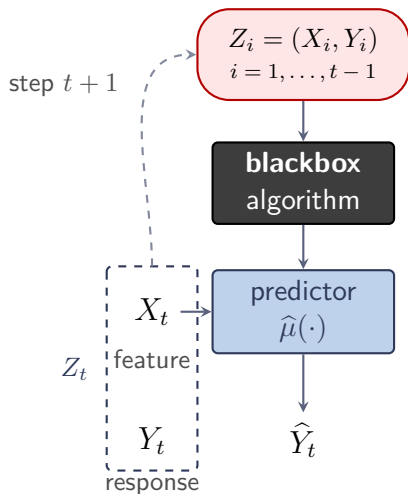
Standard online learning



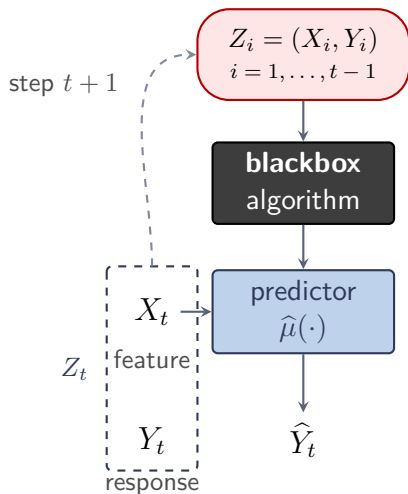
Standard online learning



Standard online learning

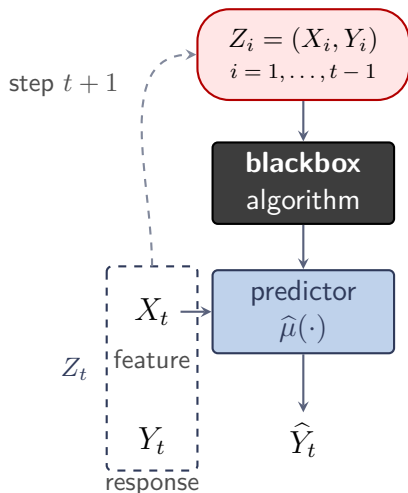


Standard online learning



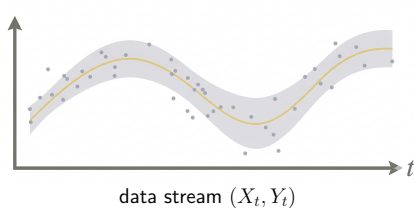
- **prediction**
ML algorithms provide a prediction $\hat{Y}_t$ at time t

Standard online learning

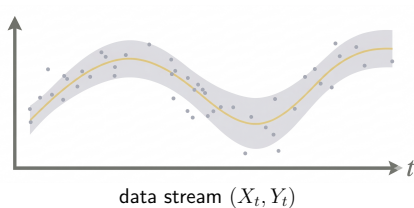


- **prediction**
ML algorithms provide a prediction $\hat{Y}_t$ at time t
- **uncertainty of prediction**
how reliable is $\hat{Y}_t$?

Online uncertainty quantification

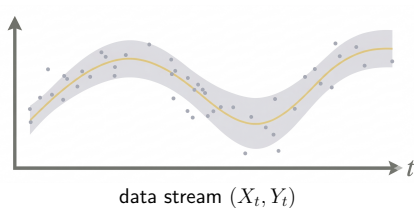


Online uncertainty quantification



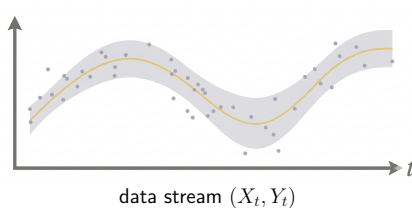
$$\hat{\mu}(\cdot)$$

Online uncertainty quantification



$$\hat{\mu}(\cdot) \longrightarrow \mathcal{C}(\cdot)$$

Online uncertainty quantification

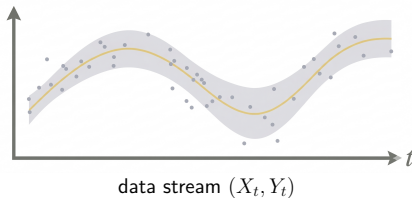


$$\hat{\mu}(\cdot) \longrightarrow \mathcal{C}(\cdot)$$

goal: construct prediction sets for streaming data in real time

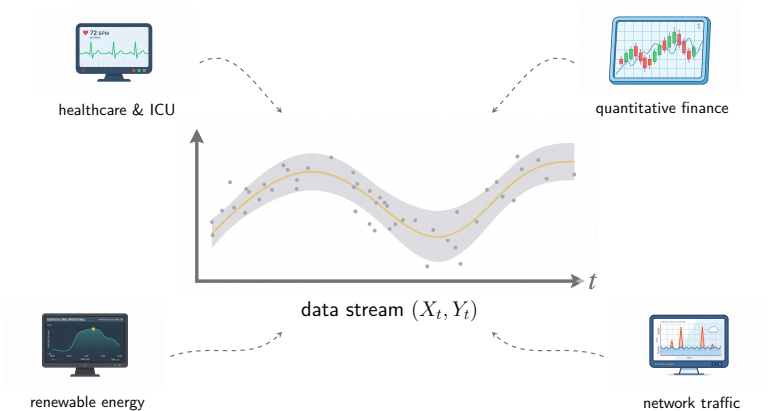
Online uncertainty quantification

challenge: data distributions & predictive models may **change unpredictably** over time.



goal: construct prediction sets for streaming data in real time

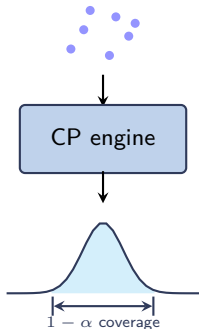
Online uncertainty quantification



goal: construct prediction sets for streaming data in real time

Conformal prediction

input data $\{Z_i = (X_i, Y_i)\}$



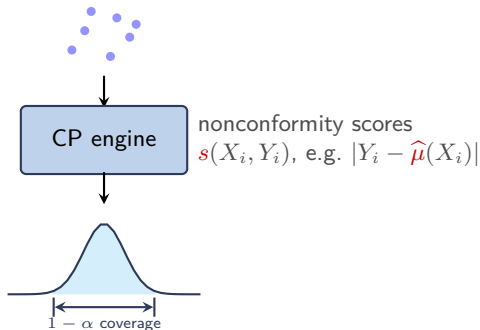
prediction set $\mathcal{C}_t(X_t)$



Vladimir Vovk

Conformal prediction

input data $\{Z_i = (X_i, Y_i)\}$



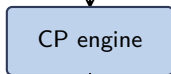
prediction set $\mathcal{C}_t(X_t)$



Vladimir Vovk

Conformal prediction

input data $\{Z_i = (X_i, Y_i)\}$



nonconformity scores
 $s(X_i, Y_i)$, e.g. $|Y_i - \hat{\mu}(X_i)|$

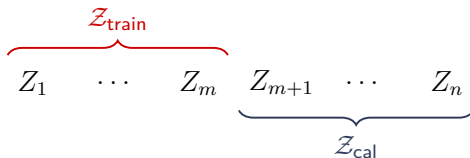


prediction set $\mathcal{C}_t(X_t)$



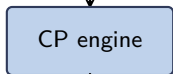
Vladimir Vovk

method 1: split conformal



Conformal prediction

input data $\{Z_i = (X_i, Y_i)\}$



nonconformity scores
 $s(X_i, Y_i)$, e.g. $|Y_i - \hat{\mu}(X_i)|$

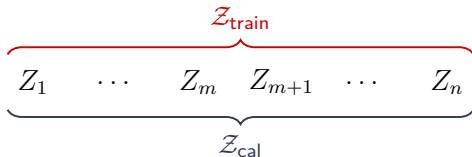


prediction set $\mathcal{C}_t(X_t)$



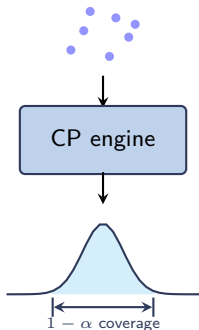
Vladimir Vovk

method 2: full conformal



Conformal prediction

input data $\{Z_i = (X_i, Y_i)\}$



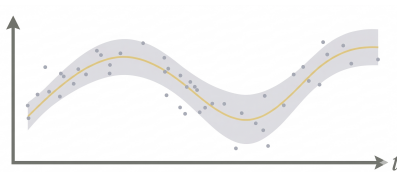
prediction set $\mathcal{C}_t(X_t)$



Vladimir Vovk

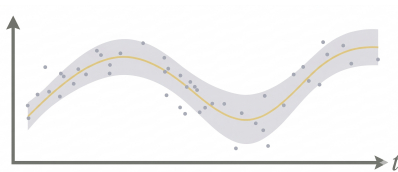
a **distribution-free** predictive inference approach, valid for any data dist. under exchangeability

Key challenge in online settings: non-exchangeability



data stream may be **nonstationary**

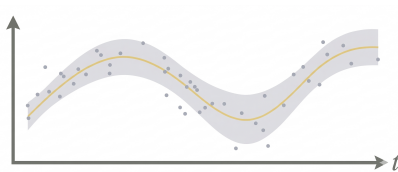
Key challenge in online settings: non-exchangeability



data stream may be **nonstationary**

~~exchangeability~~

Key challenge in online settings: non-exchangeability

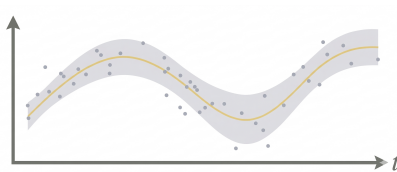


~~exchangeability~~

data stream may be **nonstationary**

question: can conformal prediction work beyond exchangeability?

Key challenge in online settings: non-exchangeability



~~exchangeability~~

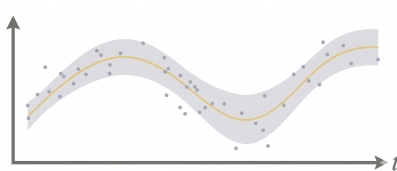
data stream may be **nonstationary**

question: can conformal prediction work beyond exchangeability?

- *NexCP*: handles *offline* non-exchangeable data [Barber et al., 2023]

• Barber et al.. *Conformal prediction beyond exchangeability*, Annals of Statistics 2023.

Key challenge in online settings: non-exchangeability



~~exchangeability~~

data stream may be **nonstationary**

question: can conformal prediction work beyond exchangeability?

- *NexCP*: handles *offline* non-exchangeable data [Barber et al., 2023]
- *ACI*: designed for arbitrary *online* data [Gibbs & Candès, 2021]

- Barber et al.. *Conformal prediction beyond exchangeability*, Annals of Statistics 2023.
- Gibbs & Candès. *Adaptive conformal inference under distribution shift*, NeurIPS 2021.

Adaptive conformal inference (ACI)

$$\alpha_t = \alpha_{t-1} + \gamma_t(\alpha - \mathbb{1}\{Y_{t-1} \notin \mathcal{C}_{t-1}(X_{t-1})\})$$

- online gradient descent applied to pinball loss
- $\alpha_t \rightarrow q_t = 1 - \alpha_t$ quantile $\rightarrow \mathcal{C}_t(X_t) = \{y : s_t(X_t, y) \leq q_t\}$

- Gibbs & Candès. *Adaptive conformal inference under distribution shift*, NeurIPS 2021.
- Angelopoulos et al. *Online conformal prediction with decaying step sizes*, ICML 2024.

Adaptive conformal inference (ACI)

$$\alpha_t = \alpha_{t-1} + \gamma_t(\alpha - \mathbb{1}\{Y_{t-1} \notin \mathcal{C}_{t-1}(X_{t-1})\})$$

- online gradient descent applied to pinball loss
- $\alpha_t \rightarrow q_t = 1 - \alpha_t$ quantile $\rightarrow \mathcal{C}_t(X_t) = \{y : s_t(X_t, y) \leq q_t\}$

Long-term coverage (LTC) guarantees for ACI

$$\underbrace{\left| \frac{1}{T} \sum_{t=1}^T \mathbb{1}\{Y_t \in \mathcal{C}_t(X_t)\} - (1 - \alpha) \right|}_{\text{time-averaged empirical cvg}} \xrightarrow{T \rightarrow \infty} 0$$

- *no assumption at all!*

- Gibbs & Candès. *Adaptive conformal inference under distribution shift*, NeurIPS 2021.
- Angelopoulos et al. *Online conformal prediction with decaying step sizes*, ICML 2024.

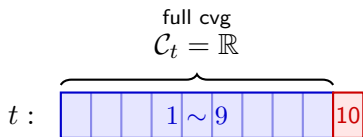
Is “long-term coverage” informative enough?

Outline

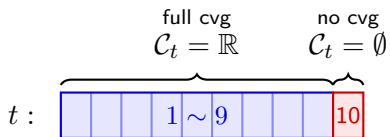
- 1 Long-term coverage vs. training-conditional regret
- 2 Problem formulation
- 3 Online conformal with pretrained scores
- 4 Online conformal with adaptively trained scores
- 5 Concluding remarks

- 1 Long-term coverage vs. training-conditional regret
- 2 Problem formulation
- 3 Online conformal with pretrained scores
- 4 Online conformal with adaptively trained scores
- 5 Concluding remarks

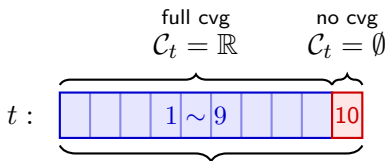
A pathological example ($\alpha = 10\%$)



A pathological example ($\alpha = 10\%$)

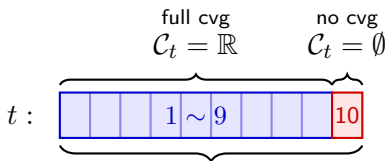


A pathological example ($\alpha = 10\%$)



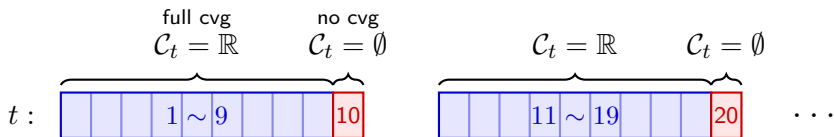
$$\frac{1}{10} \sum_{t=1}^{10} \mathbb{1}\{Y_t \in \mathcal{C}_t\} = 0.9$$

A pathological example ($\alpha = 10\%$)

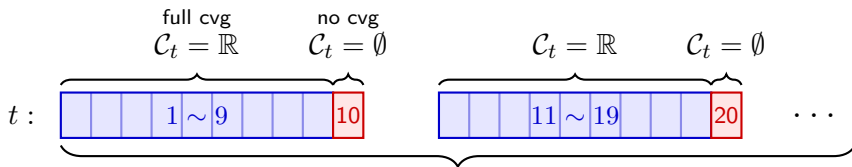


$$\frac{1}{10} \sum_{t=1}^{10} \mathbb{1}\{Y_t \in \mathcal{C}_t\} = 0.9 \quad (\text{LTC in period 1})$$

A pathological example ($\alpha = 10\%$)

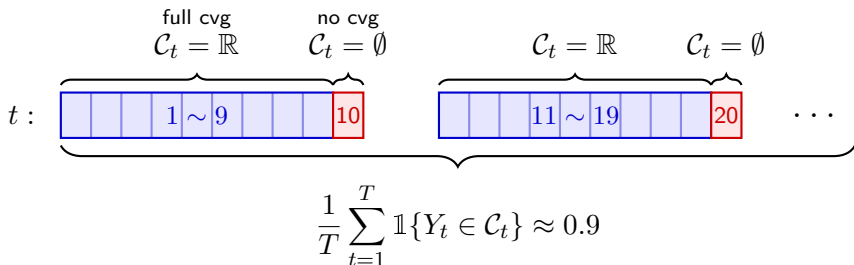


A pathological example ($\alpha = 10\%$)



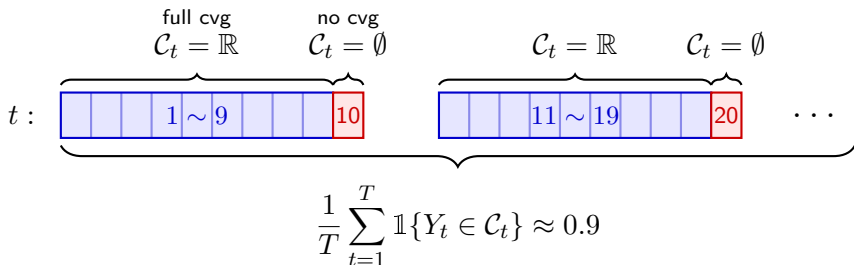
$$\frac{1}{T} \sum_{t=1}^T \mathbb{1}\{Y_t \in \mathcal{C}_t\} \approx 0.9$$

A pathological example ($\alpha = 10\%$)



- vacuous prediction set at every single time

A pathological example ($\alpha = 10\%$)



- vacuous prediction set at every single time
- perfect “long-term cvg” $\leftarrow$ cancellation of under-cvg/over-cvg

Our metric: training-conditional regret

$$\text{regret}_T := \sum_{t=1}^T \mathbb{E} \left[\left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right| \right]$$

Our metric: training-conditional regret

$$\text{regret}_T := \sum_{t=1}^T \mathbb{E} \left[\left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right| \right]$$

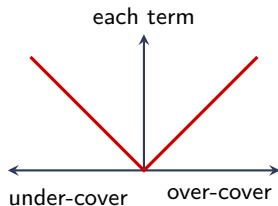


history prior to time t
($Z_1, \dots, Z_{t-1}$)

internal randomness
(for randomized alg.)

Our metric: training-conditional regret

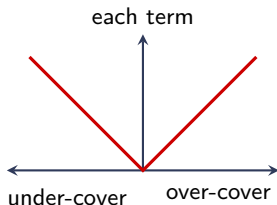
$$\text{regret}_T := \sum_{t=1}^T \mathbb{E} \left[\left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right| \right]$$



- **instantaneous validity**
 - penalizes both over- & under-coverage
(no cancellation allowed)

Our metric: training-conditional regret

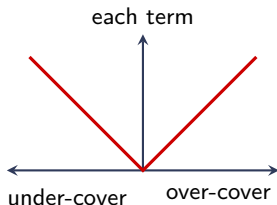
$$\text{regret}_T := \sum_{t=1}^T \mathbb{E} \left[\left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right| \right]$$



- **instantaneous validity**
 - penalizes both over- & under-coverage
(no cancellation allowed)
- **cond. on observed data**
 - stronger notion than marginal validity

Our metric: training-conditional regret

$$\text{regret}_T := \sum_{t=1}^T \mathbb{E} \left[\left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right| \right]$$



- **instantaneous validity**
 - penalizes both over- & under-coverage
(no cancellation allowed)
- **cond. on observed data**
 - stronger notion than marginal validity

Sublinear regret guarantees long-term coverage!

$\text{regret}_T = o(T)$

Other regret metrics?

(static) regret:

$$\text{AdvRegret}_T = \sum_{t=1}^T \text{loss}_t(q_t) - \min_q \sum_{t=1}^T \text{loss}_t(q)$$

- Bhatnagar et al. *Improved Regret Bounds for Online Prediction Intervals*, ICML 2023.
- Hajihashemi & Shen. *On the Regret of Online Quantile Regression with Distribution Drift*, AISTATS 2024.
- Ramalingam et al. *Bridging the Gap: Statistical Mismatch in Online Conformal Prediction*, arXiv 2025.

Other regret metrics?

(static) regret:

$$\text{AdvRegret}_T = \sum_{t=1}^T \text{loss}_t(q_t) - \min_q \sum_{t=1}^T \text{loss}_t(q)$$

- hindsight-optimal q does not account for **distribution drift**

- Bhatnagar et al. *Improved Regret Bounds for Online Prediction Intervals*, ICML 2023.
- Hajishemi & Shen. *On the Regret of Online Quantile Regression with Distribution Drift*, AISTATS 2024.
- Ramalingam et al. *Bridging the Gap: Statistical Mismatch in Online Conformal Prediction*, arXiv 2025.

Other regret metrics?

(static) regret:

$$\text{AdvRegret}_T = \sum_{t=1}^T \text{loss}_t(q_t) - \min_q \sum_{t=1}^T \text{loss}_t(q)$$

- hindsight-optimal q does not account for **distribution drift**
- lacks direct correspondence w/ classical stat. validity notion

- Bhatnagar et al. *Improved Regret Bounds for Online Prediction Intervals*, ICML 2023.
- Hajishemi & Shen. *On the Regret of Online Quantile Regression with Distribution Drift*, AISTATS 2024.
- Ramalingam et al. *Bridging the Gap: Statistical Mismatch in Online Conformal Prediction*, arXiv 2025.

This talk: design efficient online conformal algorithms
guiding principle: training-conditional regret

- 1 Long-term coverage vs. training-conditional regret
- 2 Problem formulation
- 3 Online conformal with pretrained scores
- 4 Online conformal with adaptively trained scores
- 5 Concluding remarks

Problem formulation

- **independent** data stream $Z_t := (X_t, Y_t) \stackrel{\text{ind.}}{\sim} \mathcal{D}_t, t = 1, 2, \dots$

Problem formulation

- **independent** data stream $Z_t := (X_t, Y_t) \stackrel{\text{ind.}}{\sim} \mathcal{D}_t, t = 1, 2, \dots$
- **unknown** distribution $\mathcal{D}_t$ (no prior info about drift)

Problem formulation

- **independent** data stream $Z_t := (X_t, Y_t) \stackrel{\text{ind.}}{\sim} \mathcal{D}_t, t = 1, 2, \dots$
- **unknown** distribution $\mathcal{D}_t$ (no prior info about drift)

Two settings:

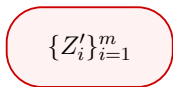
Problem formulation

- **independent** data stream $Z_t := (X_t, Y_t) \stackrel{\text{ind.}}{\sim} \mathcal{D}_t, t = 1, 2, \dots$
- **unknown** distribution $\mathcal{D}_t$ (no prior info about drift)

Two settings:

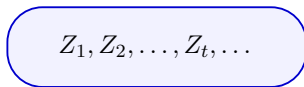
1. **pretrained scores:** $\hat{\mu}(\cdot)$ & $s(\cdot, \cdot)$ are independent of data stream

separate, ind. dataset



$\hat{\mu}(\cdot)$
(or $s(\cdot, \cdot)$)

data stream



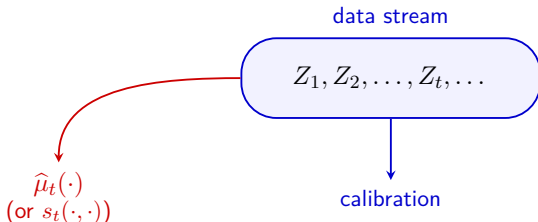
calibration

Problem formulation

- **independent** data stream $Z_t := (X_t, Y_t) \stackrel{\text{ind.}}{\sim} \mathcal{D}_t, t = 1, 2, \dots$
- **unknown** distribution $\mathcal{D}_t$ (no prior info about drift)

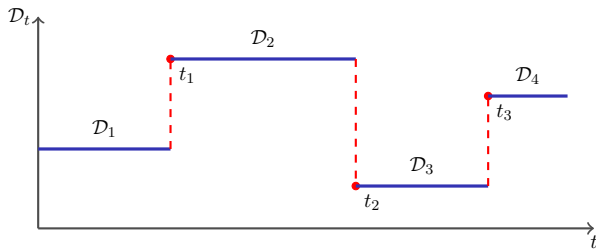
Two settings:

2. adaptively trained scores: $\hat{\mu}_t, s_t$ are trained on past data streams



Modeling distribution drift

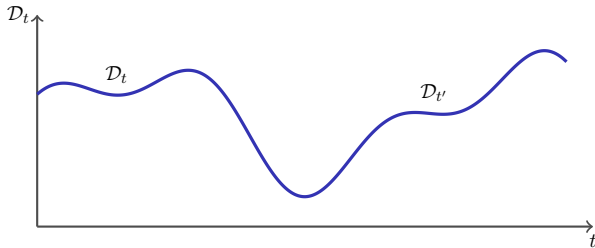
1. change-point setting:



change points $\leq N^{\text{cp}}$

Modeling distribution drift

2. smooth drift setting:



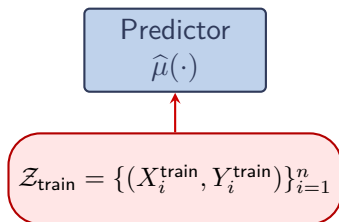
$$\sum_{t=1}^{T-1} \text{KS}(\underbrace{\mathcal{D}_t^{\text{score}}}_{\text{distribution of } s_t(X_t, Y_t)}, \mathcal{D}_{t+1}^{\text{score}}) \leq \text{KS}_T \quad \text{or} \quad \sum_{t=1}^{T-1} \text{TV}(\mathcal{D}_t, \mathcal{D}_{t+1}) \leq \text{TV}_T$$

- $\text{KS}(P, Q) = \sup_z |F_P(z) - F_Q(z)|$ (Kolmogorov-Smirnov distance)
- $\text{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ (total variation distance)

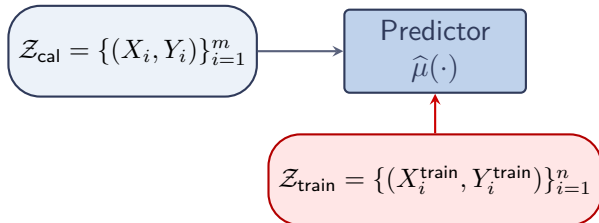
*How to achieve optimal training-conditional regret
for these settings?*

- 1 Long-term coverage vs. training-conditional regret
- 2 Problem formulation
- 3 Online conformal with pretrained scores**
- 4 Online conformal with adaptively trained scores
- 5 Concluding remarks

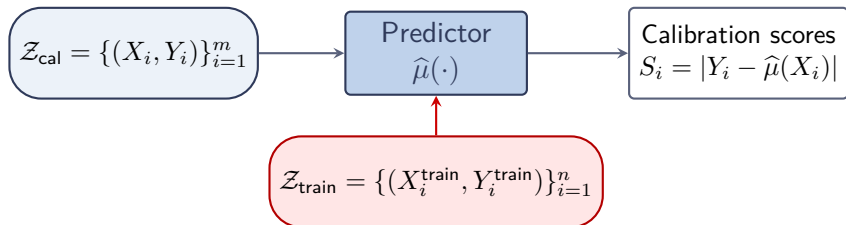
Background: split conformal



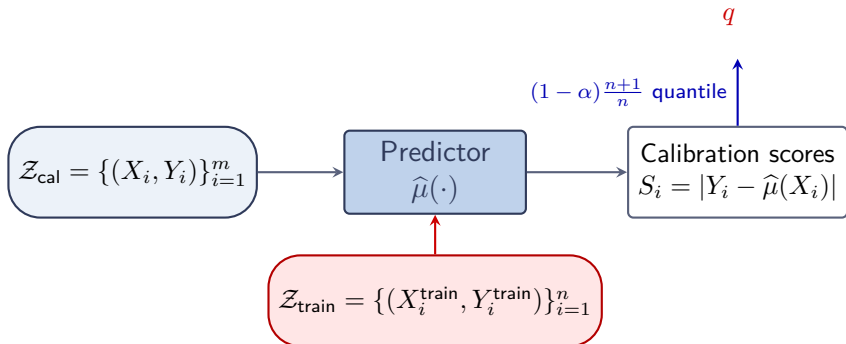
Background: split conformal



Background: split conformal

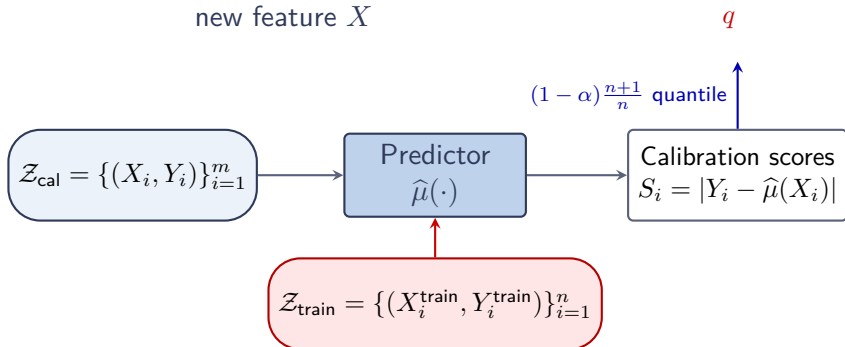


Background: split conformal

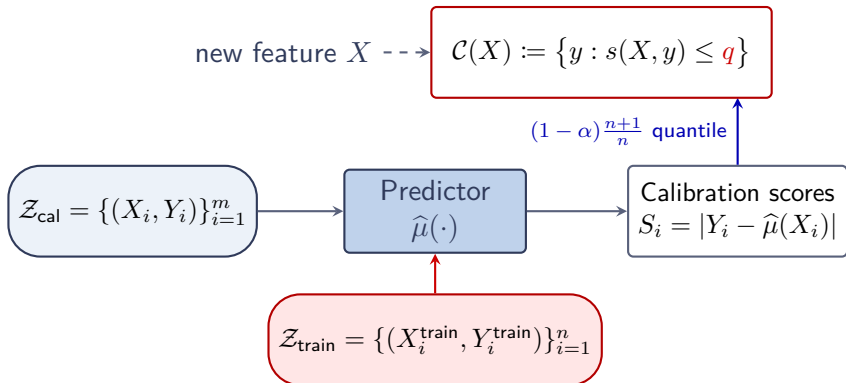


Background: split conformal

new feature X



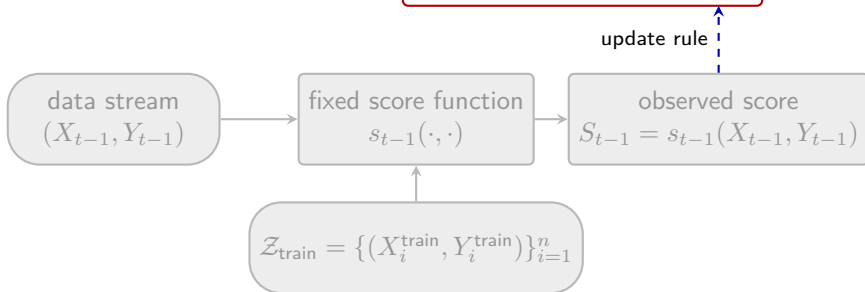
Background: split conformal



Split conformal for pretrained score setting

setup: scores $s_t(\cdot, \cdot)$ are trained on a separate, independent dataset

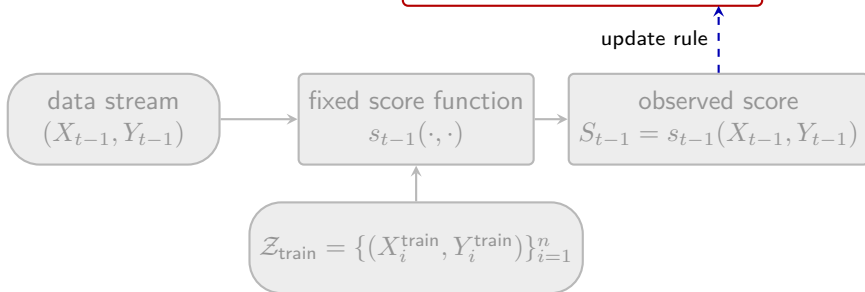
feature X_t $\dashrightarrow$ $\mathcal{C}_t(X_t) := \{y : s_t(X_t, y) \leq q_t\}$



Split conformal for pretrained score setting

setup: scores $s_t(\cdot, \cdot)$ are trained on a separate, independent dataset

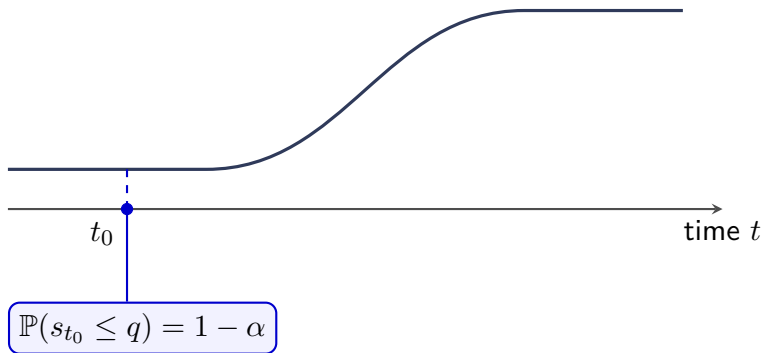
feature X_t $\dashrightarrow$ $\mathcal{C}_t(X_t) := \{y : s_t(X_t, y) \leq q_t\}$



- **main task:** compute a quantile threshold q_t at each time t

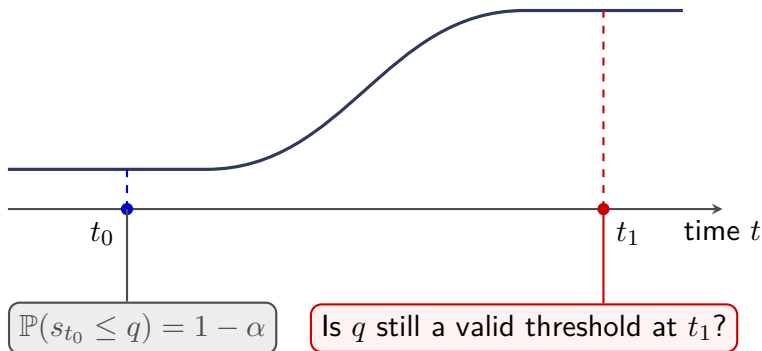
Challenge: quantile changes under distribution drift

suppose we have a perfect calibration quantile q at time $t_0 \dots$



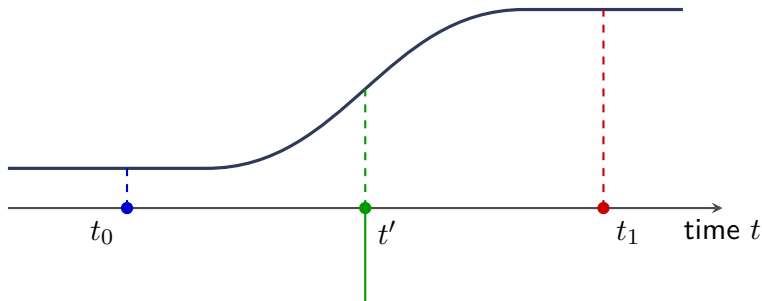
Challenge: quantile changes under distribution drift

suppose we have a perfect calibration quantile q at time $t_0 \dots$



Challenge: quantile changes under distribution drift

suppose we have a perfect calibration quantile q at time $t_0 \dots$



idea: restart and update q once drift is detected!

Detour: drift detection

a natural metric: block coverage error

$$\text{cvg-err}_q^*(s, t) := \sum_{l=s}^t \left(\mathbb{P}(s_l(X_l, Y_l) \leq q) - (1 - \alpha) \right)$$

Detour: drift detection

a natural metric: block coverage error

$$\text{cvg-err}_q^*(t_0, t) := \sum_{l=t_0}^t \left(\mathbb{P}(s_l(X_l, Y_l) \leq q) - (1 - \alpha) \right)$$

- equals 0 if no distribution drift occurs

Detour: drift detection

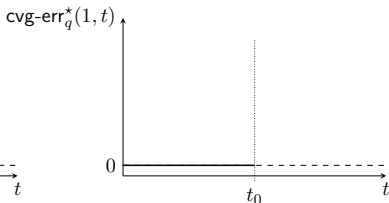
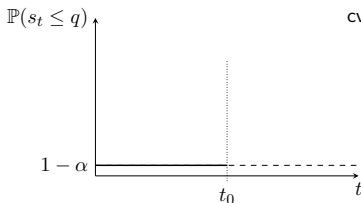
a natural metric: block coverage error

$$\text{cvg-err}_q^*(t_0, t) := \sum_{l=t_0}^t \left(\mathbb{P}(s_l(X_l, Y_l) \leq q) - (1 - \alpha) \right)$$

- equals 0 if no distribution drift occurs
- *what if distribution drift DOES occur?*

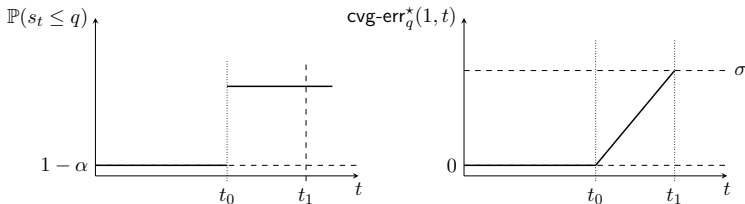
Detour: drift detection

motivating scenario 1 (abrupt shift): score distribution abruptly shifts from $\mathcal{D}_1^{\text{seg}}$ to $\mathcal{D}_2^{\text{seg}}$ at time t_0



Detour: drift detection

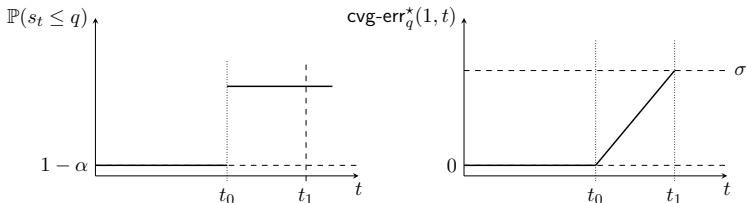
motivating scenario 1 (abrupt shift): score distribution abruptly shifts from $\mathcal{D}_1^{\text{seg}}$ to $\mathcal{D}_2^{\text{seg}}$ at time t_0



- $t_0 \rightarrow t_1$: $\text{cvg-err}_q^*(1, t)$ grows linearly

Detour: drift detection

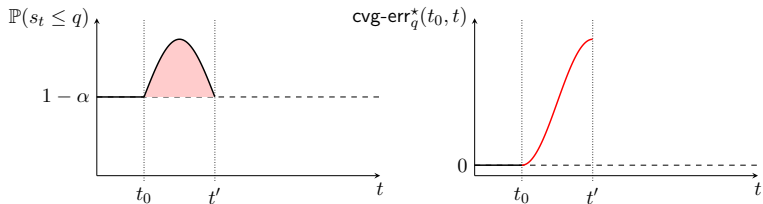
motivating scenario 1 (abrupt shift): score distribution abruptly shifts from $\mathcal{D}_1^{\text{seg}}$ to $\mathcal{D}_2^{\text{seg}}$ at time t_0



- $t_0 \rightarrow t_1$: $\text{cvg-err}_q^*(1, t)$ grows linearly
- a simple threshold σ suffices to detect drift

Detour: drift detection

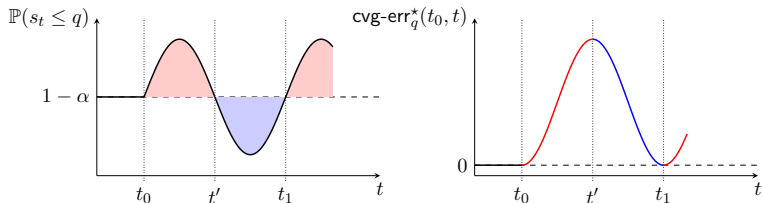
motivating scenario 2 (smooth oscillation): what if the distribution oscillates smoothly?



- $t_0 \rightarrow t'$: $\text{cvg-err}_q^*(t_0, t)$ increasing

Detour: drift detection

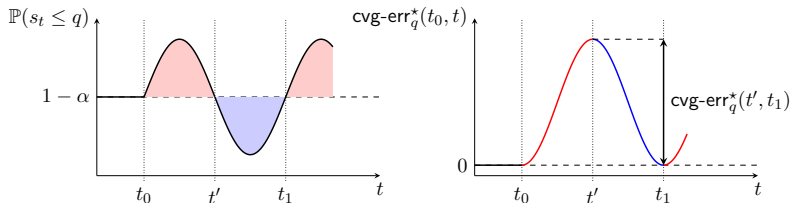
motivating scenario 2 (smooth oscillation): what if the distribution oscillates smoothly?



- $t_0 \rightarrow t'$: $\text{cvg-err}_q^*(t_0, t)$ increases
- $t' \rightarrow t_1$: drift reverses, $\text{cvg-err}_q^*(t_0, t_1)$ drops back to near 0

Detour: drift detection

motivating scenario 2 (smooth oscillation): what if the distribution oscillates smoothly?



- $t_0 \rightarrow t'$: $\text{cvg-err}_q^*(t_0, t)$ increases
- $t' \rightarrow t_1$: drift reverses, $\text{cvg-err}_q^*(t_0, t_1)$ drops back to near 0

solution: check $\text{cvg-err}_q^*(s, t)$ for all windows $[s, t] \subseteq [t_0, t_1]$

Detour: drift detection

population to empirical: in practice, monitor *empirical* block coverage error instead:

$$\text{cvg-err}_q(s, t) \quad := \quad \sum_{l=s}^t \left(\mathbb{1}\{s_l(X_l, Y_l) \leq q\} - (1 - \alpha) \right)$$

Detour: drift detection

population to empirical: in practice, monitor *empirical* block coverage error instead:

$$\begin{array}{c} \text{cvg-err}_q(s, t) := \sum_{l=s}^t (\mathbb{1}\{s_l(X_l, Y_l) \leq q\} - (1 - \alpha)) \\ \updownarrow \\ \text{cvg-err}_q^*(s, t) \end{array} \lesssim \tilde{O}(\sqrt{t - s + 1})$$

Detour: drift detection

population to empirical: in practice, monitor *empirical* block coverage error instead:

$$\begin{array}{c} \text{cvg-err}_q(s, t) \quad := \quad \sum_{l=s}^t (\mathbb{1}\{s_l(X_l, Y_l) \leq q\} - (1 - \alpha)) \\ \updownarrow \\ \text{cvg-err}_q^*(s, t) \end{array} \quad \lesssim \quad \tilde{O}(\sqrt{t - s + 1})$$

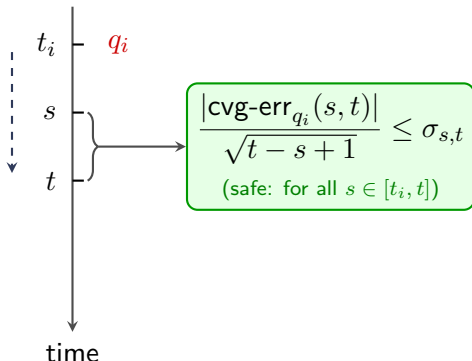
- empirical fluctuations within the scale of $\tilde{O}(\sqrt{t - s + 1})$ cannot be viewed as occurrence of drift

Proposed algorithm: DriftOCP



1. track history: maintain an active window of data points starting from the latest reset point t_i

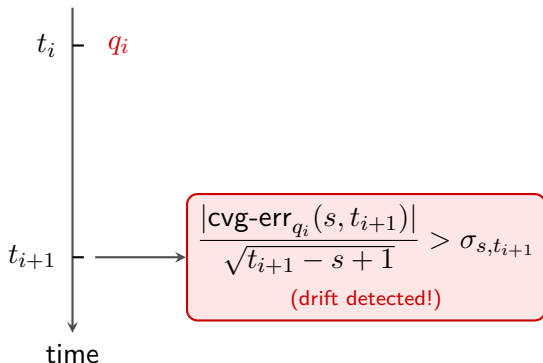
Proposed algorithm: DriftOCP



2. scan for drift: continuously check if the normalized empirical fluctuation stays below the safe threshold $\sigma_{s,t}$

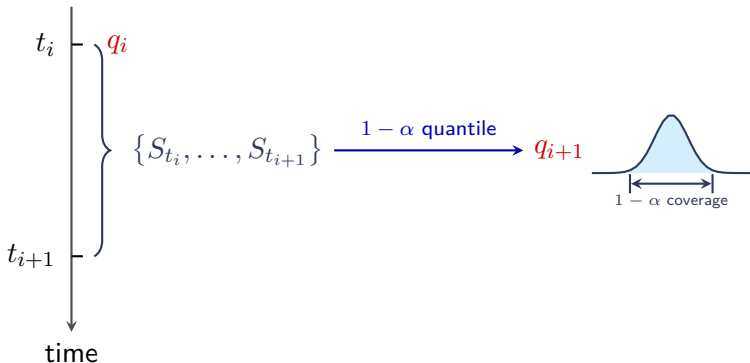
* typically $\sigma_{s,t}$ is chosen as $O(\sqrt{\log t})$

Proposed algorithm: DriftOCP



3. declare drift: if cvg-err exceeds the threshold for *any* past sub-interval, a distribution drift is declared at time t_{i+1}

Proposed algorithm: DriftOCP



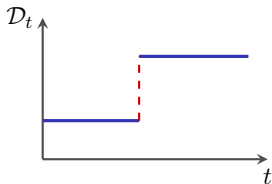
4. reset & update: discard stale history; extract scores from new window $[t_i, t_{i+1}]$ and recompute a quantile q_{i+1}

Theoretical guarantees for DriftOCP

Theorem 1 (L., Ren, Chen '26)

With appropriate choice detection thresholds, DriftOCP achieves:

- change-point setting: $\text{regret}_T \leq \tilde{O}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$*



budget:

change points $\leq N^{\text{cp}}$

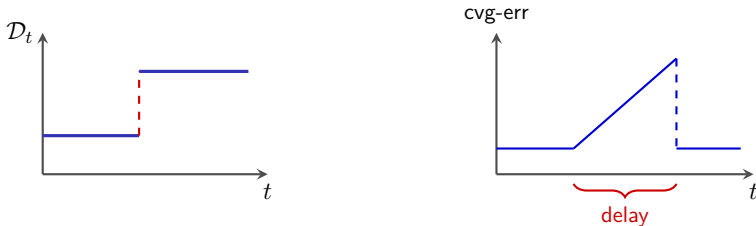
Theoretical guarantees for DriftOCP

Theorem 1 (L., Ren, Chen '26)

With appropriate choice detection thresholds, DriftOCP achieves:

1. change-point setting: $\text{regret}_T \leq \tilde{O}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$

- regret arises from (unavoidable) delay in drift detection



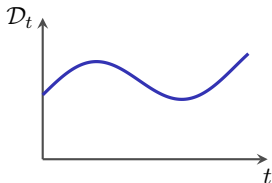
Theoretical guarantees for DriftOCP

Theorem 1 (L., Ren, Chen '26)

With appropriate choice detection thresholds, DriftOCP achieves:

1. *change-point setting*: $\text{regret}_T \leq \tilde{O}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$
2. *smooth drift setting*: $\text{regret}_T \leq \tilde{O}\left(T^{2/3} \text{KS}_T^{1/3} + \sqrt{T}\right)$

- regret arises from (unavoidable) delay in drift detection



budget:

$$\sum_{t=1}^{T-1} \text{KS}(\mathcal{D}_t^{\text{score}}, \mathcal{D}_{t+1}^{\text{score}}) \leq \text{KS}_T$$

Empirical performance of DriftOCP

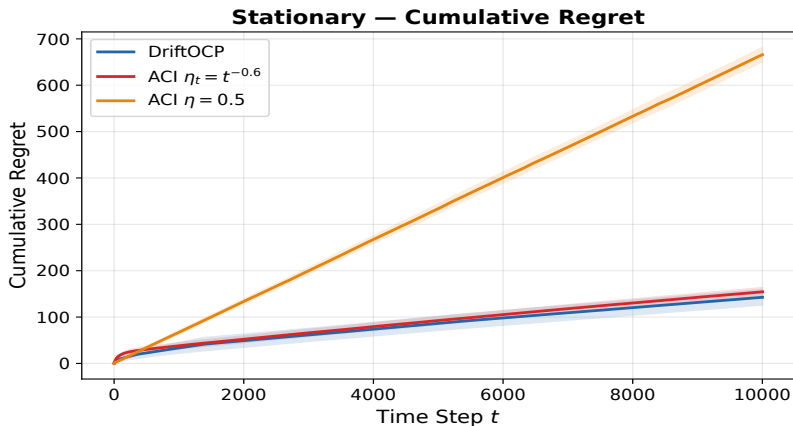


Figure 1: cumulative regret

Empirical performance of DriftOCP

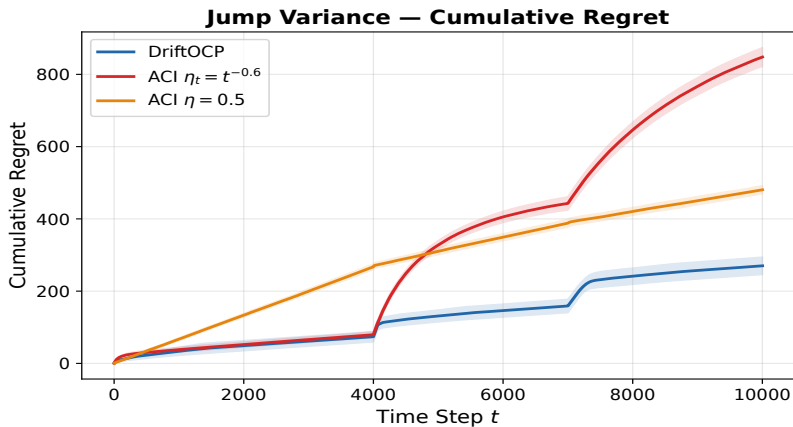


Figure 1: cumulative regret

Is regret performance of DriftOCP improvable?

Matching minimax lower bound

- *admissible algorithms*: $q_t = \pi_t(\text{past scores, rand seed})$

Theorem 2 (L., Ren, Chen '26)

For any admissible algorithm π ,

Matching minimax lower bound

- *admissible algorithms*: $q_t = \pi_t(\text{past scores, rand seed})$
- *distribution classes*:

$$\mathcal{L}_{\text{cp}} := \left\{ \{\mathcal{D}_t\} : \text{score dist. change at most } N^{\text{cp}} \text{ times} \right\}$$

Theorem 2 (L., Ren, Chen '26)

For any admissible algorithm π ,

- *change-point*: $\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_{\text{cp}}} \text{regret}_T = \tilde{\Omega}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$

Matching minimax lower bound

- *admissible algorithms*: $q_t = \pi_t(\text{past scores, rand seed})$
- *distribution classes*:

$$\mathcal{L}_{\text{sd}} := \left\{ \{\mathcal{D}_t\} : \sum_{t=1}^{T-1} \text{KS}(\mathcal{D}_t^{\text{score}}, \mathcal{D}_{t+1}^{\text{score}}) \leq \text{KS}_T \right\}$$

Theorem 2 (L., Ren, Chen '26)

For any admissible algorithm π ,

- *smooth drift*: $\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_{\text{sd}}} \text{regret}_T = \tilde{\Omega}\left(T^{2/3} \text{KS}_T^{1/3} + \sqrt{T}\right)$

Matching minimax lower bound

- *admissible algorithms*: $q_t = \pi_t(\text{past scores, rand seed})$
- *distribution classes*:

$$\mathcal{L}_{\text{sd}} := \left\{ \{\mathcal{D}_t\} : \sum_{t=1}^{T-1} \text{KS}(\mathcal{D}_t^{\text{score}}, \mathcal{D}_{t+1}^{\text{score}}) \leq \text{KS}_T \right\}$$

Theorem 2 (L., Ren, Chen '26)

For any admissible algorithm π ,

- *smooth drift*: $\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_{\text{sd}}} \text{regret}_T = \tilde{\Omega}\left(T^{2/3} \text{KS}_T^{1/3} + \sqrt{T}\right)$
- *our proposed DriftOCP is minimax optimal!*

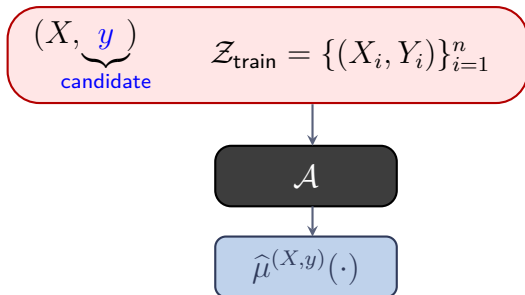
- 1 Long-term coverage vs. training-conditional regret
- 2 Problem formulation
- 3 Online conformal with pretrained scores
- 4 Online conformal with adaptively trained scores**
- 5 Concluding remarks

Background: full conformal

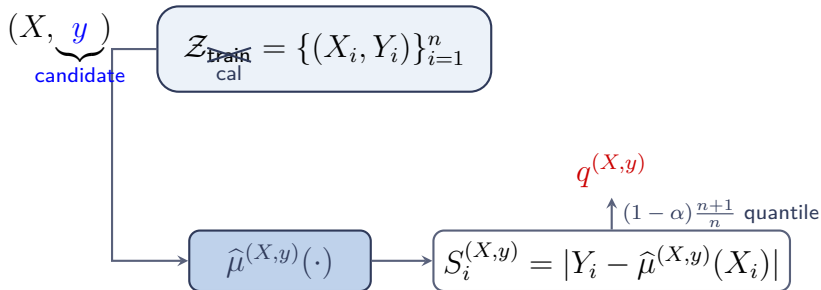
X

$$\mathcal{Z}_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^n$$

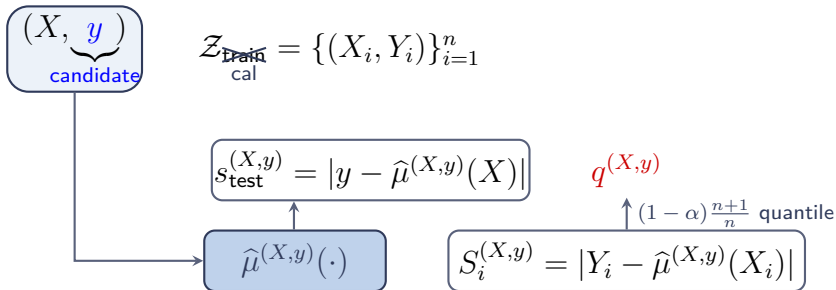
Background: full conformal



Background: full conformal



Background: full conformal



Background: full conformal

$$s_{\text{test}}^{(X,y)} = |y - \hat{\mu}^{(X,y)}(X)| \leq q^{(X,y)} \quad ?$$

Background: full conformal

$$s_{\text{test}}^{(X,y)} = |y - \hat{\mu}^{(X,y)}(X)| \leq q^{(X,y)} \quad ?$$

- ✓ **accept:** $y \in \mathcal{C}(X)$ if the inequality holds
- ✗ **reject:** $y \notin \mathcal{C}(X)$ otherwise

Background: full conformal

standard assumptions:

$$\mathcal{Z}_{\text{train}} = \{(X_i, Y_i)\}_{i < t} \leftarrow \text{- exchangeability}$$

Background: full conformal

standard assumptions:

$$\mathcal{Z}_{\text{train}} = \{(X_i, Y_i)\}_{i < t} \leftarrow \text{exchangeability}$$

$$\mathcal{A} \leftarrow \text{permutation symmetry} \quad \mathcal{A}(\{Z_i\}_{i=1}^n) \equiv \mathcal{A}(\{Z_{\sigma(i)}\}_{i=1}^n)$$

Full conformal for adaptively trained setting?

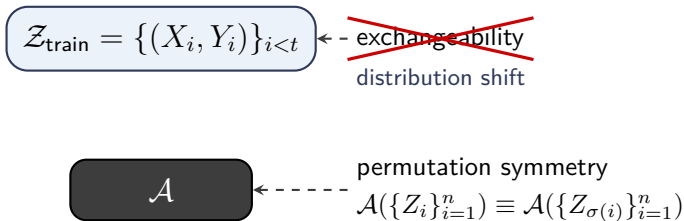
setup: fitted model $\hat{\mu}_t$ and score $s_t(\cdot, \cdot)$ depend on past data

$$\mathcal{Z}_{\text{train}} = \{(X_i, Y_i)\}_{i < t} \leftarrow \text{exchangeability}$$

$$\mathcal{A} \leftarrow \text{permutation symmetry} \quad \mathcal{A}(\{Z_i\}_{i=1}^n) \equiv \mathcal{A}(\{Z_{\sigma(i)}\}_{i=1}^n)$$

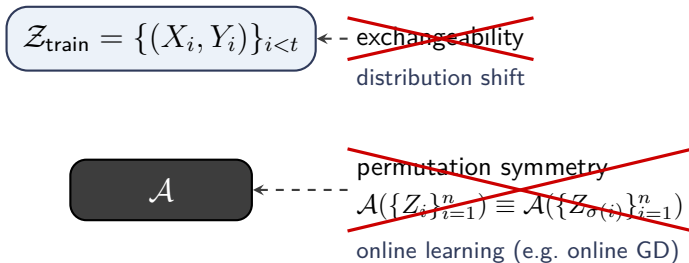
Full conformal for adaptively trained setting?

setup: fitted model $\hat{\mu}_t$ and score $s_t(\cdot, \cdot)$ depend on past data



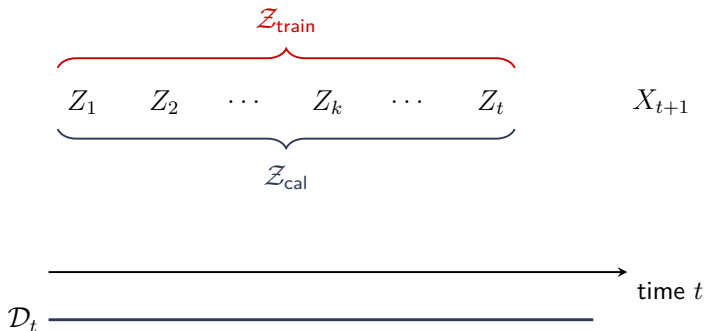
Full conformal for adaptively trained setting?

setup: fitted model $\hat{\mu}_t$ and score $s_t(\cdot, \cdot)$ depend on past data



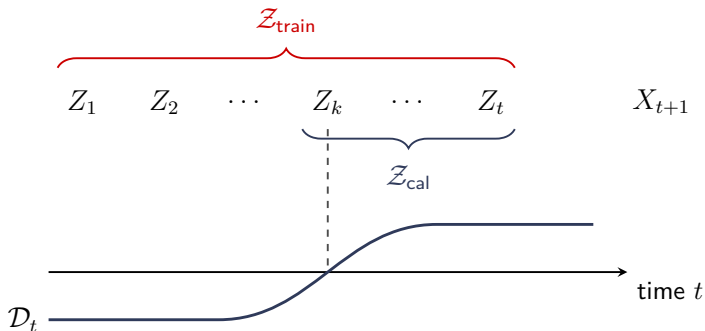
Choosing calibration set adaptively

no distribution drift:



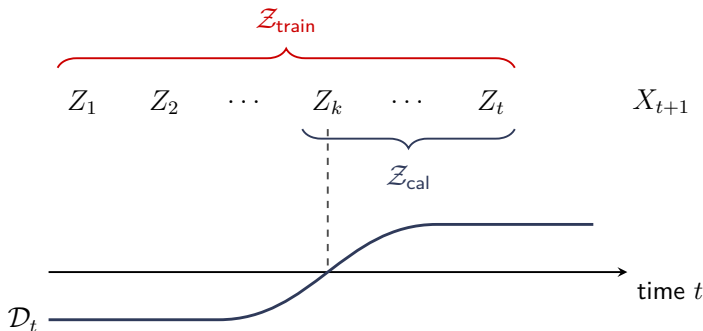
Choosing calibration set adaptively

when distribution drift occurs:



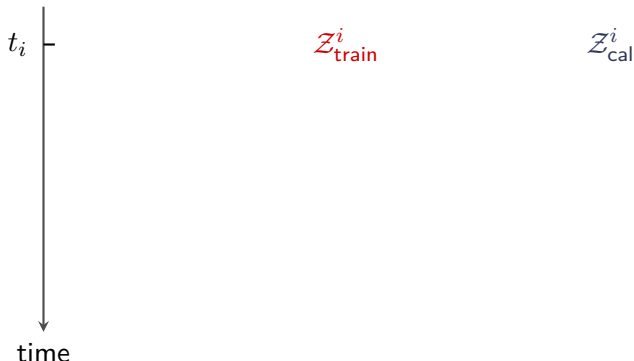
Choosing calibration set adaptively

when distribution drift occurs:



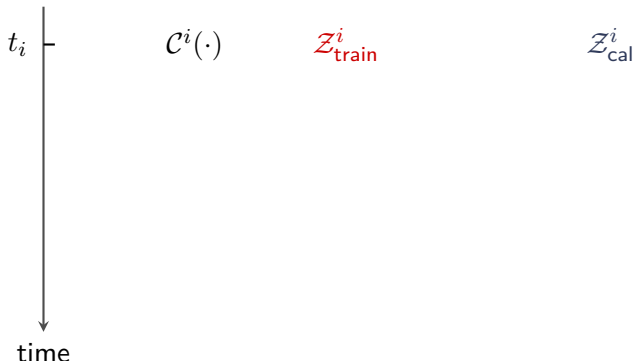
How to adjust $\mathcal{Z}_{\text{cal}}$ dynamically? **Drift detection!**

Proposed algorithm: DriftOCP-Full



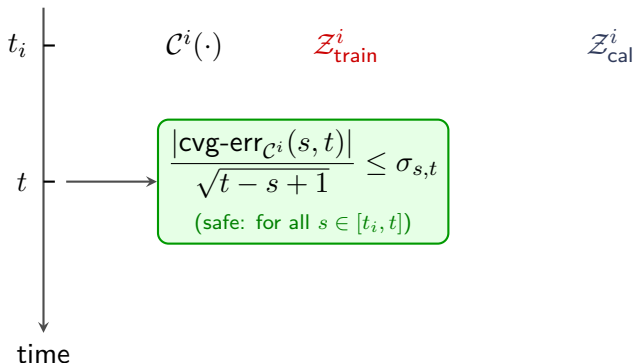
- 1. track history:** maintain an active window of data points starting from the latest reset point t_i

Proposed algorithm: DriftOCP-Full



1. track history: maintain an active window of data points starting from the latest reset point t_i

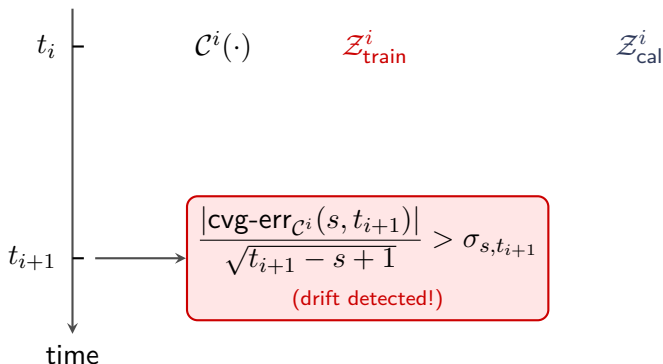
Proposed algorithm: DriftOCP-Full



2. scan for drift: continuously check if normalized empirical fluctuation stays below safe threshold $\sigma_{s,t}$

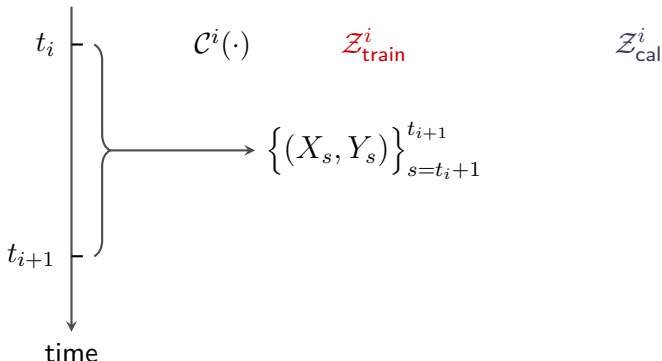
* typically $\sigma_{s,t}$ is chosen as $O(\log^3 t)$

Proposed algorithm: DriftOCP-Full



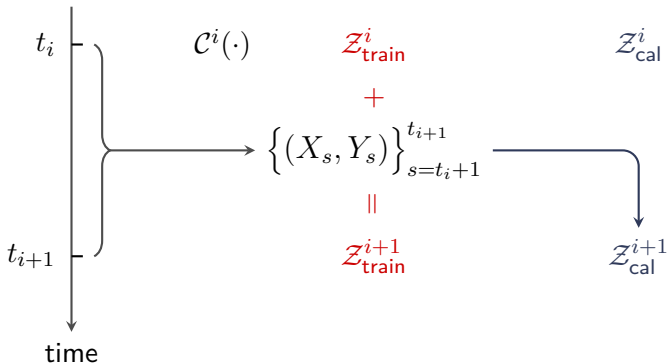
3. declare drift: if cvg-err exceeds the threshold for *any* past sub-interval, distribution shift is declared at time t_{i+1}

Proposed algorithm: DriftOCP-Full



4. form new calibration set: use data within $(t_i, t_{i+1}]$ as new calibration set

Proposed algorithm: DriftOCP-Full



4. form new calibration set: use data within $(t_i, t_{i+1}]$ as new calibration set; also augment the training set

Assumption 1 (Lipschitz continuity of conditional CDF)

For every time t , $x \in \mathcal{X}$, and any two $z, z' \in \mathbb{R}$,

$$\left| \mathbb{P}(Y_t \leq z \mid X_t = x) - \mathbb{P}(Y_t \leq z' \mid X_t = x) \right| \leq L_1 |z - z'|$$

Assumption 1 (Lipschitz continuity of conditional CDF)

For every time t , $x \in \mathcal{X}$, and any two $z, z' \in \mathbb{R}$,

$$\left| \mathbb{P}(Y_t \leq z \mid X_t = x) - \mathbb{P}(Y_t \leq z' \mid X_t = x) \right| \leq L_1 |z - z'|$$

Assumption 2 (stability of learning algorithm)

For any $\mathcal{Z}$, $\mathcal{Z}'$ w/ at most 1 different sample, the fitted model $\hat{\mu}$ obeys

$$\left| \hat{\mu}(x \mid \underbrace{\mathcal{Z}}_{\text{training data}}) - \hat{\mu}(x \mid \mathcal{Z}') \right| \leq L_2 / |\mathcal{Z}| \quad \text{for all } x \in \mathcal{X}$$

Assumption 1 (Lipschitz continuity of conditional CDF)

For every time t , $x \in \mathcal{X}$, and any two $z, z' \in \mathbb{R}$,

$$\left| \mathbb{P}(Y_t \leq z \mid X_t = x) - \mathbb{P}(Y_t \leq z' \mid X_t = x) \right| \leq L_1 |z - z'|$$

Assumption 2 (stability of learning algorithm)

For any $\mathcal{Z}$, $\mathcal{Z}'$ w/ at most 1 different sample, the fitted model $\hat{\mu}$ obeys

$$\left| \underbrace{\hat{\mu}(x \mid \mathcal{Z})}_{\text{training data}} - \hat{\mu}(x \mid \mathcal{Z}') \right| \leq L_2 / |\mathcal{Z}| \quad \text{for all } x \in \mathcal{X}$$

Examples of stable algorithms

- *constrained M-estimation*
- *linear stochastic approximation*
- *strongly convex optimization*
- *...*

Theoretical guarantees for DriftOCP-Full

Theorem 3 (L., Ren, Chen '26)

Under Assumptions 1-2 w/ $L = L_1 L_2 = \tilde{O}(1)$ and appropriate detection thresholds, DriftOCP-Full achieves

- change points setting: $\text{regret}_T \leq \tilde{O}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$*

Theoretical guarantees for DriftOCP-Full

Theorem 3 (L., Ren, Chen '26)

Under Assumptions 1-2 w/ $L = L_1 L_2 = \tilde{O}(1)$ and appropriate detection thresholds, DriftOCP-Full achieves

- 1. change points setting: $\text{regret}_T \leq \tilde{O}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$*
- 2. smooth drift setting: $\text{regret}_T \leq \tilde{O}\left((\text{TV}_T)^{\frac{1}{3}} T^{\frac{2}{3}} + \sqrt{T}\right)$*

Theoretical guarantees for DriftOCP-Full

Theorem 3 (L., Ren, Chen '26)

Under Assumptions 1-2 w/ $L = L_1 L_2 = \tilde{O}(1)$ and appropriate detection thresholds, DriftOCP-Full achieves

- 1. change points setting: $\text{regret}_T \leq \tilde{O}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$*
- 2. smooth drift setting: $\text{regret}_T \leq \tilde{O}\left((\text{TV}_T)^{\frac{1}{3}} T^{\frac{2}{3}} + \sqrt{T}\right)$*

- regret governed by cumulative total variation

$$\text{TV}_T := \sum_{t=1}^{T-1} \text{TV}(\mathcal{D}_t, \mathcal{D}_{t+1})$$

Theoretical guarantees for DriftOCP-Full

Theorem 3 (L., Ren, Chen '26)

Under Assumptions 1-2 w/ $L = L_1 L_2 = \tilde{O}(1)$ and appropriate detection thresholds, DriftOCP-Full achieves

- change points setting: $\text{regret}_T \leq \tilde{O}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$*
- smooth drift setting: $\text{regret}_T \leq \tilde{O}\left((\text{TV}_T)^{\frac{1}{3}} T^{\frac{2}{3}} + \sqrt{T}\right)$*

- regret governed by cumulative total variation

$$\text{TV}_T := \sum_{t=1}^{T-1} \text{TV}(\mathcal{D}_t, \mathcal{D}_{t+1})$$

- achieves similar regret as in pretrained score setting

Byproduct: training-cond. theory for full conformal

Theorem 4 (L., Ren, Chen '26)

Suppose $\mathcal{Z}_{\text{cal}} \subseteq \mathcal{Z}_{\text{train}}$ contain ind. samples, and Assumptions 1-2 (w/ $L = L_1 L_2$) hold. Conditional on $\mathcal{Z}_{\text{train}} \setminus \mathcal{Z}_{\text{cal}}$, for a target $Z = (X, Y) \sim \mathcal{D}$, full conformal yields, with probability at least $1 - \delta$,

$$\left| \mathbb{P}_{\mathcal{D}}(Y \in \mathcal{C}(X) \mid \mathcal{Z}_{\text{cal}}) - (1 - \alpha) \right|$$
$$\stackrel{\text{ignore log}}{\lesssim} \underbrace{\frac{L \sqrt{|\mathcal{Z}_{\text{cal}}|}}{|\mathcal{Z}_{\text{train}}|}}_{\text{stability cost}} + \underbrace{\frac{1}{\sqrt{|\mathcal{Z}_{\text{cal}}|}}}_{\text{concentration}} + \underbrace{\frac{1}{|\mathcal{Z}_{\text{cal}}|} \sum_{Z_l \in \mathcal{Z}_{\text{cal}}} \text{TV}(Z, Z_l)}_{\text{effect of distribution drift}}$$

Byproduct: training-cond. theory for full conformal

Theorem 4 (L., Ren, Chen '26)

Suppose $\mathcal{Z}_{\text{cal}} \subseteq \mathcal{Z}_{\text{train}}$ contain ind. samples, and Assumptions 1-2 (w/ $L = L_1 L_2$) hold. Conditional on $\mathcal{Z}_{\text{train}} \setminus \mathcal{Z}_{\text{cal}}$, for a target $Z = (X, Y) \sim \mathcal{D}$, full conformal yields, with probability at least $1 - \delta$,

$$\left| \mathbb{P}_{\mathcal{D}}(Y \in \mathcal{C}(X) \mid \mathcal{Z}_{\text{cal}}) - (1 - \alpha) \right|$$
$$\lesssim \underbrace{\frac{\text{ignore log } L \sqrt{|\mathcal{Z}_{\text{cal}}|}}{|\mathcal{Z}_{\text{train}}|}}_{\text{stability cost}} + \underbrace{\frac{1}{\sqrt{|\mathcal{Z}_{\text{cal}}|}}}_{\text{concentration}} + \underbrace{\frac{1}{|\mathcal{Z}_{\text{cal}}|} \sum_{Z_l \in \mathcal{Z}_{\text{cal}}} \text{TV}(Z, Z_l)}_{\text{effect of distribution drift}}$$

- beyond exchangeability

Byproduct: training-cond. theory for full conformal

Theorem 4 (L., Ren, Chen '26)

Suppose $\mathcal{Z}_{\text{cal}} \subseteq \mathcal{Z}_{\text{train}}$ contain ind. samples, and Assumptions 1-2 (w/ $L = L_1 L_2$) hold. Conditional on $\mathcal{Z}_{\text{train}} \setminus \mathcal{Z}_{\text{cal}}$, for a target $Z = (X, Y) \sim \mathcal{D}$, full conformal yields, with probability at least $1 - \delta$,

$$\left| \mathbb{P}_{\mathcal{D}}(Y \in \mathcal{C}(X) \mid \mathcal{Z}_{\text{cal}}) - (1 - \alpha) \right|$$
$$\lesssim \underbrace{\frac{\text{ignore log } L \sqrt{|\mathcal{Z}_{\text{cal}}|}}{|\mathcal{Z}_{\text{train}}|}}_{\text{stability cost}} + \underbrace{\frac{1}{\sqrt{|\mathcal{Z}_{\text{cal}}|}}}_{\text{concentration}} + \underbrace{\frac{1}{|\mathcal{Z}_{\text{cal}}|} \sum_{Z_l \in \mathcal{Z}_{\text{cal}}} \text{TV}(Z, Z_l)}_{\text{effect of distribution drift}}$$

- beyond exchangeability
- beyond permutation symmetry (use stability instead)
 - full conformal lacks training-cond. cvg w/o extra assumptions

• Bian & Barber. *Training-conditional coverage for distribution-free predictive inference*, Electron. J. Statist.

Comparison with prior work

Theorem 3.5 (Liang, Barber '23)

For γ -inflated full conformal with a symmetric algorithm $\mathcal{A}$ on n i.i.d. data $\mathcal{Z}$, with high prob.

$$\mathbb{P}_{\mathcal{D}}\left(Y \notin \hat{\mathcal{C}}^{\gamma\text{-CP}}(X) \mid \mathcal{Z}\right) - \alpha \lesssim \frac{1}{\sqrt{n}} + \frac{1}{\gamma^{1/3}} \left(\mathbb{E}_{\mathcal{D}} \left[|\hat{\mu}_n(X) - \hat{\mu}_{2n}(X)| \right] \right)^{\frac{1}{3}}$$

Comparison with prior work

Theorem 3.5 (Liang, Barber '23)

For γ -inflated full conformal with a symmetric algorithm $\mathcal{A}$ on n i.i.d. data $\mathcal{Z}$, with high prob.

$$\mathbb{P}_{\mathcal{D}}\left(Y \notin \hat{\mathcal{C}}^{\gamma\text{-CP}}(X) \mid \mathcal{Z}\right) - \alpha \lesssim \frac{1}{\sqrt{n}} + \frac{1}{\gamma^{1/3}} \left(\mathbb{E}_{\mathcal{D}} \left[|\hat{\mu}_n(X) - \hat{\mu}_{2n}(X)| \right] \right)^{\frac{1}{3}}$$

	Liang & Barber	ours (L., Ren, Chen)
coverage bound	one-sided (lower bound)	two-sided ($ \cdot \lesssim \dots$)
data assumption	strict i.i.d.	allows distribution drift
algorithm $\mathcal{A}$	must be symmetric	no symmetry required
stability metric	β^{in} (hard to verify)	replace-one (easy to verify)
CP procedure	requires γ -inflation	standard full conformal

• Liang & Barber. *Algorithmic stability implies training-conditional coverage for distribution-free prediction methods*, Ann. Statist.

Empirical performance: adaptively trained scores

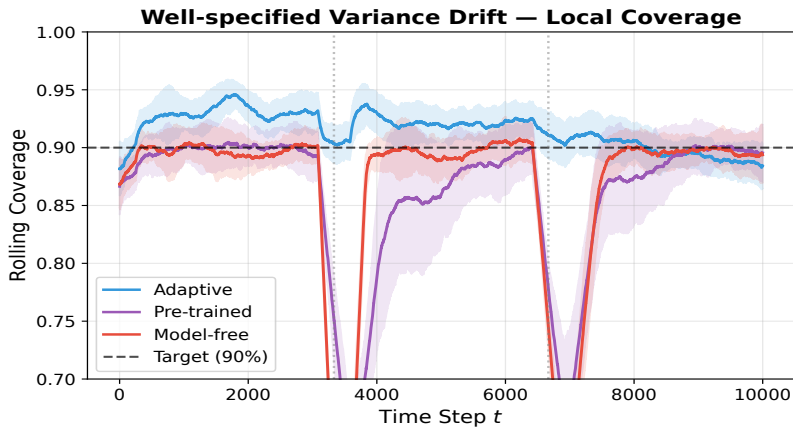


Figure 2: local coverage (well-specified variance drift)

Empirical performance: adaptively trained scores

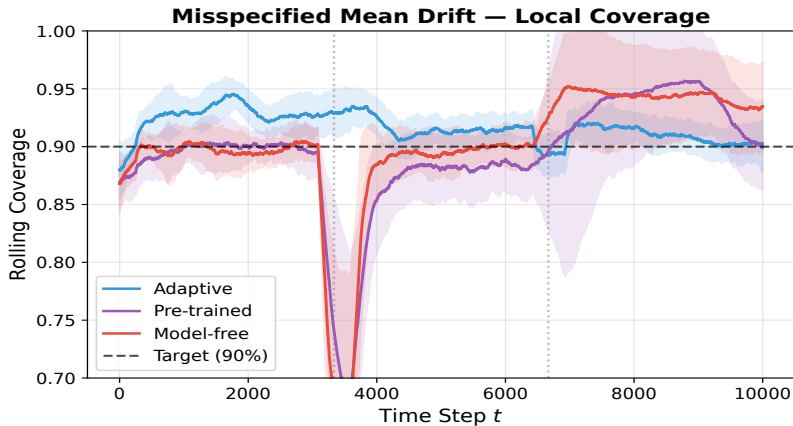


Figure 2: local coverage (misspecified mean drift)

When does DriftOCP-Full achieve optimal regret?

Minimax lower bound

- *admissible algorithm class* $\mathcal{P}_K$: $\mathcal{C}_t = \pi_t(\text{past data, rand seed})$
and is union of K intervals

Minimax lower bound

- *admissible algorithm class* $\mathcal{P}_K$: $\mathcal{C}_t = \pi_t(\text{past data, rand seed})$ and is union of K intervals
- *distribution classes*

$$\mathcal{L}_{\text{cp}} := \left\{ \{\mathcal{D}_t\} : \text{at most } N^{\text{cp}} \text{ change points} \right\}$$

$$\mathcal{L}_{\text{sd}} := \left\{ \{\mathcal{D}_t\} : \sum_{t=1}^{T-1} \text{TV}(\mathcal{D}_t, \mathcal{D}_{t+1}) \leq \text{TV}_T \right\}$$

Minimax lower bound

Theorem 5 (L., Ren, Chen '26)

For any admissible algorithm $\pi \in \mathcal{P}_K$ with $K = O(1)$:

- *change-point:* $\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_{\text{cp}}} \text{regret}_T = \tilde{\Omega}\left(\sqrt{(N^{\text{cp}} + 1)T}\right)$

Minimax lower bound

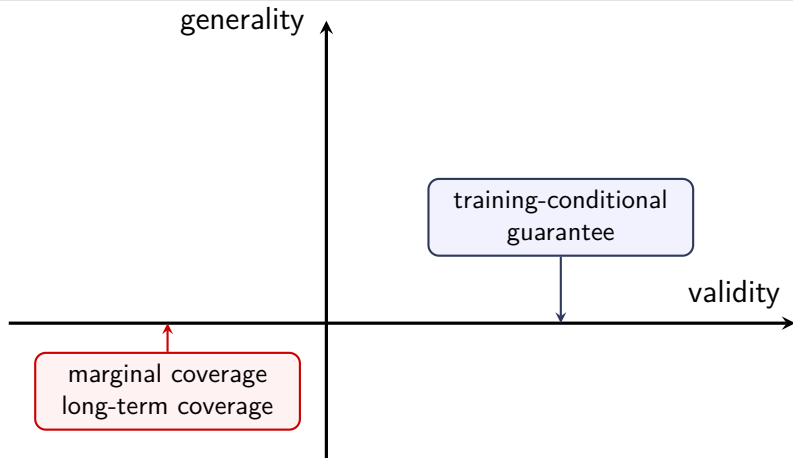
Theorem 5 (L., Ren, Chen '26)

For any admissible algorithm $\pi \in \mathcal{P}_K$ with $K = O(1)$:

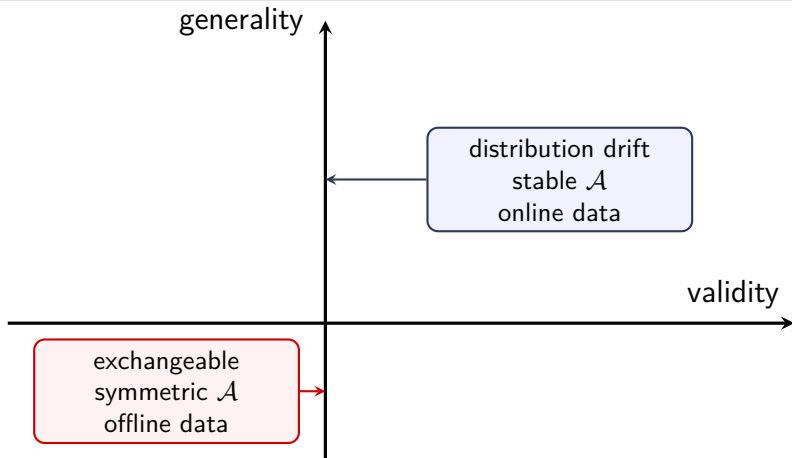
- *change-point*: $\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_{\text{cp}}} \text{regret}_T = \tilde{\Omega} \left(\sqrt{(N^{\text{cp}} + 1)T} \right)$
- *smooth drift*: $\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_{\text{sd}}} \text{regret}_T = \tilde{\Omega} \left((\text{TV}_T)^{\frac{1}{3}} T^{\frac{2}{3}} + \sqrt{T} \right)$
- *our DriftOCP-Full is minimax optimal when $K, L = \tilde{O}(1)$!*

- 1 Long-term coverage vs. training-conditional regret
- 2 Problem formulation
- 3 Online conformal with pretrained scores
- 4 Online conformal with adaptively trained scores
- 5 Concluding remarks**

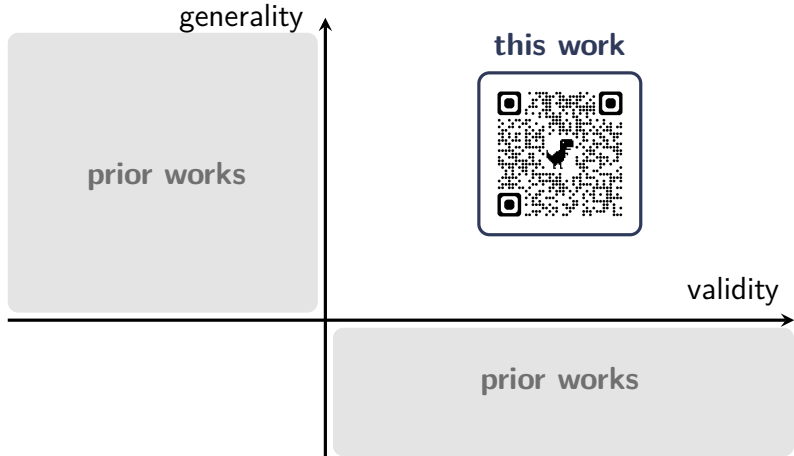
Concluding remarks



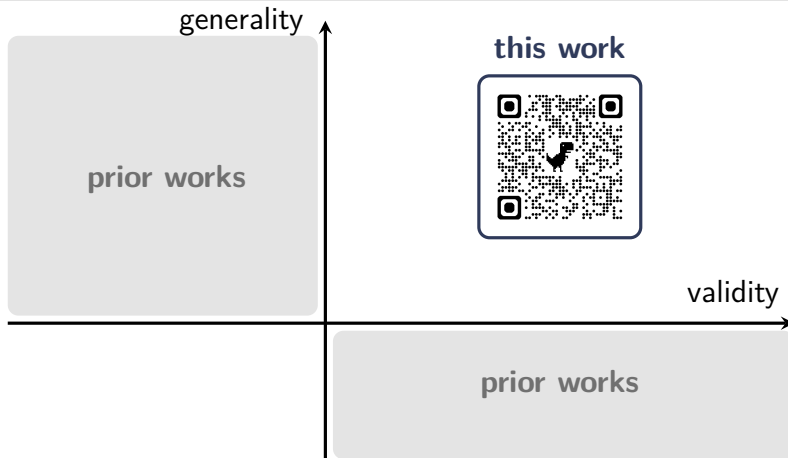
Concluding remarks



Concluding remarks

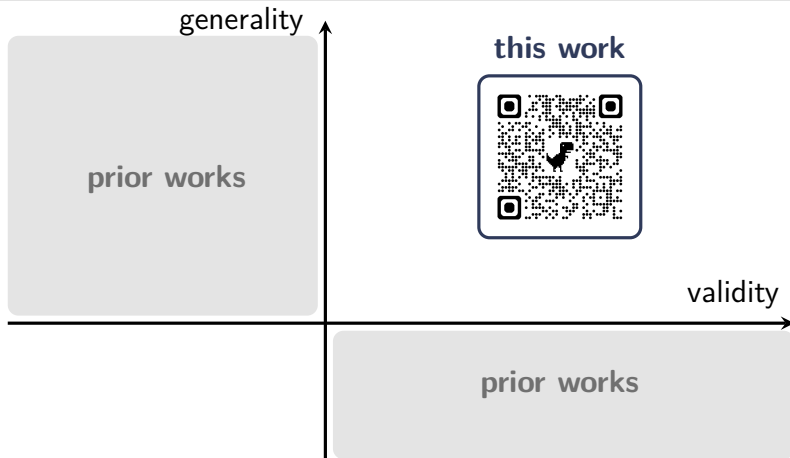


Concluding remarks



- new online conformal methods (both split and full) that attain **optimal training-cond. regret** under independent data

Concluding remarks



- new online conformal methods (both split and full) that attain **optimal training-cond. regret** under independent data
- **future directions:** beyond independence, beyond stability, ...

Training-cond. regret vs. long-term cvg

$$\text{LTC}_T \approx \left| \frac{1}{T} \sum_{t=1}^T \left(\mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right) \right|$$

Training-cond. regret vs. long-term cvg

$$\text{LTC}_T \approx \left| \frac{1}{T} \sum_{t=1}^T \left(\mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right) \right|$$

$$\frac{1}{T} \text{regret}_T = \frac{1}{T} \sum_{t=1}^T \left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right|$$

Training-cond. regret vs. long-term cvg

$$\text{LTC}_T \approx \left| \frac{1}{T} \sum_{t=1}^T \left(\mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right) \right|$$

| $\wedge$

$$\frac{1}{T} \text{regret}_T = \frac{1}{T} \sum_{t=1}^T \left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right|$$

Training-cond. regret vs. long-term cvg

$$\text{LTC}_T \approx \left| \frac{1}{T} \sum_{t=1}^T \left(\mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right) \right|$$

$\wedge$

$$\frac{1}{T} \text{regret}_T = \frac{1}{T} \sum_{t=1}^T \left| \mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{F}_{t-1}) - (1 - \alpha) \right|$$

Conclusion: regret_T avoids the cancellation of errors across time.

Satisfying the Stability Assumption

Assumption 2 ($\mathcal{O}(1/m)$ stability) naturally holds for several canonical learning algorithms.

- ✓ **Constrained M-estimation:** Empirical Risk Minimization over a constrained space Θ .

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^m \ell(\theta; Z_i)$$

Satisfying the Stability Assumption

Assumption 2 ($\mathcal{O}(1/m)$ stability) naturally holds for several canonical learning algorithms.

- ✓ **Constrained M-estimation:** Empirical Risk Minimization over a constrained space Θ .

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^m \ell(\theta; Z_i)$$

- ✓ **Linear Stochastic Approximation:** Online updates for streaming least squares (with decaying step size η_t).

$$\hat{\theta}_m = \hat{\theta}_{m-1} - \eta_m (X_m X_m^\top \hat{\theta}_{m-1} - X_m Y_m)$$

Satisfying the Stability Assumption

Assumption 2 ($\mathcal{O}(1/m)$ stability) naturally holds for several canonical learning algorithms.

- ✓ **Constrained M-estimation:** Empirical Risk Minimization over a constrained space Θ .

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^m \ell(\theta; Z_i)$$

- ✓ **Linear Stochastic Approximation:** Online updates for streaming least squares (with decaying step size η_t).

$$\hat{\theta}_m = \hat{\theta}_{m-1} - \eta_m (X_m X_m^\top \hat{\theta}_{m-1} - X_m Y_m)$$

- ✓ **Stochastic Strongly Convex Optimization:** Projected SGD for strongly convex and smooth objectives $l(\cdot, \cdot)$.

$$\hat{\theta}_m = \Pi_{\Theta} \left[\hat{\theta}_{m-1} - \eta_m \nabla \ell(\hat{\theta}_{m-1}; Z_m) \right]$$

Model-free minimax lower bounds

Theorem 6 (Minimax lower bounds)

For any admissible algorithm $\pi \in \mathcal{P}_K$,

Change points:

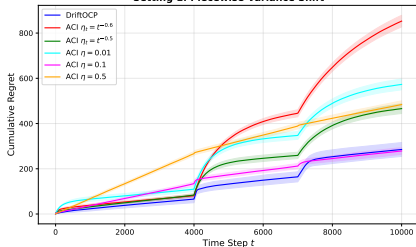
$$\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_3(N^{\text{cp}})} \text{regret}_T^\pi = \tilde{\Omega} \left(\min \left\{ \sqrt{(N^{\text{cp}} + 1)T}, \frac{T}{\sqrt{K}} \right\} \right)$$

Smooth drift:

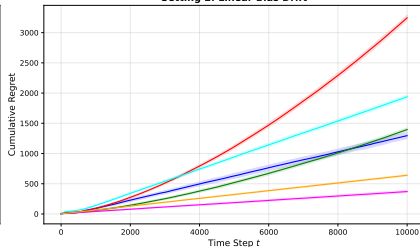
$$\sup_{\{\mathcal{D}_t\} \in \mathcal{L}_4(\text{TV}_T)} \text{regret}_T^\pi = \tilde{\Omega} \left(\min \left\{ \sqrt{T} + (\text{TV}_T)^{\frac{1}{3}} T^{\frac{2}{3}} K^{-\frac{1}{6}}, \frac{T}{\sqrt{K}} \right\} \right)$$

Empirical performance

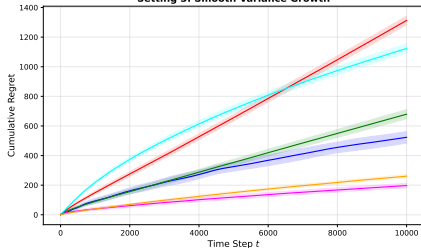
Setting 1: Piecewise Variance Shift



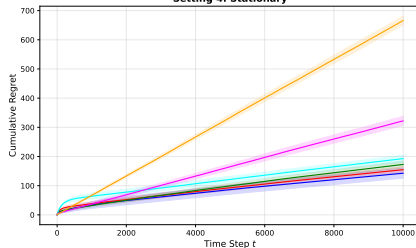
Setting 2: Linear Bias Drift



Setting 3: Smooth Variance Growth

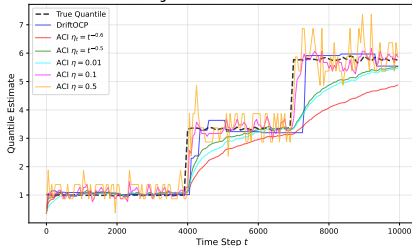


Setting 4: Stationary

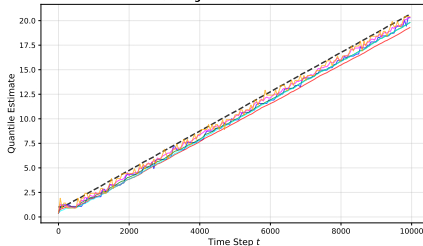


Empirical performance (cont'd)

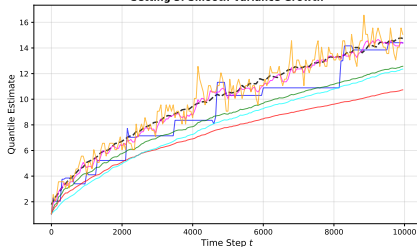
Setting 1: Piecewise Variance Shift



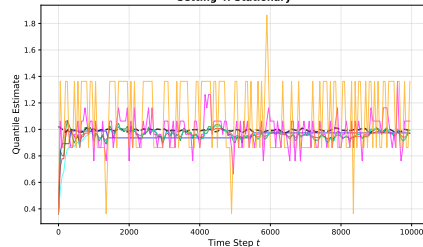
Setting 2: Linear Bias Drift



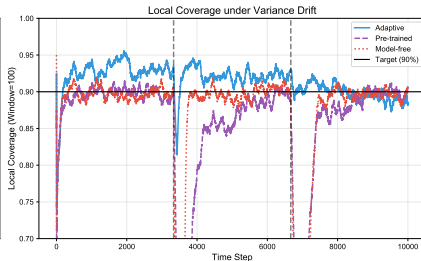
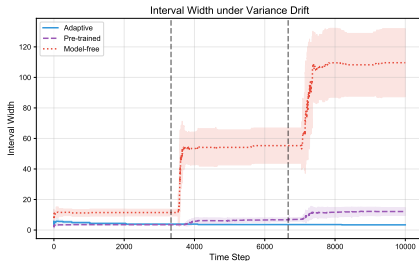
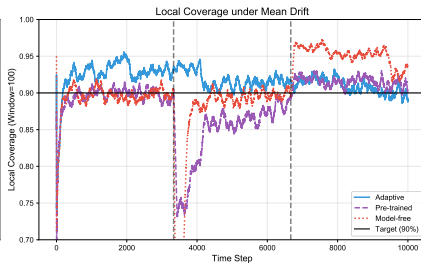
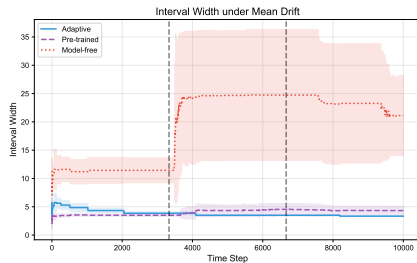
Setting 3: Smooth Variance Growth



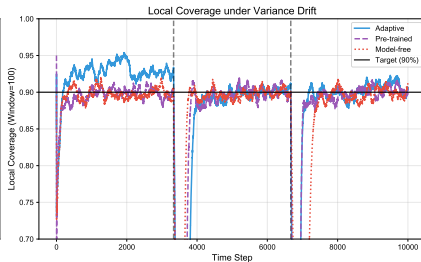
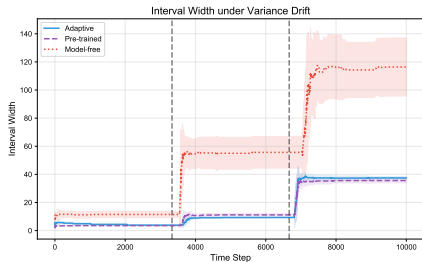
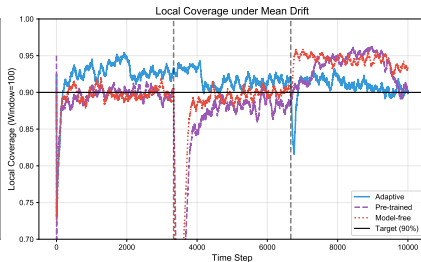
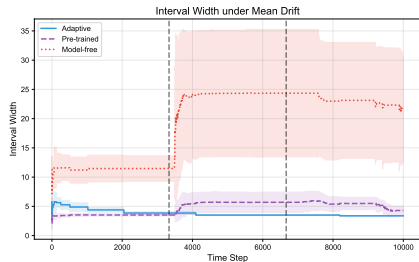
Setting 4: Stationary



Empirical performance: adaptively trained scores

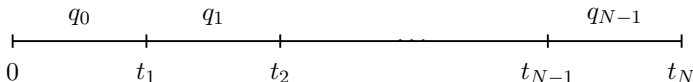


Empirical performance: adaptively trained scores (cont'd)



Proof sketch of upper bound in pretrained score

simple case: q_t^* monotonically decreasing



$$\begin{aligned}\text{regret in } i\text{-th stage} &= \sum_{t=t_{i-1}}^{t_i} \mathbb{P}(S_t \leq q_{i-1} \mid Z_{t-1}) - (1 - \alpha) \\ &= \text{cvg-err}_{q_{i-1}}^*(t_{i-1}, t_i) \approx \text{cvg-err}_{q_{i-1}}(t_{i-1}, t_i) + \mathcal{O}\left(\sqrt{t_i - t_{i-1}}\right) \\ &= \mathcal{O}\left(\sqrt{t_i - t_{i-1}}\right)\end{aligned}$$

$$\text{regret}_T \lesssim \sum_{i=1}^N \sqrt{t_i - t_{i-1}}, \quad N = \begin{cases} N^{\text{cp}}, & \text{change-point;} \\ (\text{KS}_T)^{1/3} T^{1/6}, & \text{smooth drift} \end{cases}$$

Proof sketch of lower bound in pretrained score

- **1. Time Partition:** Partition the entire horizon $[T]$ into m consecutive blocks $\mathcal{I}_1, \dots, \mathcal{I}_m$.
- **2 & 3. Hard Instance Construction:** $H_j \sim \text{Ber}(0.5)$:

$$\mathcal{D}_t^{\text{score}} = \begin{cases} \text{Exp}(1), & H_j = 0 \\ \text{Exp}(1 + \varepsilon), & H_j = 1 \end{cases} \quad \text{for all } t \in \mathcal{I}_j.$$

- **4. Expected Coverage Error:** When $4\varepsilon^2|\mathcal{I}_j| \leq 1$:

$$\mathbb{E} \left[\left| \mathbb{P}_H(s > q_t \mid q_t) - \alpha \right| \right] \geq \frac{\alpha^{1+\varepsilon} \varepsilon \log(1/\alpha)}{16}.$$

- **5. Aggregation:** Summing the errors over all m blocks:

$$\mathbb{E} \left[\sum_{t=1}^T \left| \mathbb{P}(s_t > q_t \mid q_t) - \alpha \right| \right] \geq \frac{T\varepsilon\alpha^{1+\varepsilon}}{16} \log \frac{1}{\alpha}.$$

Proof sketch of upper bound in adaptively trained score

McDiarmid's Inequality (Simplified)

If : $|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq \frac{L}{n}$ Then,

$$\mathbb{P}\left(\left|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]\right| \gtrsim \frac{L}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}}\right) \geq 1 - \delta.$$

$$\hat{\mu}(\cdot \mid \mathcal{Z}_{\text{train}}) \approx \tilde{\mu}(\cdot) + \mathcal{O}\left(\frac{L}{\sqrt{|\mathcal{Z}_{\text{train}}|}}\right)$$

$$|Y - \hat{\mu}(X \mid \mathcal{Z}_{\text{train}})| \approx \underbrace{|Y - \tilde{\mu}(X)|}_{\text{split setting!}}$$

Why Restrict to $\mathcal{P}_K$?

An “Irregular” Construction: Suppose $Y \in [0, 1]$. For each $n \geq 1$, define:

$$\mathcal{C}_n := \bigcup_{i=0}^{n-1} \left[\frac{i}{n}, \frac{i + (1 - \alpha)}{n} \right]$$

Intuition: Partition $[0, 1]$ into n equal subintervals and retain the same $(1 - \alpha)$ -fraction of each subinterval.

Proposition 4.2

Let $\{\mathcal{C}_n\}_{n \geq 1}$ be defined as above. If the distribution $\mathcal{D}$ of Y on $[0, 1]$ admits a Riemann-integrable density, then

$$\lim_{n \rightarrow \infty} \left| \mathbb{P}(Y \in \mathcal{C}_n) - (1 - \alpha) \right| = 0.$$

Proof sketch of lower bound in adaptively trained score

