Optimal training-conditional regret for online conformal prediction

Jiadong Liang* Zhimei Ren* Yuxin Chen*

February 18, 2026

Abstract

We study online conformal prediction for non-stationary data streams subject to unknown distribution drift. While most prior work studied this problem under adversarial settings and/or assessed performance in terms of gaps of time-averaged marginal coverage, we instead evaluate performance through training-conditional cumulative regret. We specifically focus on independently generated data with two types of distribution shift: abrupt change points and smooth drift.

When non-conformity score functions are pretrained on an independent dataset, we propose a split-conformal–style algorithm that leverages drift detection to adaptively update calibration sets, which provably achieves minimax-optimal regret. When non-conformity scores are instead trained online, we develop a full-conformal–style algorithm that again incorporates drift detection to handle non-stationarity; this approach relies on stability—rather than permutation symmetry—of the model-fitting algorithm, which is often better suited to online learning under evolving environments. We establish non-asymptotic regret guarantees for our online full conformal algorithm, which match the minimax lower bound under appropriate restrictions on the prediction sets. Numerical experiments corroborate our theoretical findings.

Keywords: online conformal prediction, training-conditional regret, distribution drift, minimax optimality

Contents

*Department of Statistics and Data Science, the Wharton School, University of Pennsylvania; email: {jdl97,zren,yuxinc}@wharton.upenn.edu.

1 Introduction

Conformal prediction, also known as conformal inference, has emerged as a versatile, distribution-free framework for quantifying uncertainty in modern data science (Vovk et al., 1999; Papadopoulos et al., 2002; Vovk et al., 2005; Angelopoulos et al., 2023, 2024b). What sets it apart is its ability to offer rigorous, finite-sample coverage guarantees under minimal distribution assumptions, allowing practitioners to treat complex machine learning models as black boxes while still producing reliable measures of uncertainty. In its classical formulation, we observe n training data taking the form of n feature-response pairs $\{(X_i, Y_i)\}_{1\le i\le n} \subset \mathcal{X} \times \mathbb{R}$, and are given a test point $X_{n+1} \in \mathcal{X}$ for which the corresponding response Y_{n+1} is unknown. The aim is to construct a prediction set $\widehat{\mathcal{C}}(X_{n+1})$ that is likely to cover Y_{n+1}. Conformal prediction achieves this objective in a distribution-free fashion, provided the data $\{(X_i, Y_i)\}_{1\le i\le n+1}$ are *exchangeable* (Angelopoulos et al., 2023, 2024b).

While the ability to accommodate exchangeable data applies to wide-ranging practical scenarios, there is no shortage of scenarios that naturally violate exchangeability. One notable example arises when the data distributions drift over time, as is often the case with sequential or online data (Zhou et al., 2025; Fannjiang et al., 2022). This motivates a flurry of recent studies exploring online conformal prediction, with the objective to extend the conformal prediction framework to accommodate sequentially arriving data streams (e.g., Vovk et al. (2009); Weinstein and Ramdas (2020); Gibbs and Candes (2021); Bastani et al. (2022); Zaffran et al. (2022); Bhatnagar et al. (2023); Lin et al. (2022); Feldman et al. (2022); Auer et al. (2023); Sun and Yu (2023); Xu and Xie (2023b,a); Xu et al. (2024); Gibbs and Candès (2024); Han et al. (2024a); Angelopoulos et al. (2023, 2024a, 2025); Bao et al. (2024); Lee and Matni (2024); Yang et al. (2024); Podkopaev et al. (2024); Su et al. (2024); Zhang et al. (2024b); Ramalingam et al. (2025); Sale and Ramdas (2025); Humbert et al. (2025)).

1.1 Online conformal prediction

Setting the stage, consider a sequential data stream $\{(X_t, Y_t)\}_{1\le t\le T}$ generated by a dynamic process, where $X_t \in \mathcal{X}$ denotes the feature (or covariate) at time t and $Y_t \in \mathbb{R}$ the corresponding response. The data-generating distribution is allowed to drift over time; namely, the distribution of (X_t, Y_t), denoted by $\mathcal{D}_t$, may vary with t. At each time t, the task is to use the previously observed data $\{(X_s, Y_s)\}_{s<t}$, together with the newly observed feature X_t, to construct a prediction set $\mathcal{C}_t(X_t)$ that is likely to contain the as-yet-unobserved response Y_t. More precisely, for a prescribed miscoverage level $\alpha \in (0, 1)$, a desirable prediction set $\mathcal{C}_t(X_t)$ would satisfy

$$\mathbb{P}\left\{ Y_t \in \mathcal{C}_t(X_t) \;\middle|\; \{(X_s, Y_s)\}_{s<t} \right\} \ge 1 - \alpha. \tag{1}$$

Central to conformal prediction is the non-conformity score function $s_t(\cdot, \cdot)$, which is computed at time t and may sometimes depend on past observations $\{(X_\tau, Y_\tau)\}_{\tau:\tau<t}$. For the most part, the score $s_t(x, y)$ measures the extent to which a data point $(x, y) \in \mathcal{X} \times \mathbb{R}$ deviates from the prediction of a fitted model. A canonical example is the absolute residual score $s_t(x, y) = |y - \widehat{\mu}_t(x)|$, where $\widehat{\mu}_t(\cdot)$ denotes a predictive model trained by an arbitrary machine learning algorithm (for instance, a neural network, or a nonparametric estimator). A widely studied class of prediction intervals takes the form

$$\mathcal{C}_t(x) \coloneqq \{y : s_t(x, y) \le q_t\} \tag{2}$$

for some adaptively chosen threshold q_t, in which case prediction interval construction amounts to dynamically adjusting $\{q_t\}$ given the non-conformity scores.

The online nature of the above problem has motivated a recent line of work to reframe (1) as an online decision-making task and leverage techniques from online learning to address it. A prominent example is *Adaptive Conformal Inference (ACI)*, proposed by Gibbs and Candes (2021). In a nutshell, the ACI algorithm sequentially calibrates the quantile estimates via the iterative update rule:

$$q_{t+1} = q_t + \eta_t\big(\mathbb{1}\{s_t(X_t, Y_t) > q_t\} - \alpha\big), \tag{3}$$

which can be interpreted as an instance of the online subgradient method applied to optimize the quantile loss (or pinball loss).

1.2 Prior coverage guarantees and their inadequacy

To establish theoretical validity, a substantial body of prior work developed coverage guarantees for online conformal prediction methods. For instance, Gibbs and Candes (2021) demonstrated that the ACI algorithm achieves some sort of *time-averaged coverage* without imposing any assumption on the data-generating mechanism; more formally, they proved that, with a suitable constant learning rate schedule, ACI satisfies

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\big(Y_t \in \mathcal{C}_t(X_t)\big)}_{\text{empirical long-term coverage frequency}} \;\to\; 1 - \alpha \tag{4}$$

as T grows, which holds even when the data stream is generated adversarially. Building on this result, subsequent work has extended time-averaged coverage results to a broader family of algorithms (e.g., Zaffran et al. (2022); Angelopoulos et al. (2024a); Bhatnagar et al. (2023); Zhang et al. (2024a)).

Note, however, that controlling the empirical long-term coverage frequency in (4) does not, by itself, preclude vacuous solutions. As noted in prior studies (e.g., Bastani et al. (2022); Bhatnagar et al. (2023)) and further elaborated in Section 2.3, one can easily construct prediction sets that fulfill property (4) while failing to incorporate any information of the underlying data distributions. In other words, achieving convergence of empirical long-term coverage frequency does not ensure reliable coverage at any individual time, nor does it guarantee that the prediction sets are informative and efficient.

To remedy the above issue, a line of subsequent work (e.g., Bhatnagar et al. (2023); Gibbs and Candès (2024); Hajihashemi and Shen (2024); Ramalingam et al. (2025); Zhang et al. (2024a,b)) shifted focus towards *regret-based analysis*, drawing heavily from the online learning literature (Shalev-Shwartz, 2012;

Hazan et al., 2016). While multiple notions of regret have been explored in this strand of work, they are primarily formulated for adversarial online settings, where the underlying data-generating distributions are left completely unspecified. Consequently, these regret metrics often lack a direct correspondence with standard conformal validity targets, such as training-conditional coverage. Furthermore, several prior works (e.g., Bhatnagar et al. (2023); Hajihashemi and Shen (2024); Ramalingam et al. (2025)) evaluated the cumulative performance gap relative to global quantile optimized in hindsight (i.e., the quantile computed based on all data), which is, however, not well-suited to non-stationary environments with drifting data distributions.

It is important to emphasize again that the adoption of empirical long-term coverage frequency and adversarial regret largely stems from the objective to dispense with distributional assumptions, thereby maximizing the "distribution-free" nature of online predictive inference. However, if one is willing to impose more structure on the data-generating mechanism, it may become possible to derive coverage guarantees that align more closely with classical validity notions. While several prior work (Gibbs and Candes, 2021; Han et al., 2024a; Zaffran et al., 2022; Xu and Xie, 2023a; Angelopoulos et al., 2024a; Humbert et al., 2025) had investigated more specialized settings—such as independent data with drifting distributions, hidden Markov models—the optimality of the resulting theoretical guarantees remain largely unexplored.

1.3 This paper

In this work, we make progress by focusing on the following non-adversarial scenario:

- *A non-adversarial setting with independent data:* The data $\{(X_t, Y_t)\}_{1\le t\le T}$ are *independently* generated but otherwise distribution-free. The distribution of (X_t, Y_t), denoted by $\mathcal{D}_t$, is allowed to drift over time, but the predictive inference algorithm has no prior knowledge of the distributional drift.

The independence assumption enables us to move beyond performance metrics like time-averaged marginal coverage and adversarial regret, and instead adopt a regret metric that aligns more closely with classical statistical validity. Informally, we focus on the following *training-conditional* cumulative regret metric

$$\mathsf{regret}_T := \sum_{t=1}^{T} \mathbb{E}\left[\left|\mathbb{P}\left(Y_t \in \mathcal{C}_t(X_t) \;\middle|\; \text{past data}\right) - (1-\alpha)\right|\right], \tag{5}$$

whose precise definition is given in Section 2.2. This metric measures, at each time t, the deviation of the coverage probability conditional on past observations from the target level, and then aggregates these deviations over time. The emphasis on training-conditional (sample-conditional) validity is standard in conformal prediction (e.g., Vovk, 2012; Bian and Barber, 2023; Amann et al., 2023; Liang and Barber, 2025).

Within this framework, we pay particular attention to two forms of distribution drift: (i) *the change-point setting*, where the data distribution is piecewise stationary with several abrupt change points; (ii) *the smooth drift setting*, where the distributions evolve continuously and smoothly over time, subject to an upper bound on its aggregate variation. Note that the predictive inference algorithm operates without prior knowledge of the drift structure. Our main contributions are summarized as follows.

Online conformal prediction with pretrained scores. Consider first the scenario in which the non-conformity score functions are pretrained on a separate, independent dataset—a common setting in online conformal prediction where split-conformal-style methods are naturally applicable. We propose an online conformal prediction algorithm, dubbed DriftOCP (see Algorithm 2), which leverages drift detection subroutines to adaptively update calibration sets—the set of data used for calibrating q_t—over time. Our algorithm is computationally lightweight, horizon-independent, and adapts efficiently to the distribution drift. We provide non-asymptotic theoretical guarantees by establishing regret upper bounds for DriftOCP that match the minimax lower bounds (up to a logarithmic factor) in both the change-point and smooth drift settings. Numerical experiments across a range of distribution-shift scenarios further demonstrate the efficacy of DriftOCP, showing that it adapts effectively to diverse data-generating mechanisms.

Online conformal prediction with adaptively trained scores. Next, consider a more challenging scenario in which both the predictive models and the non-conformity score functions are trained online, potentially depending on past observations. To enhance data efficiency without data splitting, we adopt the full

conformal paradigm, and put forward an online full conformal prediction algorithm called DriftOCP-full (see Algorithm 4), which integrates drift detection subroutines to tackle non-stationarity. Rather than assuming permutation symmetry of the model fitting algorithm—which is often violated in online learning—we focus instead on stable learning algorithms, and establish non-asymptotic upper bounds on the training-conditional cumulative regret of DriftOCP-full. We further demonstrate the optimality of our approach by deriving matching minimax lower bounds (up to a log factor) under appropriate restrictions on the prediction sets. Notably, our training-conditional lower bound applies universally to *all* prediction methods regardless of their specific construction—a result that was previously out of reach. Empirically, we benchmark several conformal prediction methods and validate the plausibility of constructing prediction sets using sequentially fitted models.

1.4 Notation

We now gather a set of notations used throughout the paper. For any $a, b \in \mathbb{R}$, denote $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$, $(a)_+ = a \vee 0$, and $(a)_- = a \wedge 0$. For any integer n, let $[n] \coloneqq \{1, \ldots, n\}$. For $x \in \mathbb{R}$, we use $\lceil x \rceil$ to denote the smallest integer greater than or equal to x, and $\lfloor x \rfloor$ the largest integer less than or equal to x. For two nonnegative functions f and g, we write $f \lesssim g$ (equivalently, $f = O(g)$ and $g = \Omega(f)$) if there exists a universal constant $C > 0$ such that $f \leq Cg$. We write $f \gtrsim g$ if $g \lesssim f$, and $f \asymp g$ if both $f \lesssim g$ and $g \lesssim f$ hold. The notation $f = \widetilde{O}(g)$ and $g = \widetilde{\Omega}(f)$ is defined analogously, up to additional logarithmic factors. For the set $\mathbb{R}$ of real numbers, we denote by $\mathcal{B}(\mathbb{R})$ the Borel sets on it. We denote by $\|v\|_2$ the Euclidean norm of a vector $v \in \mathbb{R}^d$. For a matrix $A \in \mathbb{R}^{m \times d}$, we use $\|A\| \coloneqq \sup_{\|x\|_2 = 1} \|Ax\|_2$ for its spectral norm. For two probability distributions P and Q defined on a measurable space $(\Xi, \mathcal{F})$, we denote by $\mathsf{TV}(P, Q)$ their total-variation (TV) distance, i.e.,

$$\mathsf{TV}(P, Q) \coloneqq \sup_{A \in \mathcal{F}} \big\{ \, |P(A) - Q(A)| \, \big\}.$$

Suppose P and Q admit probability density functions p and q on Ξ, respectively. We define the Kullback–Leibler (KL) divergence from Q to P by

$$\mathsf{KL}(P \,\|\, Q) \coloneqq \int_{\Xi} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x,$$

whenever the integral is well-defined. Furthermore, if P and Q are two probability distributions defined on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$, we denote by $\mathsf{KS}(P, Q)$ their Kolmogorov–Smirnov (KS) distance, i.e.,

$$\mathsf{KS}(P, Q) \coloneqq \sup_{x \in \mathbb{R}} \Big\{ \, \big| P\big((-\infty, x]\big) - Q\big((-\infty, x]\big) \big| \, \Big\}. \tag{6}$$

For random objects $Z \sim P$ and $\widetilde{Z} \sim Q$, we overload the notation by letting $\mathsf{TV}(Z, \widetilde{Z})$, $\mathsf{KL}(Z \,\|\, \widetilde{Z})$ and $\mathsf{KS}(Z, \widetilde{Z})$ denote $\mathsf{TV}(P, Q)$, $\mathsf{KL}(P \,\|\, Q)$ and $\mathsf{KS}(P, Q)$, respectively. Also, for any sequence of objects $\{Z_i\}_{i \geq 1}$, we adopt the notation $Z_{k:m} = \{Z_k, \ldots, Z_m\}$ for any $m \geq k \geq 1$.

2 Problem formulation and key metrics

2.1 Settings

Consider a sequence of T *independent* data points arriving sequentially, denoted by $Z_t = (X_t, Y_t) \in \mathcal{X} \times \mathbb{R}$, $t = 1, \ldots, T$, where the set $\mathcal{X}$ represents the feature domain. At each time t, the feature X_t is revealed first, and the response Y_t becomes available after a prediction set has been formed. Throughout this paper, we use $Z_{1:t} = \{(X_s, Y_s)\}_{s \leq t}$ to denote the set of all data up to time t.

Procedure. An online conformal prediction procedure operates as follows. Given the data $\{(X_s, Y_s)\}_{s<t}$ observed prior to time t and the newly arrived feature X_t, an online conformal prediction algorithm—denoted by π—seeks to construct a prediction set

$$\mathcal{C}_t \;=\; \mathcal{C}_t^{\pi}(X_t; \{(X_s, Y_s)\}_{s<t}) \subseteq \mathbb{R},$$

designed to contain Y_t with probability exceeding—and ideally close to—the target level $1-\alpha$. Here, we often write $\mathcal{C}_t$ for brevity if it is clear from the context. The set $\mathcal{C}_t$ is typically built with the aid of a non-conformity score function $s_t(\cdot,\cdot)$ along with a fitted predictive model $\widehat{\mu}_t(\cdot)$. In this paper, we consider two practically important scenarios, distinguished by how the predictive models and non-conformity scores are trained.

- *Online conformal prediction with pretrained scores.* In this scenario, both the score functions and the predicted models are pretrained on a separate, independent dataset or data stream. As a result, the $s_t(\cdot,\cdot)$'s are independent of the online data stream on which the prediction sets are constructed, while still being allowed to evolve over time.
- *Online conformal prediction with adaptively trained scores.* In this scenario, we allow both the score functions and the predictive models to be trained online, possibly depending on the past observations of the data stream. Therefore, the $s_t(\cdot,\cdot)$'s may be statistically dependent on $\{(X_s, Y_s)\}_{s<t}$.

For both scenarios, an ideal online conformal prediction algorithm would adapt efficiently to the dynamic environment while allowing for tractable computation.

Distribution shift over time. Denote by $\mathcal{D}_t$ (resp. $\mathcal{D}_{1:t}$) the distribution of $Z_t = (X_t, Y_t)$ (resp. $Z_{1:t}$). We allow $\mathcal{D}_t$ to vary over time, which generally violates the exchangeability assumption. In this work, we pay particular attention to the following two distribution drift scenarios.

(i) The change-point setting. This concerns the scenario where the data stream is, in some sense, piecewise stationary. Formally, assume the existence of N^{cp} (*a priori* unknown) change points, denote by

$$1 = \tau_0 < \tau_1 < \cdots < \tau_{N^{\mathsf{cp}}} < \tau_{N^{\mathsf{cp}}+1} = T+1, \tag{7}$$

such that for each $k = 0, \ldots, N^{\mathsf{cp}}$,

$$\begin{cases} s_t(X_t, Y_t) \sim \mathcal{D}^{\mathsf{score}}_{k,\mathsf{seg}}, & \tau_k \le t < \tau_{k+1}, \quad \text{when scores are pretrained,} \\ (X_t, Y_t) \sim \mathcal{D}_{k,\mathsf{seg}}, & \tau_k \le t < \tau_{k+1}, \quad \text{when scores are trained online,} \end{cases} \tag{8}$$

where $\mathcal{D}^{\mathsf{score}}_{k,\mathsf{seg}}$ (resp. $\mathcal{D}_{k,\mathsf{seg}}$) represents the score (resp. data) distribution over the k-th time segment $[\tau_k, \tau_{k+1})$. In words, the distribution of interest remains fixed within each time segment, but may change abruptly at the change points $\tau_1, \ldots, \tau_{N^{\mathsf{cp}}}$. It is assumed that the number and locations of the change points, as well as the associated data distributions, are arbitrary and unknown to the online conformal prediction algorithm.

(ii) The smooth drift setting. In contrast to the above change-point setting that is well suited to modeling infrequent but potentially abrupt distributional jumps, the second setting targets the scenario in which $\mathcal{D}_t$ evolves continuously and smoothly over time. To quantify the overall extent of such distributional variation, we rely on the following two metrics.

- *Cumulative data variation:* this metric measures the aggregate total-variation distance between consecutive data distributions:
$$\mathsf{TV}_T \coloneqq \sum_{t=1}^{T-1} \mathsf{TV}(\mathcal{D}_t, \mathcal{D}_{t+1}). \tag{9}$$
- *Cumulative score variation:* in contrast to TV_T, which is defined based on data distributions, this metric is score-based and tracks the cumulative Kolmogorov-Smirnov distance of consecutive score distributions:
$$\mathsf{KS}_T \coloneqq \sum_{t=1}^{T-1} \mathsf{KS}\big(\mathcal{D}^{\mathsf{score}}_t, \mathcal{D}^{\mathsf{score}}_{t+1}\big), \tag{10}$$
where $\mathcal{D}^{\mathsf{score}}_t$ denotes the distribution of $s_t(X_t, Y_t)$ under data distribution $(X_t, Y_t) \sim \mathcal{D}_t$.

Notably, the score-based metric KS_T can be viewed as a particular instance of the more general cumulative data variation TV_T. In fact, similar quantities have been adopted in prior studies on online learning under data distribution shift (Besbes et al., 2014, 2019; Cheung et al., 2019; Zhao et al., 2020). In this smooth drift setting, our aim is to design online conformal prediction algorithms whose performance can adapt gracefully to such cumulative variations.

2.2 Key metrics: training-conditional coverage and cumulative regret

To assess the performance of an online conformal prediction procedure π, a natural metric is the *training-conditional coverage rate*, defined as

$$\mathsf{cvg}_t = \mathsf{cvg}_t^\pi(Z_{1:t-1}) \coloneqq \mathbb{P}\big(Y_t \in \mathcal{C}_t^\pi(X_t; Z_{1:t-1}) \,\big|\, Z_{1:t-1}\big), \tag{11}$$

where we often write cvg_t for brevity. This metric quantifies the probability of successful coverage conditional on all past data (note that the prediction set $\mathcal{C}_t$ is often constructed based on past observations). Compared with marginal coverage, training-conditional coverage is a stronger notion that ensures most of the test points are covered given the constructed prediction set $\mathcal{C}_t$.

Ideally, one would anticipate cvg_t to match the target level $1-\alpha$. The deviation between the nominal and actual coverage at time t—which may be interpreted as the "regret" incurred at time t—is quantified by the *training-conditional coverage gap* metric defined as

$$\mathsf{cvg\text{-}gap}_t = \mathsf{cvg\text{-}gap}_t^\pi(Z_{1:t-1}) \coloneqq \big|\mathsf{cvg}_t^\pi(Z_{1:t-1}) - (1-\alpha)\big|. \tag{12}$$

The *training-conditional cumulative regret*—hereafter often abbreviated as cumulative regret, or simply regret—of algorithm π is then defined as

$$\mathsf{regret}_T = \mathsf{regret}_\pi(\mathcal{D}_{1:T}, T) \coloneqq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{Z_{1:t-1}\sim\mathcal{D}_{1:t-1}} \big[\mathsf{cvg\text{-}gap}_t^\pi(Z_{1:t-1})\big], \tag{13}$$

which aggregates the training-conditional coverage gaps over time and captures the deviation from the target coverage rate. Here and throughout, we often suppress the explicit dependence on past data and distributions and write $\mathsf{cvg\text{-}gap}_t$ and regret_T when it is clear from the context. Importantly, this cumulative regret notion bridges predictive inference (through coverage guarantees) and online learning (through regret analysis).

Another metric is the long-term coverage rate defined as

$$\mathsf{lt\text{-}cvg}_T = \mathsf{lt\text{-}cvg}_\pi\big(\mathcal{D}_{1:T}, T\big) \coloneqq \frac{1}{T}\sum_{t=1}^{T} \mathbb{P}\big(Y_t \in \mathcal{C}_t^\pi(X_t; Z_{1:t-1})\big), \tag{14}$$

which is often abbreviated by $\mathsf{lt\text{-}cvg}_T$ and can be viewed as an expected version of the empirical long-term coverage frequency in (4). This metric reflects the *time-averaged* coverage probability of a procedure over a horizon T. A large body of prior work (e.g., Gibbs and Candes (2021); Bastani et al. (2022); Angelopoulos et al. (2024a)) studied how far the long-term coverage of a procedure deviates from the target level by looking at the quantity $\mathsf{lt\text{-}cvg}_T - (1-\alpha)$. Note, however, that this quantity captures only the gap between the average coverage probability and the nominal level, rather than the average of the coverage gaps over time (i.e., *gap of average* versus *average of gaps*); as a result, it does not necessarily reflect variations across individual times.

2.3 Why training-conditional cumulative regret?

The training-conditional cumulative regret defined above offers a meaningful criterion for evaluating online conformal prediction algorithms. Unlike long-term coverage metrics like (14), cumulative regret remains informative under distributional drift by aggregating coverage gaps over the entire horizon. The fact below summarizes some basic connections between long-term coverage and regret; the proof can be found in Section A.

Fact 2.1. *The following connections between long-term coverage rate and cumulative regret hold.*

(i) The long-term coverage rate of any online conformal prediction algorithm satisfies

$$\big|\mathsf{lt\text{-}cvg}_T - (1-\alpha)\big| \;\le\; \frac{\mathsf{regret}_T}{T}.$$

(ii) Consider any $0 < \alpha \le 1/2$. There exists an online conformal prediction algorithm such that:

– *For every* $t = 1, \ldots, T$*, its prediction set* $\mathcal{C}_t$ *is either* $\varnothing$ *or* $\mathbb{R}$*, and satisfies* $\mathsf{lt\text{-}cvg}_T = 1 - \alpha$;

– *The regret is lower bounded by* $\mathsf{regret}_T \geq \alpha T$.

On the one hand, Fact 2.1 asserts that sublinear regret (i.e., $\mathsf{regret}_T = o(T)$) guarantees faithful calibration of the long-term coverage rate. On the other hand, Fact 2.1 indicates that the converse fails to hold—as already observed previously (e.g., Bastani et al. (2022); Bhatnagar et al. (2023); Gibbs and Candès (2024))—an algorithm can achieve perfectly calibrated long-term coverage while still incurring training-conditional regret that grows linearly with T. More specifically, long-term coverage does not distinguish an algorithm that consistently achieves coverage close to $1 - \alpha$ from one whose average coverage only coincidentally approaches $1 - \alpha$. For this reason, regret_T serves as a more informative performance measure for online conformal prediction.

3 Online conformal prediction with pretrained scores

In this section, we study online conformal prediction with pretrained score functions, and put forward an algorithm that achieves minimax-optimal regret (up to logarithmic factors) for both the change-point and the smooth drift settings. To be precise, we impose the following assumption throughout this section.

Assumption 3.1 (Pretrained scores)**.** *Suppose the non-conformity score functions* $\{s_t(\cdot,\cdot)\}_{t=1}^T$ *are trained on a separate dataset. Conditional on* $\{s_t(\cdot,\cdot)\}_{t=1}^T$*, the samples* $\{(X_t, Y_t)\}_{t=1}^T$ *are independently generated.*

In words, the data used to pretrain the scores—such as an offline dataset or a different data stream—are separate from, and independently generated of, the data stream for which we construct conformal prediction sets. Consequently, the procedures studied in this section have the flavor of split conformal methods (Vovk et al., 2005). It is also noteworthy that the score functions are allowed to be time-varying.

3.1 Algorithm

Let us motivate our algorithmic ideas and describe the proposed procedure for handling distribution shifts over time. Intuitively, when the data distributions drift significantly while the online conformal prediction algorithm continues to rely on stale quantile estimates, miscoverage can occur frequently, resulting in loss of regret optimality. To remedy this issue, a natural strategy is to continuously monitor the empirical coverage and promptly reset the quantile estimates once they become statistically unreliable. This idea underlies our algorithm design.

3.1.1 Motivating examples

To formalize the above intuition, we begin by examining two simplified cases. A metric that we shall pay particular attention to is the following block coverage error over the time interval $[s, t]$:

$$\mathsf{cvg\text{-}err}_q^\star(s,t) \coloneqq \sum_{l=s}^{t} \Big(\mathbb{P}\big(s_l(X_l, Y_l) \leq q\big) - (1-\alpha) \Big) \tag{15}$$

with q a given threshold; we elucidate how this metric allows us to detect distribution shift below.

A simple case with 1 change point. Before time t, there is a unique change point $t_0 < t$:

- for every $1 \leq l \leq t_0$, the score $s_l(X_l, Y_l)$ is independently drawn from the distribution $\mathcal{P}_1^{\mathsf{seg}}$;
- for every $t_0 < l \leq t$, the score $s_l(X_l, Y_l)$ is independently drawn from the distribution $\mathcal{P}_2^{\mathsf{seg}}$.

The threshold q is taken to be the $(1-\alpha)$-quantile of $\mathcal{P}_1^{\mathsf{seg}}$. Below, we write $s_l = s_l(X_l, Y_l)$ for brevity.

Figure 1 provides a schematic illustration of this simple scenario. The left panel plots $\mathbb{P}(s_t \leq q) - (1-\alpha)$ as t varies. By construction, $\mathbb{P}(s_t \leq q) - (1-\alpha) = 0$ before the change point t_0. At time t_0, the score distribution shifts from $\mathcal{P}_1^{\mathsf{seg}}$ to $\mathcal{P}_2^{\mathsf{seg}}$, causing $\mathbb{P}(s_t \leq q) - (1-\alpha)$ to jump to a nonzero value. This jump reflects the miscalibration induced by applying the pre-change cutoff q to the post-change distribution. The

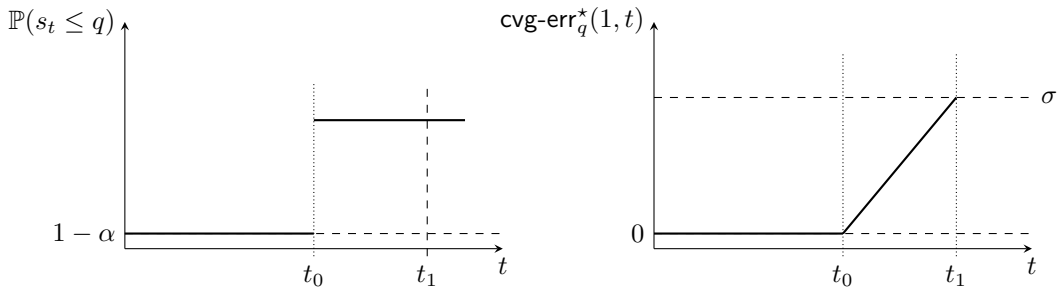

Figure 1: The case with a single change point at t_0. (Left) pointwise coverage $\mathbb{P}(s_t \le q)$ vs. t; (right) block coverage error $\mathsf{cvg\text{-}err}_q^\star(1,t)$ vs. t along with a detection threshold σ.

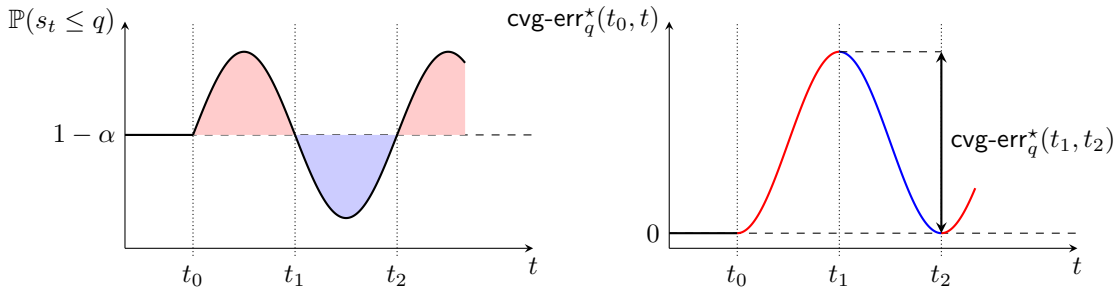

Figure 2: Schematic illustration of a case with smooth, oscillating distribution shifts. (Left) pointwise coverage $\mathbb{P}(s_t \le q)$ vs. t; (right) block coverage error $\mathsf{cvg\text{-}err}_q^\star(1,t)$ vs. t.

right panel plots $\mathsf{cvg\text{-}err}_q^\star(1,t)$ versus t, illustrating the cumulative effect of these pointwise deviations. We have $\mathsf{cvg\text{-}err}_q^\star(1,t) = 0$ prior to t_0, after which the bias accumulates over time and $\mathsf{cvg\text{-}err}_q^\star(1,t)$ grows linearly. If we fix a threshold $\sigma > 0$ and declare a distributional change once $\mathsf{cvg\text{-}err}_q^\star(1,t) > \sigma$, then for σ sufficiently small the detection time t_1 will occur shortly after t_0. This illustrates how a simple block-coverage statistic can enable timely detection of distributional drift.

A case with smooth and oscillating distribution shifts. Consider another simple example, where the score distributions evolve smoothly over time and the instantaneous deviations $\mathbb{P}(s_t \le q) - (1-\alpha)$ oscillate in sign. The variation of $\mathbb{P}(s_t \le q)$ vs. t is displayed in Figure 2(left), where $\mathbb{P}(s_t \le q)$ crosses the reference level $1-\alpha$ multiple times, with the signed deviation being positive on some sub-intervals and negative on others.

Such oscillations cause a cancellation effect. As illustrated in Figure 2(right), the block coverage error $\mathsf{cvg\text{-}err}_q^\star(t_0,t)$ may initially increase but subsequently return to 0 as positive and negative contributions offset one another. Consequently, monitoring deviations from a single starting point t_0 can fail to detect distribution drift. Motivated by this, our proposed solution is to scan over different starting times within a time window and track the maximum deviation. As shown in Figure 2(right), the block $[t_0, t_2]$ exhibits a large deviation $\mathsf{cvg\text{-}err}_q^\star(t_1,t_2)$, even though deviations measured from t_0 cancel out. This maximum-deviation statistic is therefore capable of detecting smooth and oscillating distribution shifts.

3.1.2 The proposed procedure: DriftOCP

We are now positioned to present the proposed online conformal prediction procedure in the presence of pretrained scores, beginning with a distribution drift detection subroutine.

Subroutine: detection of distribution drift (DriftDetect). Thus far, we have illustrated the potential utility of $\mathsf{cvg\text{-}err}_q^\star(s,t)$ in the face of distribution shift. Given that this quantity is not accessible in

Algorithm 1: DRIFTDETECT$(q; t_0, t_1; \sigma)$

input: quantile q; time window $[t_0, t_1] \subseteq [T]$; detection threshold σ.
for $j = t_0, \cdots, t_1$ **do** // scan the entire time window.
  compute $Z_{j,t_1} \leftarrow \dfrac{\left|\mathsf{cvg\text{-}err}_q(j, t_1)\right|}{\sqrt{t_1 - j + 1}}$ (cf. (16)). // construct detection statistics.
  if $Z_{j,t_1} > \sigma$ **then**
    return true. // declare detection of drift once this statistic exceeds the threshold.
return false // no distribution drift has been detected.

practice, we propose to approximate it via the following empirical block coverage error:

$$\mathsf{cvg\text{-}err}_q(s,t) \coloneqq \sum_{l=s}^{t} \big(\mathbb{1}\{s_l(X_l, Y_l) \le q\} - (1-\alpha)\big). \tag{16}$$

Assuming statistical independence, the central limit theorem implies that $\mathsf{cvg\text{-}err}_q(s,t)$ fluctuates around $\mathsf{cvg\text{-}err}^\star_q(s,t)$ with uncertainty on the order of $(t-s+1)^{1/2}$. Moreover, under stationarity of the scores within $[s,t]$, one has $\mathsf{cvg\text{-}err}^\star_q(s,t) = 0$. Consequently, testing whether the normalized empirical fluctuation $\left|\mathsf{cvg\text{-}err}_q(s,t)\right| / \sqrt{t-s+1}$ exceeds a suitably chosen threshold σ provides a natural criterion for detecting distribution shifts within the time interval $[s,t]$. The intuition is formalized in the subroutine described in Algorithm 1, denoted by DRIFTDETECT$(q; t_0, t_1; \sigma)$, which scans the window $[t_0, t_1]$ for statistically significant departure from stationarity. The subroutine plays a pivotal role in our main procedure.

Full procedure. We now describe several key components of our main procedure. The complete procedure, called DRIFTOCP (short for *online conformal prediction with drift detection*), is summarized in Algorithm 2.

- *Stage-wise decomposition.* The entire time horizon is divided into a sequence of *stages* in a data-driven manner, where we use n to index stages. A new stage is initiated whenever the subroutine DRIFTDETECT signals a substantial distribution drift. Within each stage, the score distributions are treated as *approximately stationary*.
- *Decomposition into rounds within each stage.* Provided no distribution drift is detected, each stage is further partitioned into a sequence of *rounds*. Following the standard doubling trick (e.g., Cesa-Bianchi and Lugosi (2006, Chapter 2.3), Lattimore and Szepesvári (2020, Chapter 6)), we let the round lengths grow geometrically, which eliminates the need for prior knowledge of the horizon length. We use r to index rounds. For round r of stage n, all data from the preceding round are used to update the quantile estimate $q_{n,r}$, which in turn determines the prediction set at any time τ within the current round:
$$\mathcal{C}_\tau = \{y : s_\tau(X_\tau, y) \le q_{n,r}\}.$$
- *Drift detection within each round.* Let $\tau_{n,r}$ represent the time at which round r of stage n begins. During this round, incoming samples are monitored via DRIFTDETECT$\big(q_{n,r};\, \tau_{n,r},\, \tau_{n,r}+t;\, \sigma_{n,r}\big)$, so that each call to the subroutine DRIFTDETECT operates on a block beginning at the onset of the current round.

We would also like to highlight several appealing features of Algorithm 2. First, it is horizon-free, meaning that the procedure does not require any prior knowledge of the horizon length T; as we shall see later, our algorithm achieves the desirable *anytime* regret—a terminology commonly adopted in the online learning literature (Lattimore and Szepesvári, 2020) to emphasize its horizon-free nature. Second, it is computationally lightweight. Each new observation triggers at most one drift detection subroutine over the current window and each round only updates the quantile estimate once. In particular, the computational cost at each time t scales linearly with the length of the current scanning window, instead of recalibrating over many candidate lookback windows as in some prior work (see Section 3.4). Moreover, the drift detection subroutine is inexpensive in practice, since the underlying detection statistics can be maintained incrementally. Finally, we emphasize that Algorithm 2 operates without any prior knowledge of the underlying distributional drift—such as the number and locations of change points, or the degree of cumulative variation—highlighting its adaptation to unknown and evolving data-generating mechanisms.

Algorithm 2: Online Conformal Prediction with Drift Detection (DriftOCP)

input: target coverage level $1-\alpha$; detection thresholds $\{\sigma_{n,r}\}_{n,r=1}^{\infty}$.
initialize: $n \leftarrow 1$, $r \leftarrow 1$, $\tau \leftarrow 0$, $\tau_{1,1} \leftarrow 1$, $q_{1,1} \leftarrow 0$.
while *true* **do**
  for $t = 1, \cdots, T_r\ (:= 3^r)$ **do** // round r contains at most $T_r = 3^r$ time points.
    $\tau \leftarrow \tau + 1$. // update global time index.
    observe feature X_τ and score function $s_\tau(\cdot,\cdot)$; set $X_{n,r,t} \leftarrow X_\tau$, $s_{n,r,t}(\cdot,\cdot) \leftarrow s_\tau(\cdot,\cdot)$.
    construct prediction set $\mathcal{C}_\tau \leftarrow \{y : s_\tau(X_\tau, y) \le q_\tau\}$ with $q_\tau = q_{n,r}$.
    observe Y_τ; set $Y_{n,r,t} \leftarrow Y_\tau$. // response is observed after the prediction set is formed.
    drift $\leftarrow$ DriftDetect$(q_{n,r};\ \tau_{n,r},\ \tau_{n,r} + t - 1;\ \sigma_{n,r})$. // call Algorithm 1.
    if *drift is true* **then**
      $\widetilde{q} \leftarrow q_{n,r}$, $n \leftarrow n+1$, $r \leftarrow 1$. // update stage index and round index.
      $q_{n,1} \leftarrow \widetilde{q}$, $\tau_{n,1} \leftarrow \tau + 1$. // initialization for new stage.
      break. // enter next stage.
  if *drift is false* **then**
    $q_{n,r+1} \leftarrow \mathrm{argmin}_q \left| \sum_{j=1}^{T_r} \left(\mathbb{1}\{s_{n,r,j}(X_{n,r,j}, Y_{n,r,j}) > q\} - \alpha \right) \right|$. // update quantile estimate.
    $r \leftarrow r+1$, $\tau_{n,r} \leftarrow \tau + 1$. // start time of next round.

3.2 Theoretical guarantees

Next, we establish non-asymptotic upper bounds on the training-conditional regret for the proposed Algorithm 2, encompassing both the change-point and smooth drift settings introduced in Section 2.1.

Theorem 3.1. *Suppose that Assumption 3.1 holds. If we set the detection thresholds as* $\sigma_{n,r} := 24\sqrt{\log(4\tau_{n,r})}$ *for every stage-round index pair* (n, r)*, then Algorithm 2 achieves*

$$\mathsf{regret}_T \le \begin{cases} \widetilde{O}(\sqrt{(N^{\mathsf{cp}}+1)T}) & \text{for the change-point setting;} \\ \widetilde{O}\left(\sqrt{T} + (\mathsf{KS}_T)^{\frac{1}{3}} T^{\frac{2}{3}}\right) & \text{for the smooth drift setting.} \end{cases} \tag{17}$$

The proof of this theorem is provided in Section B.1. For the change-point setting, Theorem 3.1 reveals that the regret scales proportionally to the square root of the number of change points N^{cp}, in addition to the $\sqrt{T}$ dependence on the time horizon—a scaling in T that arises commonly in online learning (Shalev-Shwartz, 2012; Lattimore and Szepesvári, 2020). In contrast, for the smooth drift setting, our regret bound contains a term $(\mathsf{KS}_T)^{1/3}T^{2/3}$, whose dependence on T is worse than the $\sqrt{T}$ scaling. This suggests that the dominant source of regret may stem from the temporal evolution of the underlying score distribution, underscoring the important role of real-time adaptation. We also emphasize that the coverage gap depends on the KS distance between the *score* distributions rather than those of the *raw data*. As discussed in Barber et al. (2023), this distinction can lead to much tighter guarantees, since the scores may be far closer in distribution than the underlying data. Encouragingly, these regret bounds match the minimax lower bound (up to logarithmic factors), as we shall demonstrate next.

3.3 Minimax lower bound

To examine the optimality of Algorithm 2, this subsection develops minimax lower bounds on the cumulative regret, tailored to the class of online algorithms with pretrained scores. We begin by specifying the admissible algorithms and distribution classes of interest, which are necessary for the lower-bound analysis.

- *Admissible algorithms.* For any online conformal prediction algorithm π with pretrained scores, let π_t denote its rule for selecting the quantile threshold q_t (cf. (2)) at time t. Let $U \sim \mathsf{Unif}(0,1)$ be an auxiliary random seed, independent of the data stream. We consider a family $\mathcal{Q}$ of non-anticipating algorithms $\pi = \{\pi_t\}_{t\ge 1}$, where $\pi_1 : [0,1] \to \mathbb{R}$ and $\pi_t : \mathbb{R}^{t-1} \times [0,1] \to \mathbb{R}$ $(t \ge 2)$ are measurable

mappings. For each t, π specifies the quantile threshold

$$q_t = \begin{cases} \pi_1(U), & \text{if } t = 1, \\ \pi_t(s_{t-1}, \ldots, s_1, U), & \text{if } t \geq 2, \end{cases}$$

where we remind the reader that $s_t = s_t(X_t, Y_t)$. Each π_t depends only on the past scores and the random seed U, hence algorithms in $\mathcal{Q}$ are score-based, non-anticipatory, and possibly randomized.

- *Distribution classes.* We introduce two *score-based* distribution classes, corresponding to the two settings in Section 2.1: for given budgets $N^{\mathsf{cp}} \in \mathbb{Z}^+$ and $\mathsf{KS}_T > 0$, define

$$\mathcal{L}_1(N^{\mathsf{cp}}) := \Big\{(\mathcal{D}_1, \ldots, \mathcal{D}_T) : \ (\mathcal{D}_1^{\mathsf{score}}, \ldots, \mathcal{D}_T^{\mathsf{score}}) \text{ change at most } N^{\mathsf{cp}} \text{ times.}\Big\}; \tag{18a}$$

$$\mathcal{L}_2(\mathsf{KS}_T) := \Big\{(\mathcal{D}_1, \ldots, \mathcal{D}_T) : \ \sum_{t=1}^{T-1} \mathsf{KS}(\mathcal{D}_t^{\mathsf{score}}, \mathcal{D}_{t+1}^{\mathsf{score}}) \leq \mathsf{KS}_T\Big\}. \tag{18b}$$

For a distribution class $\mathcal{L}$, the worst-case regret of algorithm $\pi \in \mathcal{Q}$ is defined as

$$\mathsf{regret}_\pi(\mathcal{L}, T) := \sup_{(\mathcal{D}_1, \ldots, \mathcal{D}_T) \in \mathcal{L}} \mathsf{regret}_\pi(\mathcal{D}_{1:T}, T). \tag{19}$$

Armed with these definitions and notation, we are ready to present our minimax lower bounds.

Theorem 3.2. *Consider any fixed $\alpha \in (0, 1)$. Suppose that Assumption 3.1 holds. For any admissible algorithm $\pi \in \mathcal{Q}$, its worst-case regret (cf. (19)) satisfies*

$$\mathsf{regret}_\pi\big(\mathcal{L}_1(N^{\mathsf{cp}}), T\big) = \Omega\Big(\sqrt{(N^{\mathsf{cp}} + 1)T}\Big);$$
$$\mathsf{regret}_\pi\big(\mathcal{L}_2(\mathsf{KS}_T), T\big) = \Omega\Big(\sqrt{T} + (\mathsf{KS}_T)^{1/3} T^{2/3}\Big).$$

Evidently, the minimax regret lower bounds in Theorem 3.2 match the achievable regret of Algorithm 2 in Theorem 3.1 (modulo some logarithmic factors), thereby confirming the regret optimality of our proposed procedure in a minimax sense. The proof is postponed to Section B.2.

3.4 Comparisons with prior art

Gibbs and Candes (2021) introduced ACI with a time-invariant stepsize schedule, and established guarantees in terms of the time-averaged long-run coverage frequency (cf. (4)), which hold irrespective of the data generating mechanism but do not imply valid coverage at individual time points. Stronger (asymptotic) guarantees were also established under stationary hidden Markov models. Building on this work, Angelopoulos et al. (2024a) studied ACI with decaying stepsizes and proved asymptotically exact (pointwise) coverage under i.i.d. data; these results, however, are asymptotic in nature and do not readily extend to settings with distribution drift. The analysis for the empirical long-term coverage frequency has further motivated the studies of a new perspective in online learning called "gradient equilibrium," which yields a useful framework for several other statistical applications (Angelopoulos et al., 2025). Relatedly, Bastani et al. (2022) generalized the notion of long-term coverage frequency by proposing an approach that achieves *multi-valid coverage* guarantees even in adversarial settings. Their guarantees, however, are stated in terms of empirical frequencies along the realized sequence and do not yield training-(and-calibration)-conditional coverage guarantees.

Several recent works Pournaderi and Xiang (2024); Humbert et al. (2025) began to investigate training-conditional guarantees for online conformal prediction. Nevertheless, these results either do not account for distribution shift or rely on fairly strong assumptions (e.g., a uniform upper bound on the pre-/post-drift density ratio). Assuming independently trained scores (as in Assumption 3.1), Han et al. (2024a) derived training-conditional guarantees for online conformal prediction under distribution drift via adaptive lookback-window selection for quantile calibration. Their results focus on *last-step* (terminal-time) validity, whereas we study training-conditional *cumulative* regret. Our method is also computationally more efficient: at time t,

their procedure requires t quantile estimates and t^2 empirical CDF evaluations, while Algorithm 2 needs at most one quantile estimate and t empirical coverage computations. The same authors also studied model assessment and selection under distribution drift (Han et al., 2024b).

When it comes to lower bound analysis, Areces et al. (2024); Duchi (2025) discussed minimax lower bounds for the training-conditional coverage error in the presence of independently trained score functions. Compared to their results, Theorem 3.2 moves beyond coverage guarantees under worst-case covariate shift and explicitly accounts for the effect of distribution drift over time.

4 Online conformal prediction with adaptively trained scores

We now turn our attention to the scenario in which the non-conformity scores and the predictive models are allowed to be trained online based on past observations. More precisely, we make the following assumptions throughout this section.

Assumption 4.1 (Online-trained scores)**.** *Suppose that the non-conformity scores are constructed online. At each time t, the score functions may depend on the past data $\{(X_t, Y_t)\}_{s<t}$, but not on any data observed at or after time t. The data $\{(X_t, Y_t)\}_{t=1}^T$ are independently generated.*

Given the flexibility to adaptively update the score functions, we adopt the full conformal paradigm (Vovk et al., 2005), which leverages all available data for both score construction and quantile estimation, without resorting to data splitting. While this full conformal approach enables more efficient use of the data, it also introduces intricate statistical dependence across time, making it challenging to detect distributional drift and to establish training-conditional coverage. We develop several technical innovations to address these challenges.

4.1 Algorithm

We first review the standard full conformal prediction method, and then describe how it can be adapted to streaming data with distribution drift.

Review: (batch) full conformal prediction. Imagine we are given two datasets,

$$\mathcal{Z}^{\mathsf{train}} := \left\{(X_i^{\mathsf{train}}, Y_i^{\mathsf{train}})\right\}_{i=1}^n \qquad \text{and} \qquad \mathcal{Z}^{\mathsf{cal}} := \left\{(X_i^{\mathsf{cal}}, Y_i^{\mathsf{cal}})\right\}_{i=1}^m,$$

which may overlap. While it is common to take $\mathcal{Z}^{\mathsf{train}} = \mathcal{Z}^{\mathsf{cal}}$ to maximize data efficiency, we allow $\mathcal{Z}^{\mathsf{train}}$ and $\mathcal{Z}^{\mathsf{cal}}$ to differ, a flexibility that will be useful for subsequent algorithmic development. The test point contains the feature X^{test}.

The training dataset $\mathcal{Z}^{\mathsf{train}}$ is used to train predictive models via a learning algorithm $\mathcal{A}$, yielding

$$\widehat{\mu}^{(X^{\mathsf{test}},y)}(\cdot) := \mathcal{A}\left(\mathcal{Z}^{\mathsf{train}}; (X^{\mathsf{test}}, y)\right) \tag{20}$$

for every candidate response $y \in \mathbb{R}$, whereas the calibration dataset $\mathcal{Z}^{\mathsf{cal}}$ is used to construct the prediction set with the aid of the fitted models $\widehat{\mu}$. Importantly, the fitted model $\widehat{\mu}^{(X^{\mathsf{test}},y)}$ depends on the hypothesized response y (cf. (20)), and may need to be refitted for each y under consideration. For each (X^{test}, y), we define a set of non-conformity scores (or residual scores) as:

$$s_i^{(X^{\mathsf{test}},y)} := \left|Y_i^{\mathsf{cal}} - \widehat{\mu}^{(X^{\mathsf{test}},y)}(X_i^{\mathsf{cal}})\right|, \qquad i = 1, \ldots, m, \tag{21a}$$

$$s_{\mathsf{test}}^{(X^{\mathsf{test}},y)} := \left|y - \widehat{\mu}^{(X^{\mathsf{test}},y)}(X^{\mathsf{test}})\right|. \tag{21b}$$

The full conformal prediction set is then taken to be:

$$\mathcal{C}(X^{\mathsf{test}}) := \left\{ y : s_{\mathsf{test}}^{(X^{\mathsf{test}},y)} \le \mathsf{Quantile}_{1-\alpha}\left(\frac{1}{m+1}\left[\delta\left(s_{\mathsf{test}}^{(X^{\mathsf{test}},y)}\right) + \sum_{i=1}^m \delta\left(s_i^{(X^{\mathsf{test}},y)}\right)\right]\right)\right\}, \tag{22}$$

where $\delta(a)$ denotes a point mass (i.e., the Dirac measure) at a, and $\mathsf{Quantile}_{1-\alpha}(P)$ denotes the $(1-\alpha)$-quantile of distribution P. In words, this prediction set contains all candidate values whose residual scores do not exceed the $(1-\alpha)$-quantile of the empirical distribution formed by the m calibration residuals together with the candidate's own residual. When $\mathcal{Z}^{\mathsf{train}} = \mathcal{Z}^{\mathsf{cal}}$, under exchangeability of the data and permutation symmetry of the model fitting algorithm $\mathcal{A}$, classical results (e.g., Vovk et al. (2005); Lei et al. (2018); Barber et al. (2023); Liang and Barber (2025)) guarantee finite-sample validity of this full conformal procedure.

Our algorithm: online full conformal prediction with drift detection (DriftOCP-full). When distributional drift occurs over time, the assumption of exchangeability breaks down, invalidating the coverage guarantees of the full conformal algorithm described above. Building on the key algorithmic ideas introduced in Section 3.1, we extend full conformal methods to online settings with temporal distribution drift.

We refer to the proposed algorithm as DriftOCP-full (short for *online full conformal prediction with drift detection*), and present the full procedure in Algorithm 4. We first isolate several key features of DriftOCP-full that parallel those of DriftOCP.

- *Drift detection subroutine DriftDetect+.* We continue to employ a drift detection subroutine to identify the occurrence of a distribution drift. We introduce a slightly extended version of Algorithm 1, formalized in Algorithm 3 and referred to as DriftDetect+. In essence, DriftDetect+ differs from DriftDetect only in that it replaces $\mathsf{cvg\text{-}err}_q(s,t)$—defined in (16) based on quantiles of the non-conformity score—with the more general definition of empirical block coverage error

$$\mathsf{cvg\text{-}err}_{\mathcal{C}}(s,t) := \left| \sum_{l=s}^{t} \left(\mathbb{1}\{Y_l \in \mathcal{C}(X_l)\} - (1-\alpha) \right) \right|. \tag{23}$$

- *Decomposition into stages and rounds.* Akin to Algorithm 2, we partition the entire time horizon into stages—with the aid of the subroutine DriftDetect+ in Algorithm 3 in a data-driven manner—and further decompose each stage into rounds. Within each stage, the data distributions are treated as approximately stationary. As before, we use n and r to index stages and rounds, respectively. We will repeatedly use the following notation:
  - $T_r = 3^r$: the number of time points in round r of each stage, chosen to grow geometrically in r so as to avoid requiring prior knowledge of the horizon length T.
  - r_n: the last round of stage n. We adopt the convention that $(n,0) = (n-1, r_{n-1})$.
  - $\tau_{n,r}$: the time index—measured in the original horizon $\{1,\dots,T\}$—corresponding to the first time point of round r in stage n. We adopt the convention that $\tau_{n,r_n+1} := \tau_{n+1,1}$ and $\tau_{n,0} := \tau_{n-1,r_{n-1}}$.
  - $X_{n,r,t}$ and $Y_{n,r,t}$: the feature and the response arriving at the t-th time point of round r in stage n.

Next, we highlight several full conformal components of DriftOCP-full that extend batch full conformal methods to online settings.

- *Training and calibration sets for round r of stage n.* When constructing the prediction set at any time within round r of stage n, we choose the training and calibration sets as

$$\mathcal{Z}^{\mathsf{train}}_{n,r} := \underbrace{\{(X_i,Y_i)\}_{i=1}^{\tau_{n,r}-1}}_{\text{all data before current round}} \quad \text{and} \quad \mathcal{Z}^{\mathsf{cal}}_{n,r} := \underbrace{\{(X_i,Y_i)\}_{i=\tau_{n,r-1}}^{\tau_{n,r}-1}}_{\text{all data in preceding round}}. \tag{24}$$

  In words, the training set $\mathcal{Z}^{\mathsf{train}}_{n,r}$ comprises all samples observed prior to the current round, while the calibration set $\mathcal{Z}^{\mathsf{cal}}_{n,r}$ consists of all samples collected during the immediately preceding round (i.e., round $r-1$ of stage n). Intuitively, the data in preceding round are treated as stationary in distribution and are therefore well suited for calibration, whereas all earlier data—regardless of whether distribution shifts have occurred—can be leveraged for model training.

- *Fitted models, scores, and prediction sets.* The construction of the prediction set follows the standard full conformal method described in (20)-(22). Consider any time point within round r of stage n. For a

Algorithm 3: DRIFTDETECT+$(\mathcal{C}; t_0, t_1; \sigma)$

input: set-valued function $\mathcal{C}(\cdot)$; time window $[t_0, t_1] \subseteq [T]$; detection threshold σ.
for $j = t_0, \cdots, t_1$ **do** // scan the entire time window.
  compute $Z_{j,t_1} \leftarrow \dfrac{\left|\mathsf{cvg\text{-}err}_{\mathcal{C}}(j, t_1)\right|}{\sqrt{t_1 - j + 1}}$ (cf. (23)). // construct detection statistics.
  if $Z_{j,t_1} > \sigma$ **then**
    return true. // declare detection of drift once this statistic exceeds the threshold.
return false // no distribution drift has been detected.

Algorithm 4: ONLINE FULL CONFORMAL PREDICTION WITH DRIFT DETECTION (DRIFTOCP-FULL)

input: target coverage level $1 - \alpha$; detection thresholds $\{\sigma_{n,r}\}_{n,r=1}^{\infty}$.
initialize: $n \leftarrow 1$, $r \leftarrow 1$, $\tau \leftarrow 0$, $\tau_{1,1} \leftarrow 1$, $\mathcal{C}_{1,1}(x) \leftarrow \mathbb{R}$; $\forall x \in \mathcal{X}$.
while *true* **do**
  construct set-valued function $\mathcal{C}_{n,r}(\cdot)$ by (27). // prediction-set forming strategy for this round.
  for $t = 1, \cdots, T_r$ $(:= 3^r)$ **do** // round r contains at most $T_r = 3^r$ time points.
    $\tau \leftarrow \tau + 1$. // update global time index.
    observe feature X_τ; set $X_{n,r,t} \leftarrow X_\tau$.
    construct prediction set $\mathcal{C}_\tau \leftarrow \mathcal{C}_{n,r}(X_\tau)$; take $\mathcal{C}_{n,r,t} \leftarrow \mathcal{C}_\tau$. // form prediction set
    observe Y_τ; set $Y_{n,r,t} \leftarrow Y_\tau$. // response is observed after the prediction set is formed.
    drift $\leftarrow$ DRIFTDETECT+$(\mathcal{C}_{n,r}; \tau_{n,r}, \tau_{n,r} + t - 1; \sigma_{n,r})$. // call Algorithm 3.
    if *drift is true* **then**
      $n \leftarrow n + 1$, $r \leftarrow 1$. // update stage index and round index.
      break. // enter next stage.
  if *drift is false* **then**
    $r \leftarrow r + 1$, $\tau_{n,r} \leftarrow \tau + 1$. // start time of next round.

feature X observed in this round and an imputed response y, we invoke a learning algorithm $\mathcal{A}$ to fit a predictive model

$$\widehat{\mu}^{(X,y)}(\cdot) := \mathcal{A}\left(\mathcal{Z}_{n,r}^{\mathsf{train}}; (X, y)\right). \tag{25}$$

The non-conformity (or residual) scores are then computed for all T_{r-1} points observed in the immediately preceding round (i.e., $\mathcal{Z}_{n,r}^{\mathsf{cal}}$) as well as the hypothesized test point (X, y), yielding

$$s_i^{(X,y)} := \left|Y_{n,r-1,i}^{\mathsf{cal}} - \widehat{\mu}^{(X,y)}(X_{n,r-1,i}^{\mathsf{cal}})\right|, \qquad i = 1, \ldots, T_{r-1}, \tag{26a}$$

$$s_{\mathsf{test}}^{(X,y)} := \left|y - \widehat{\mu}^{(X,y)}(X)\right|. \tag{26b}$$

Given these scores, we form the prediction set based on feature X as

$$\mathcal{C}_{n,r}(X) := \left\{ y : s_{\mathsf{test}}^{(X,y)} \leq \mathsf{Quantile}_{1-\alpha}\left(\frac{1}{T_{r-1}+1} \left[\delta\left(s_{\mathsf{test}}^{(X,y)}\right) + \sum_{i=1}^{T_{r-1}} \delta\left(s_i^{(X,y)}\right) \right] \right) \right\}, \tag{27}$$

which collects all candidate responses y for which the test residual $s_{\mathsf{test}}^{(X,y)}$ does not exceed the target quantile of the combined calibration and test scores, in direct analogy to the standard full conformal framework. The prediction-set construction strategy $\mathcal{C}_{n,r}(\cdot)$ remains fixed throughout the current round, since the same training and calibration sets are used for all time points within this round.

4.2 Theoretical guarantees under stability assumptions

We now turn to the regret performance of the proposed Algorithm 4. A dominant fraction of prior full conformal theory relies on a permutation symmetry assumption of the model fitting algorithm, namely, that the fitted predictor $\widehat{\mu}$ remains invariant under arbitrary reordering of the training samples. However, many online learning algorithms, such as online gradient descent with time-varying learning rates, do not produce predictors that are exactly permutation invariant. Enforcing permutation symmetry in these cases would oftentimes require, at each time step, retraining the model from scratch on all previously observed data, thereby incurring a substantial computational burden. To better accommodate online model fitting algorithms, we instead rely on two different assumptions—one concerning the Lipschitz continuity of the conditional response distribution, and the other pertaining to the stability of the learning algorithm—replacing the permutation symmetry requirement.

Assumption 4.2 (Lipschitz continuity of conditional response distribution)**.** *There exists a quantity $L_1 > 0$ such that, for every time $t \geq 1$ and every $x_t \in \mathcal{X}$, the function $g_{x_t}(z) \coloneqq \mathbb{P}(Y_t \leq z \mid X_t = x_t)$ is L_1-Lipschitz continuous w.r.t. z.*

Assumption 4.3 (Stability of learning algorithm)**.** *Let $\mathcal{Z} = \{z_1, \cdots, z_m\}$ be a training set of size m, and let $\widehat{\mu}(\cdot \,|\, \mathcal{Z})$ represent the predictive model returned by algorithm $\mathcal{A}$ when trained on $\mathcal{Z}$. We assume that $\widehat{\mu}(\cdot \,|\, \cdot)$ is a measurable function. For any $i \in [m]$ and any replacement sample w, define $\mathcal{Z}_{i,w} = \{z_1, \cdots, z_{i-1}, w, z_{i+1}, \cdots, z_m\}$, which differs from $\mathcal{Z}$ only in its i-th element. We assume that there exists a constant $L_2 > 0$ such that, for an arbitrary m, one has*

$$\left|\widehat{\mu}\big(x \mid \mathcal{Z}\big) - \widehat{\mu}\big(x \mid \mathcal{Z}_{i,w}\big)\right| \leq \frac{L_2}{m} \qquad \text{for all } x, w, \mathcal{Z}, \text{ and } i \in [m]. \tag{28}$$

Assumptions 4.2 and 4.3 are commonly used in full conformal prediction literature (e.g., Barber et al. (2021); Ndiaye (2022); Steinberger and Leeb (2023); Liang and Barber (2025); Lee and Zhang (2025)). In fact, Assumption 4.2 is fairly standard in statistical modeling; a common example concerns the setting $Y_t = m_t(X_t) + \varepsilon_t$, where ε_t is generated independently of X_t and admits a density uniformly bounded above by L_1. In addition, Assumption 4.3 formalizes a sort of stability requirement of $\widehat{\mu}$: perturbing a single training example alters the predictive output by at most $O(1/m)$ (assuming a constant L_2). To help illustrate the practical relevance of Assumption 4.3, we single out a few canonical parametric examples that can be readily analyzed within our framework:

- *constrained M-estimation*: see Section D.1;
- *linear stochastic approximation*: see Section D.2;
- *stochastic strongly convex optimization*: see Section D.3.

The interested reader is referred to Section D for detailed verification of Assumption 4.3 in these examples.

Armed with the above assumptions, we establish regret upper bounds for the proposed online full conformal algorithm.

Theorem 4.1. *Suppose that Assumption 4.1 holds, and that Assumptions 4.2 and 4.3 hold with quantities L_1 and L_2, respectively. Let $L = L_1 L_2$. If we set the drift detection thresholds as $\sigma_{n,r} \coloneqq 10 \log^3(40\tau_{n,r})$ for every stage-round index pair (n, r), then Algorithm 4 achieves*

$$\mathsf{regret}_T \leq \begin{cases} \widetilde{O}\big(\sqrt{(N^{\mathsf{cp}} + L + 1)T}\,\big) & \text{for the change-point setting;} \\ \widetilde{O}\big(\sqrt{(L+1)T} + (\mathsf{TV}_T)^{\frac{1}{3}} T^{\frac{2}{3}}\big) & \text{for the smooth drift setting.} \end{cases} \tag{29}$$

Despite the adaptive, online training of the non-conformity score functions, the training-conditional regret attained by our online full conformal prediction algorithm takes a form similar to that achieved with pretrained scores, provided that $L = O(1)$. A main difference is that the score-based Kolmogorov–Smirnov distance appearing in the pretrained-score scenario (see Theorem 3.1) is replaced here by the total-variation distance w.r.t. data distributions, since the scores are now trained based on the observed data. Our result is fully non-asymptotic, which stands in stark contrast to several prior works (e.g., Angelopoulos et al. (2024a)) that focused on asymptotic coverage guarantees (i.e., $T \to \infty$ with other parameters held fixed).

Byproduct: training-conditional coverage for batch full conformal methods. En route to establishing the regret upper bound of DRIFTOCP-FULL, we need to address the challenge of achieving training-conditional coverage when scores are trained in-sample using a possibly non-symmetric learning algorithm. Our analysis leads to new training-conditional coverage results for batch full conformal methods, which is stated below and may be of independent interest. The proof is deferred to Section C.1.

Proposition 4.1. *Consider any integers $n \geq m$. Let $Z_{1:m}^{\mathsf{cal}}$ be a calibration dataset and $Z_{1:n}^{\mathsf{train}}$ a dataset used for model fitting. We assume that the calibration dataset is a subset of the training dataset, and in particular, $Z_{1:m}^{\mathsf{cal}} = Z_{1:m}^{\mathsf{train}}$. The samples in $\{Z_{1:m}^{\mathsf{cal}}\} \cup \{Z_{m+1:n}^{\mathsf{train}}\}$ are independently generated. Construct the full conformal prediction set $\mathcal{C}(\cdot) = \mathcal{C}(\cdot \mid Z_{1:m}^{\mathsf{cal}}; Z_{1:n}^{\mathsf{train}})$ as in Eqn. (22). Consider a target pair $Z = (X, Y) \sim \mathcal{D}$. Suppose the distribution $\mathcal{D}$ and the fitted model satisfy Assumptions 4.2 and 4.3 with coefficients L_1 and L_2, respectively, and denote $L_1 L_2$ as L. Then for any $\delta \in (0,1)$, conditional on any realization $Z_{m+1:n}^{\mathsf{train}} = z_{m+1:n}^{\mathsf{train}}$, we have*

$$\left|\mathbb{P}_{\mathcal{D}}\left(Y \in \mathcal{C}(X) \,\middle|\, Z_{1:m}^{\mathsf{cal}}\right) - (1-\alpha)\right| \leq \frac{52L\sqrt{m\log(45n/\delta)}}{n} + 25\sqrt{\frac{\log(40/\delta)}{m}} + \frac{2}{m}\sum_{l=1}^{m} \mathsf{TV}(Z, Z_l^{\mathsf{cal}}) \tag{30}$$

with probability exceeding $1-\delta$ (with respect to the randomness only in $Z_{1:m}^{\mathsf{cal}}$).

Remark 4.1. *In particular, in the most common case where the training and calibration sets coincide (so that $m = n$ and $Z_{1:n}^{\mathsf{cal}} = Z_{1:n}^{\mathsf{train}}$), this result asserts that the standard full conformal method (cf. (22)) achieves*

$$\left|\mathbb{P}_{\mathcal{D}}\left(Y \in \mathcal{C}(X) \,\middle|\, Z_{1:n}^{\mathsf{train}}\right) - (1-\alpha)\right| \lesssim \max\{L, 1\}\sqrt{\frac{\log(n/\delta)}{n}} + \frac{1}{n}\sum_{l=1}^{n} \mathsf{TV}(Z, Z_l^{\mathsf{cal}}) \tag{31}$$

with probability greater than $1-\delta$.

Proposition 4.1 establishes a training–conditional concentration bound for full conformal residuals that holds for a fixed batch of data and a stable learner (no online structure is used). This result captures the effect of using a data-dependent predictor inside the full conformal construction, and will play a crucial role in establishing our training-conditional regret bound (when combined with the stage/round decomposition and drift–detection analysis). In addition, Proposition 4.1 generalizes existing results on training-conditional coverage for full-conformal-type approach; more detailed comparisons with prior results are provided in Section 4.4.

4.3 Minimax lower bound

We now complement Theorem 4.1 with a lower bound, which serves to better evaluate the optimality of our proposed procedure. Before proceeding, it is important to note that, while Theorem 3.2 already establishes a regret lower bound, that result hinges upon a specific way of constructing the prediction set—namely, one based on quantile estimation of pretrained non-conformity scores. In practice, however, a broader class of methods is available, including the online full conformal approach, which can induce substantially more complex and structurally different prediction sets. As a result, Theorem 3.2 does not provide an appropriate lower bound for the settings considered in this section. We develop a new lower bound for this broader class of algorithms below.

Lower bound. We start by specifying the scope of the problem.

- *Admissible algorithms.* Since the prediction-set construction considered in this section no longer relies on a given set of non-conformity score functions, the first step is to redefine the class of admissible algorithms accordingly. Denote by $\mathsf{Map}(\mathcal{X}, \mathcal{B}(\mathbb{R}))$ the set of mappings from $\mathcal{X}$ to $\mathcal{B}(\mathbb{R})$ (i.e., this forms the set of prediction-set construction functions). Let $U \sim \mathrm{Unif}(0,1)$ be a random variable independent of the data stream. Let $\pi_1 : [0,1] \to \mathsf{Map}(\mathcal{X}, \mathcal{B}(\mathbb{R}))$ and, for $t \geq 2$, let $\pi_t : (\mathcal{X}, \mathbb{R})^{t-1} \times [0,1] \to \mathsf{Map}(\mathcal{X}, \mathcal{B}(\mathbb{R}))$. Given a sequence of data $\{Z_i\}_{i=1}^{t-1} = \{(X_i, Y_i)\}_{i=1}^{t-1}$ prior to time t, we define $\mathcal{C}_t \in \mathsf{Map}(\mathcal{X}, \mathcal{B}(\mathbb{R}))$ to be the set-valued mapping induced by algorithm π_t at time t, namely,

$$\mathcal{C}_t(\cdot) = \begin{cases} \pi_1(U), & \text{if } t = 1, \\ \pi_t(Z_{t-1}, \ldots, Z_1, U), & \text{if } t \geq 2. \end{cases} \tag{32}$$

The collection of mappings $\pi = \{\pi_t\}_{t\geq 1}$ that generate such set-valued functions $\{\mathcal{C}_t(\cdot) : t = 1, 2, \cdots\}$ constitutes a class of algorithms denoted by $\mathcal{P}$. Moreover, we restrict attention to a structured subclass of $\mathcal{P}$ in which each prediction set is expressible as a finite union of intervals.

Definition 4.1 (K-interval procedure)**.** *For every integer $K \geq 1$, define the K-interval algorithm class $\mathcal{P}_K \subseteq \mathcal{P}$ as*

$$\mathcal{P}_K := \left\{\pi \in \mathcal{P} \,\middle|\, \text{for all } t \in [T] \text{ and } x \in \mathcal{X} :\ \mathcal{C}_t(x) \text{ is the union of at most } K \text{ intervals.}\right\} \tag{33}$$

We shall discuss the practical relevance of this algorithm subclass momentarily.

- *Distribution class.* Analogous to the score-based distribution class $\mathcal{L}_1(N^{\mathsf{cp}})$ (see (18a)) that pertains to the change-point setting, we introduce a closely related distribution class—defined directly in terms of the data distributions—that permits at most N^{cp} change points:

$$\mathcal{L}_3(N^{\mathsf{cp}}) := \left\{(\mathcal{D}_1, \ldots, \mathcal{D}_T) :\ (\mathcal{D}_1, \ldots, \mathcal{D}_T) \text{ change at most } N^{\mathsf{cp}} \text{ times.}\right\}; \tag{34a}$$

Additionally, we define another TV-based distribution class concerning the smooth drift setting: for a given budget $\mathsf{TV}_T > 0$, define

$$\mathcal{L}_4(\mathsf{TV}_T) \ := \ \left\{(\mathcal{D}_1, \ldots, \mathcal{D}_T) :\ \sum_{t=1}^{T-1} \mathsf{TV}(\mathcal{D}_t, \mathcal{D}_{t+1}) \leq \mathsf{TV}_T \right\}. \tag{34b}$$

Moreover, for a distribution class $\mathcal{L}$, the worst-case regret of algorithm $\pi \in \mathcal{P}_K$ is defined as

$$\mathsf{regret}_\pi(\mathcal{L}, T, K) := \sup_{(\mathcal{D}_1, \ldots, \mathcal{D}_T) \in \mathcal{L}} \mathsf{regret}_\pi(\mathcal{D}_{1:T}, T), \tag{35}$$

where we make explicit the dependency on K. We can now present our minimax lower bound that accommodates online conformal prediction with adaptive training.

Theorem 4.2. *Consider any fixed constant $\alpha \in (0, 1/2]$. Suppose that Assumption 4.1 holds. For any admissible algorithm $\pi \in \mathcal{P}_K$, the worst-case regret under π has the following lower bound:*

$$\mathsf{regret}_\pi\big(\mathcal{L}_3(N^{\mathsf{cp}}), T, K\big) = \widetilde{\Omega}\left(\min\left\{\sqrt{(N^{\mathsf{cp}}+1)T}, \frac{T}{\sqrt{K}}\right\}\right);$$
$$\mathsf{regret}_\pi\big(\mathcal{L}_4(\mathsf{TV}_T), T, K\big) = \widetilde{\Omega}\left(\min\left\{\sqrt{T} + (\mathsf{TV}_T)^{\frac{1}{3}} T^{\frac{2}{3}} K^{-\frac{1}{6}}, \frac{T}{\sqrt{K}}\right\}\right).$$

The proof of Theorem 4.2 is deferred to Section C.4. Clearly, when K is a finite constant, the regret bound in (29) matches this minimax lower bound up to a logarithmic factor, provided $L = O(1)$ (meaning that the learning algorithm is stable and the conditional response distribution is smooth).

Why restricted to $\mathcal{P}_K$? We now elucidate the rationale for restricting attention to the algorithm subclass $\mathcal{P}_K$. In brief, imposing structural constraints on the algorithm class is necessary to formulate a meaningful minimax problem. Without such restrictions, one can design "irregular" procedures that achieve asymptotically perfect marginal coverage while using essentially no information about the data-generating process. To illustrate this point, suppose $Y \in [0, 1]$. For each $n \geq 1$, define

$$\mathcal{C}_n := \bigcup_{i=0}^{n-1} \left[\frac{i}{n}, \frac{i + (1-\alpha)}{n}\right], \tag{36}$$

obtained by partitioning $[0, 1]$ into n equal subintervals and retaining the same $(1-\alpha)$-fraction of each subinterval. For sufficiently regular distributions, $\mathcal{C}_n$ captures roughly a $(1-\alpha)$-fraction of the total probability mass, largely independent of the actual shape of the density. The following proposition formalizes this observation.

Proposition 4.2. *Let $\{\mathcal{C}_n\}_{n\geq 1}$ be defined by Eqn. (36). If the distribution $\mathcal{D}$ of Y on $[0,1]$ admits a Riemann-integrable density, then*

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

This example shows that, in the absence of geometric constraints, marginal coverage alone does not preclude vacuous procedures. Restricting attention to $\mathcal{P}_K$, where each prediction interview is a union of at most K intervals, excludes such uninformative construction and yields a more meaningful lower bound.

Implications beyond the online setting. Although Theorem 4.2 is stated for the online setting, its proof proceeds by first establishing a lower bound on the per-round contribution to the cumulative regret, and then constructing a distribution sequence that allocates the available distribution drift budget in a way that realizes these per-round bottlenecks. As a byproduct, our arguments readily yield an *offline* lower bound for training-conditional coverage error over the algorithm class $\mathcal{P}_K$ (see Definition 4.1). We record this consequence below, which may be of independent interest.

Proposition 4.3. *Fix any $\alpha\in(0,1/2]$. Let $\mathcal{S}$ be the collection of distributions on $\mathcal{X}\times\mathbb{R}$ that admit a density. Let $\{(X_i,Y_i)\}_{i=1}^n$ be i.i.d. draws from some $\mathcal{D}\in\mathcal{S}$, and let $U\sim\mathrm{Unif}(0,1)$ be independent of the data. Consider any algorithm π that maps $\{(X_i,Y_i)\}_{i=1}^n$ and U to a set-valued function $\widehat{\mathcal{C}}(\cdot):\mathcal{X}\to\mathcal{B}(\mathbb{R})$ such that, for each $x\in\mathcal{X}$, the set $\widehat{\mathcal{C}}(x)$ is a union of at most K intervals. Then we have*

$$\sup_{\mathcal{D}\in\mathcal{S}}\mathbb{E}\left[\left|\mathop{\mathbb{P}}_{(X,Y)\sim\mathcal{D}}\left(Y\in\widehat{\mathcal{C}}(X)\,\middle|\,\{(X_i,Y_i)\}_{i=1}^n;U\right)-(1-\alpha)\right|\right]\;=\;\widetilde{\Omega}\left(\min\left\{\frac{1}{\sqrt{K}},\frac{1}{\sqrt{n}}\right\}\right),$$

where the outer expectation is taken over the training sample $\{(X_i,Y_i)\}_{i=1}^n$ and the internal randomization U.

Proposition 4.3—which is a direct consequence of Lemma C.6 given in Section C.4—is independent of the online setting and provides a lower bound for training-conditioned validity of full conformal prediction in the offline regime. The result places no parametric restriction on the prediction set $\widehat{\mathcal{C}}(\cdot)$, and is therefore fundamentally different from those information-theoretic lower bounds in classical parametric estimation problems. The bound in Proposition 4.3 holds for a fixed algorithm and considers the worst case over a class of distributions, complementing the result of Bian and Barber (2023), which instead fixes the distribution and takes the worst case over a class of algorithms. Moreover, relative to prior work, our bound explicitly characterizes the learning limit in terms of the structural complexity K of the prediction sets (i.e., the number of intervals). Determining the optimal K-dependence in training-conditional lower bounds remains an interesting open direction, which we leave for future work. We view this lower bound as a baseline that may be useful more broadly in the study of conformal inference beyond the online setting.

4.4 Comparisons with prior art

Existing training-conditional guarantees. Prior literature has established training-conditional coverage guarantees for split conformal methods (Vovk, 2012). More recently, Bian and Barber (2023) showed that such guarantees can be achieved in a distribution-free manner by K-fold CV+ when the sample size is sufficiently large relative to the number of folds, but not achievable by full conformal methods or jackknife+. Training-conditional coverage guarantees for full conformal methods and jackknife+ have instead largely been obtained under stability-type assumptions; see, e.g., Liang and Barber (2025); Amann et al. (2023) and Pournaderi and Xiang (2024). Proposition 4.1 strengthens these results in three complementary ways.

First, Liang and Barber (2025) expressed their bounds through an m-stability quantity $\beta^{\mathsf{out}}_{m,n-1}$ (see Definition 3.1 therein), yielding a coverage error of order $O\left(\sqrt{\frac{\log(1/\delta)}{\min\{m,n\}}}+\left(\beta^{\mathsf{out}}_{m,n-1}\right)^{\frac{1}{3}}\right)$ (see their Theorems 3.2 and 4.1). To obtain $o_p(1)$ error, one needs both $m\to\infty$ and $\beta^{\mathsf{out}}_{m,n-1}\to 0$, effectively requiring the fitted predictor $\widehat{\mu}(\cdot)$ to stabilize as more data arrive. As noted by Amann et al. (2023, Lemma B.7), such stabilization can fail under distribution shift, where $\widehat{\mu}$ typically adapts to the evolving data distribution. In contrast, Proposition 4.1 avoids an explicit m-stability requirement and remains applicable in drifting scenarios.

Second, Amann et al. (2023, Proposition A.2) established a training-conditional *under*-coverage bound that inflates the prediction set by an additional slack $\Delta>0$, leading to conservativeness and potentially

wide prediction sets. Their confidence dependence scales as $\delta^{-1/2}$, whereas Proposition 4.1 attains a sharper logarithmic $\log(1/\delta)$ dependence without requiring an explicit inflation parameter.

Third, Pournaderi and Xiang (2024, Theorem 6) provided a training-conditional coverage guarantee for full conformal prediction under covariate shift. Their result, however, does not address shifts in the conditional distribution $Y \mid X$. Moreover, relative to Proposition 4.1, their analysis relies on additional structural assumptions, including a uniform upper bound on the train–test density ratio and a parametric model for the fitted predictor with a bi-Lipschitz dependence on its parameters, which could narrow the range of settings in which the bound can be verified.

Verification of stability conditions. Stability analyses for both empirical loss minimizers and stochastic optimization methods have been developed in, e.g., Barber et al. (2021), Ndiaye (2022) and Lee and Zhang (2025). Compared to our work (mainly our results in Section D), these prior results typically differ in several important respects. First, they did not accommodate constrained optimization problems. Second, they often assume that, for every data realization z, the loss $\ell(\cdot, z)$ is uniformly strongly convex, which most naturally holds for explicitly regularized ERM objectives; in contrast, our verification only requires strong convexity of the *population* risk L on $\mathcal{C}$. Third, while Lee and Zhang (2025) (see their Example 2) examined stochastic optimization methods, their analysis is essentially offline—the algorithm is rerun multiple times on a fixed dataset under different permutations—and is derived for fixed stepsizes, which does not yield a stability coefficient that vanishes with the sample size. Such vanishing stability is crucial in our online full conformal analysis in order to obtain training-conditional coverage guarantees.

5 Numerical experiments

5.1 Experiments: online conformal prediction with pretrained scores

In this subsection, we evaluate the performance of the proposed DriftOCP algorithm against the ACI method under various distribution shift scenarios, assuming the presence of pretrained score functions. The experimental setup is described below.

Data generation. Consider a regression setting with a data stream $\{(X_t, Y_t)\}_{t\geq 1}$. The feature vector X_t is in $\mathbb{R}^d$ with $d = 5$, and each component is generated by $X_{t,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The response variable satisfies

$$Y_t = 2X_{t,1} + X_{t,2} + \mu_t + \sigma_t \cdot \varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

where μ_t and σ_t are varying parameters. We examine four distribution shift cases as follows.

- **Setting 1 (piecewise variance shift):** $\mu_t = 0$ and

$$\sigma_t = \begin{cases} 0.5, & \text{if } t < 4000, \\ 2.0, & \text{if } 4000 \leq t < 7000, \\ 3.5, & \text{if } t \geq 7000. \end{cases}$$

  This setting simulates abrupt changes in noise level, representing sudden regime shifts.

- **Setting 2 (linear bias drift):** $\sigma_t = 0.5$ and $\mu_t = \kappa \cdot t$ with $\kappa = 0.002$, so that $\mu_T = 20$ at $T = 10000$. This represents smooth temporal drift in the conditional mean.

- **Setting 3 (smooth variance growth):** $\mu_t = 0$ and $\sigma_t = \sqrt{1 + 0.008t}$. This model continuously increases variability over time.

- **Setting 4 (no distribution drift):** $\mu_t = 0$ and $\sigma_t = 0.5$ for all t. This serves as a baseline where no distribution shift occurs.

Experimental protocol. For each setting, we use a training set of size $n = 500$ drawn from the initial distribution ($t = 0$) to fit a random forest regressor (Breiman, 2001) with 100 trees, implemented in scikit-learn (Pedregosa et al., 2011), as the pre-trained prediction model. The pretrained predictive model $\widehat{\mu}$ remains fixed throughout the online prediction phase and is not updated.

The non-conformity score at each time step t is taken to be the absolute residual between the observed and predictive responses, namely, $s_t := |Y_t - \widehat{\mu}(X_t)|$. The initial quantile $\widehat{q}_0$ is set to be the $(1-\alpha)$-th empirical quantile of the training residuals $\{|Y_i - \widehat{\mu}(X_i)|\}_{i=1}^n$.

The test horizon is $T = 10{,}000$ time steps. We set the target miscoverage level to be $\alpha = 0.1$. All experiments are repeated 40 times with different random seeds, and we report the mean and standard deviation of cumulative regret.

Numerical evaluation of cumulative regret. We measure performance using the cumulative regret defined in Eqn. (13). However, since the true miscoverage probability $\mathbb{P}(s_t > q_t \mid q_t)$ is intractable, we estimate it via Monte Carlo simulation. Specifically, for each time step t, we pre-generate a fixed evaluation set of $M = 500$ independent samples $\{(X_t^{(m)}, Y_t^{(m)})\}_{m=1}^M$ from the true distribution $\mathcal{D}_t$ at that time step. The instantaneous coverage rate (defined in Eqn. (11)) is then estimated as

$$\widehat{\mathsf{cvg}}_t = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\left\{|Y_t^{(m)} - \widehat{\mu}(X_t^{(m)})| \le q_t\right\}.$$

The cumulative regret up to time T is then calculated as

$$\widehat{\mathsf{Regret}}_T = \sum_{t=1}^{T} \left|\widehat{\mathsf{cvg}}_t - (1-\alpha)\right|.$$

Methods for comparison. We compare the following algorithms numerically.

- **DriftOCP (Algorithm 2):** We use a drift detection threshold in Algorithm 2 of $\sigma_{n,r} = 4$. To avoid false positives from high-variance estimates, we require a minimum window size of $t \ge 10$ before any drift detection can be declared.
- **ACI with decaying stepsizes (Angelopoulos et al., 2024a):** $\eta_t = (t+1)^{-\gamma}$ for $\gamma \in \{0.5, 0.6\}$.
- **ACI with fixed stepsizes (Gibbs and Candes, 2021):** $\eta_t \equiv \eta \in \{0.01, 0.1, 0.5\}$.

Results. Figure 3 summarizes both the regret and calibration dynamics across the four data-generating settings. The top row plots the cumulative regret over time, while the bottom row tracks the corresponding evolution of the calibration quantiles; the black dashed curve in the bottom row is an approximation of the ground-truth quantile, obtained via repeated simulations at each time point. Taken together, the two rows highlight the tuning trade-off of ACI: a large constant stepsize reacts quickly to distributional changes, but yields highly variable quantile trajectories even under stationarity, leading to substantial cumulative regret; conversely, smaller or decaying stepsizes stabilize the quantile updates in stationary periods, yet may adapt too slowly after distribution shifts and consequently lag behind the moving target (most notably in Setting 1). As a result, the optimal stepsize for ACI differs across various settings, making it difficult to select a priori in practice. In contrast, DriftOCP adapts to different regimes in a data-driven manner, achieving stable tracking during stationary segments and rapid re-alignment following change points. This behavior leads to uniformly controlled regret across regimes, comparable to that attained by hindsight-optimal tuning of ACI.

5.2 Experiments: online conformal prediction with adaptively trained scores

This section examines how our methods interact with different ways of generating non-conformity scores in the presence of distribution drift. In particular, we pair our drift-aware recalibration mechanism with (i) a covariate-agnostic score, (ii) a fixed pretrained model, and (iii) an adaptively updated fitted model. The numerical comparisons illustrate tangible gains from adaptively updated models in terms of predictive efficiency and validity.

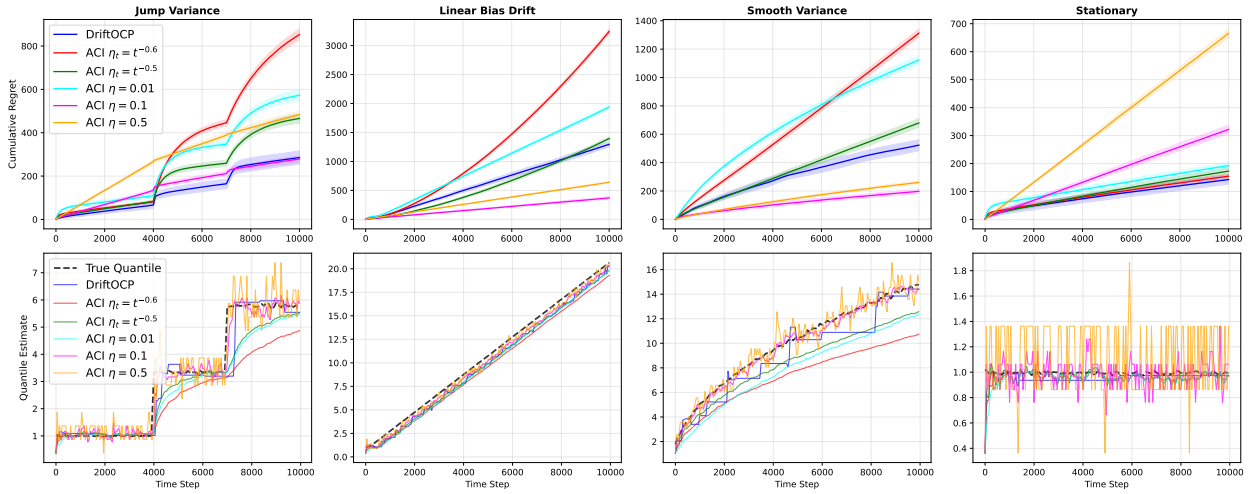

Figure 3: **Cumulative regret and calibration quantiles under four data-generating settings.** Top row: cumulative regret trajectories. Bottom row: calibration-quantile evolution; the black dashed curve indicates an approximation to the ground-truth quantile obtained via repeated simulations at each time point. Across settings, ACI exhibits a clear stepsize trade-off: ACI with large constant stepsizes adapts quickly but produces volatile quantile updates and suboptimal performance under stationarity, whereas ACI with smaller or decaying stepsizes yields more stable updates at the cost of slower adaptation to distributional changes. In comparison, DriftOCP is stable within stationary time segments and adapts rapidly to distribution shifts, yielding consistently controlled regret. Curves are averaged over 20 runs; shaded bands indicate ± 1 standard deviation.

Data generation. We consider an online regression stream $\{(X_t, Y_t)\}_{t\geq 1}$ with feature dimension $d = 10$. Throughout these experiments, we construct prediction sets using a *linear* predictor, and the non-conformity score function is computed from linear regression residuals. We then consider two data-generating models for (X_t, Y_t) to distinguish well-specified learning from misspecification:

- **Well-specified case:** $Y_t = X_t^\top \beta^* + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, 1)$, so the linear predictor can be correctly specified;
- **Misspecified case:** $Y_t = X_t^\top \beta^* + \frac{1}{100}\|X_t\|_2^2 + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, 1)$, where the additional quadratic term introduces a mild deviation from linearity.

In both cases, the true coefficient $\beta^* \in \mathbb{R}^d$ is sampled from $\mathcal{N}(0, I_d)$ and is subsequently held fixed across simulation repetitions. For each model specification, we introduce piecewise-stationary covariate shifts with change points at $t \in \{3333, 6667\}$. Specifically, we consider:

- **Mean shifts:** $X_t \sim \mathcal{N}(\mu_t \mathbf{1}_d, I_d)$, where

$$\mu_t = \begin{cases} 0, & \text{if } t \leq 3333; \\ 3, & \text{if } 3333 < t \leq 6667; \\ -2, & \text{if } t > 6667; \end{cases}$$

- **Variance shifts:** $X_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$, where

$$\sigma_t = \begin{cases} 1, & \text{if } t \leq 3333; \\ 5, & \text{if } 3333 < t \leq 6667; \\ 10, & \text{if } t > 6667. \end{cases}$$

In each run of the experiments, we first draw $n_{\text{pretrain}} = 100$ independent observations to fit an initial ridge regression model, and then draw an additional $n_{\text{train}} = 500$ observations to initialize the calibration quantile for the non-conformity scores. We subsequently generate an online data stream of length $T = 10{,}000$.

Score construction strategies. We compare three score construction strategies that are paired with drift-aware recalibration. The first uses our full-conformal variant tailored to online optimization, while the latter two use DriftOCP with pretrained score functions:

- **DriftOCP-full + online SGD:** the score is formed using a sequentially updated fitted model, $s_t = |Y_t - X_t^\top \widehat{\beta}_t|$, where $\widehat{\beta}_t$ is updated by online SGD with stepsize $\eta_t = 0.01/\sqrt{t}$;
- **DriftOCP + pretrained ridge:** the score uses a fixed ridge-regression predictor (Hoerl and Kennard, 1970) with regularization parameter $\lambda = 1.0$, implemented in scikit-learn (Pedregosa et al., 2011), as the pretrained model;
- **DriftOCP + absolute response:** a covariate-agnostic baseline $s_t = |Y_t|$.

All methods employ the same drift-detection threshold $\sigma = 4$ and the same doubling-round structure. The target miscoverage level is $\alpha = 0.1$ for all settings.

Results. We report the prediction interval width and the local coverage rate computed over a sliding window of 100 time steps. As shown in Figure 4, the results are averaged over 20 independent runs, with shaded regions indicating ± 1 standard deviation. The top row plots the prediction interval width over time, while the bottom row shows the local coverage rate computed using a rolling window of 100 steps. The horizontal dashed line marks the target level $1 - \alpha = 0.9$ and the vertical dashed lines correspond to the change points.

Across all four settings—well-specified or misspecified models, and under either mean or variance drift—the adaptive-score method with online SGD consistently achieves the most favorable tradeoff, producing the narrowest intervals while maintaining stable coverage around the target level. In contrast, the pretrained-score baseline tends to be sensitive to mismatches between the pretraining and test covariate distributions, resulting in wider intervals and a higher degree of coverage fluctuations after the change points. The model-free baseline ($s_t = |Y_t|$) is in general conservative and produces substantially wider intervals than the other two methods throughout. It is also sensitive to distribution shift, exhibiting undercoverage at change points.

We also observe a transient effect at the beginning of the data stream: the adaptive method exhibits slightly biased local coverage and inflated widths early on, which is as expected since the fitted model is still in its initialization phase and the online updates are relatively volatile. As more data arrive, the adaptively fitted model stabilizes, and the resulting score becomes better calibrated, after which the method tracks the target coverage tightly even after distributional shifts.

Finally, note that the adaptive method relies on an online SGD-trained predictor, whose trajectory depends on the data order and thus does not satisfy permutation symmetry. The strong empirical performance of this non-symmetric learning pipeline provides additional evidence supporting our training-conditional guarantees, which do not require symmetry of the underlying fitted model.

6 Additional related work

In this section, we briefly discuss a small sample of other related papers. To start with, a substantial body of work has established theoretical coverage guarantees for conformal prediction (Angelopoulos and Bates, 2021; Angelopoulos et al., 2024b). The majority of these results, however, rely on the assumption that the data are exchangeable (most notably, i.i.d. observations) (e.g., Vovk et al. (2005); Vovk (2015); Lei et al. (2018); Barber et al. (2021)). A growing literature has investigated how conformal prediction procedures can be modified to retain validity when exchangeability is violated, aiming to preserve meaningful coverage guarantees under various forms of distributional shift. For instance, Tibshirani et al. (2019); Barber et al. (2023) developed weighted split conformal methods that restore *marginal* validity under a weighted-exchangeability condition. Their methods rely on importance weights tied to the data distribution; in a distribution-free setting, one must either use a non-data-dependent weight (as in Barber et al. (2023)) or impose stronger structural assumptions (e.g., invariance of $Y \mid X$ as in (Tibshirani et al., 2019)). In parallel, Podkopaev and Ramdas (2021); Si et al. (2024) developed split conformal methods under label shift, when the marginal distribution of Y differs across environments while the conditional distribution $P(X \mid Y)$ remains invariant. In another line of work (Chernozhukov et al., 2018; Cauchois et al., 2024; Ai and Ren, 2024; Gui et al., 2024; Aolaritei

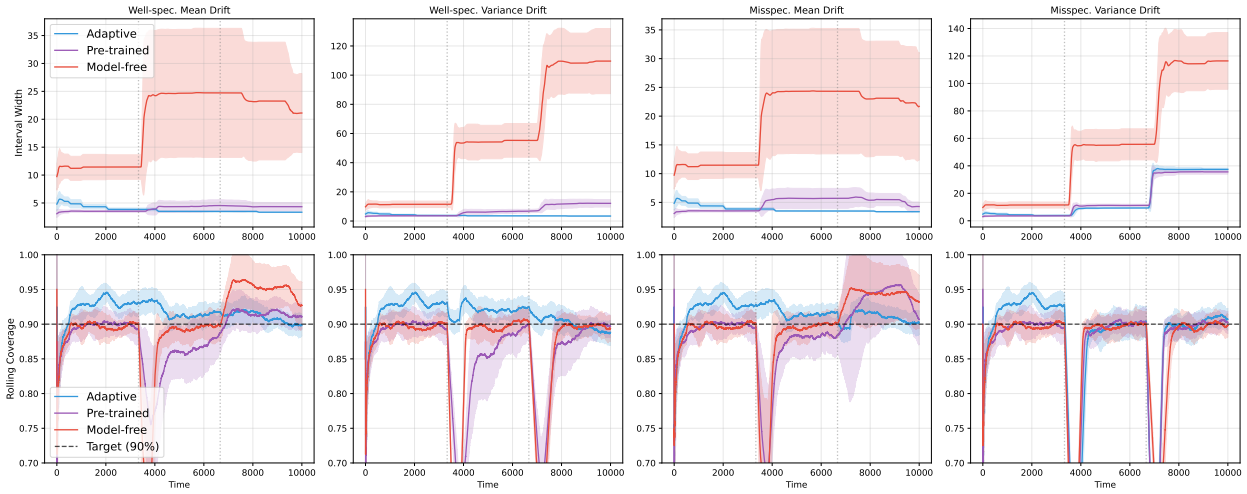

Figure 4: **Online conformal prediction with different score constructions.** Top row: prediction-interval width over time. Bottom row: local coverage rate computed with a rolling window of 100 steps; the horizontal line marks the target level $1-\alpha = 0.9$. Vertical dashed lines indicate change points at $t = 3333$ and $t = 6667$. Columns correspond to the four settings (well-specified vs. misspecified model, each under mean vs. variance drift). The adaptive-score method (online SGD) yields noticeably shorter intervals and more stable coverage under variance drift, whereas the pretrained-score method is sensitive to a mismatch between the pretraining and test covariate distributions. The model-free baseline ($s_t = |Y_t|$) is in general conservative and produces wide prediction intervals; it is also sensitive to distribution shift, exhibiting undercoverage at change points. Curves are averaged over 20 runs, with shaded bands indicating ± 1 standard deviation.

et al., 2025), distribution shifts between training and test environments are tackled from a distributionally robust optimization perspective. Meanwhile, a growing literature developed conformal prediction methods for time-series data (Zaffran et al., 2022; Xu and Xie, 2021, 2023b,a; Xu et al., 2024; Chen et al., 2024; Cleaveland et al., 2024; Stocker et al., 2025).

For temporally dependent, non-exchangeable sequences, Oliveira et al. (2024) showed that split conformal retains *approximate marginal* validity up to an explicit penalty term controlled by decoupling/mixing conditions (including β-mixing), with sharper results subsequently obtained in Barber and Pananjady (2025). These results, however, do not yield sharp training-(and-calibration)-conditional concentration bounds and typically require stationarity-type dependence assumptions. In contrast, training-(and-calibration)-conditional coverage guarantees beyond exchangeability are comparatively rarer.

In addition, a line of work connected ACI-style calibration with ideas from online learning and studied regret bounds under various performance criteria (Bhatnagar et al., 2023; Gibbs and Candès, 2024; Zhang et al., 2024a; Ramalingam et al., 2025; Liu et al., 2026), which, however, fell short of ensuring training-conditional coverage. Moreover, the idea of ACI has been extended for broader settings, such as risk control (Feldman et al., 2022; Farinhas et al., 2024), stochastic control for time series (Yang et al., 2024), and parametric quantile calibration (Areces et al., 2025), among other things.

7 Discussion

In this work, we have developed two online conformal prediction methods that adapt efficiently to temporal distribution drift, producing prediction sets that are both valid and informative. For scenarios involving pretrained score functions, our DriftOCP algorithm leverages an efficient drift detection subroutine to update the calibration set sequentially, achieving regret bounds that are minimax optimal (up to logarithmic factors) across both change-point and smooth drift regimes. For scenarios where the predictive models (and hence the score functions) are trained adaptively from prior observations, we propose DriftOCP-full, a full-conformal-style online algorithm that enjoys strong regret guarantees under stability assumptions; for this setting, we

further establish matching minimax lower bounds under suitable restrictions on the prediction sets. Unlike much of the prior work that focused on metrics like empirical long-term coverage frequency or adversarial regret, our analysis exploits additional independence assumption across data while remaining otherwise distribution-free, which has enabled training-conditional regret guarantees. The proposed algorithms are horizon-free, computationally efficient, and supported by fully non-asymptotic, minimax-optimal theoretical guarantees.

Our results naturally suggest several directions for future investigation. To begin with, while our current theoretical development hinges upon independence across data samples, many online predictive inference problems involve temporally dependent observations, as commonly encountered in, say, time-series settings (Xu and Xie, 2021; Zaffran et al., 2022; Oliveira et al., 2024). A natural and challenging direction is therefore to extend our online conformal methods to tackle temporally dependent, non-stationary Markovian environments, while still delivering rigorous training-conditional coverage guarantees. On another front, the training-conditional guarantees in Section 4 rely on stability assumptions for the underlying model-fitting algorithms. When the desirable stability conditions fail to hold or are difficult to verify—e.g., in certain nonparametric, or deep learning models (Gibbs and Candès, 2025; Lei et al., 2011; Romano et al., 2019)—it is fundamentally important to develop new online full conformal methods that ensure both validity and efficiency without stability. Finally, our algorithmic and analysis frameworks might shed light on other online statistical problems, such as online multicalibration (Collina et al., 2026).

Acknowledgments

Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the ONR grant N00014-25-1-2344, the NSF grants 2221009 and 2218773, the Wharton AI & Analytics Initiative's AI Research Fund, and the Amazon Research Award. Z. Ren is supported by the NSF grant DMS-2413135 and Wharton Analytics. Y. Chen would like to thank Jiahao Ai for extensive discussion about adaptive conformal inference.

A Proof of Fact 2.1

Proof of (i). By the triangle inequality, we obtain

$$\begin{aligned}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}(Y_t\in\mathcal{C}_t)-(1-\alpha)\right| &\le \frac{1}{T}\sum_{t=1}^{T}\left|\mathbb{P}(Y_t\in\mathcal{C}_t)-(1-\alpha)\right| \\ &\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{C}_t}\left[\left|\mathbb{P}\big(Y_t\in\mathcal{C}_t\mid\mathcal{C}_t\big)-(1-\alpha)\right|\right] \le \frac{\mathsf{regret}_T}{T}.\end{aligned}$$

Proof of (ii). For any time point t, let us construct $\mathcal{C}_t$ as follows:

$$\mathcal{C}_t := \begin{cases}\varnothing, & \text{with probability } \alpha,\\ \mathbb{R}, & \text{with probability } 1-\alpha.\end{cases}$$

On the one hand, it can be easily derived that, for any $1\le t\le T$,

$$\mathbb{P}(Y_t\in\mathcal{C}_t)=\alpha\mathbb{1}\{Y_t\in\varnothing\}+(1-\alpha)\mathbb{1}\{Y_t\in\mathbb{R}\}=1-\alpha,$$

thus implying that

$$\mathsf{lt\text{-}cvg}_T=\frac{1}{T}\sum_{t=1}^{T}\mathbb{P}(Y_t\in\mathcal{C}_t)=1-\alpha.$$

On the other hand, it is seen that

$$\begin{aligned}&|\mathbb{P}(Y_t\in\varnothing)-(1-\alpha)|=1-\alpha;\\ &|\mathbb{P}(Y_t\in\mathbb{R})-(1-\alpha)|=|1-(1-\alpha)|=\alpha.\end{aligned}$$

Therefore, the following holds naturally

$$|\mathbb{P}(Y_t \in \mathcal{C}_t \mid \mathcal{C}_t) - (1-\alpha)| \geq \min\{1-\alpha, \alpha\} = \alpha.$$

Summing over all $t = 1, \ldots, T$ and recalling the definition of regret_T, we complete the proof.

B Detailed proofs in Section 3

Before proceeding, let us introduce some convenient notation.

B.1 Proof of Theorem 3.1

We now turn to the proof of Theorem 3.1. At a high level, we shall first establish a per-round regret bound, then aggregate these bounds within each stage, and finally sum over all stages to obtain regret bounds over the entire time horizon $[T]$.

B.1.1 Additional notation

To facilitate presentation for the analysis of Algorithm 2, we introduce some additional notation. First, write

$$s_{n,r,l} = s_{n,r,l}(X_{n,r,l}, Y_{n,r,l}) \qquad \text{and} \qquad s_t = s_t(X_t, Y_t)$$

as long as it is clear from the context. In addition, for any $1 \leq i < j \leq T$, we define the "typical" event:

$$\mathcal{A}(i,j) \coloneqq \left\{ \sup_{x \in \mathbb{R}} \left\{ \left| \sum_{t=i}^{j} \big(\mathbb{1}\{s_t \leq x\} - \mathbb{P}(s_t \leq x) \big) \right| \right\} \leq 6\sqrt{(j-i+1)\log j} \right\}. \tag{37a}$$

To see that this is a high-probability event, invoking Lemma E.4 with $\delta = j^{-6}$ and using the fact

$$\frac{4}{\sqrt{j-i+1}} + \sqrt{\frac{6\log j}{2(j-i+1)}} \leq 4\sqrt{\frac{\log j}{j-i+1}} + \sqrt{\frac{3\log j}{j-i+1}} < 6\sqrt{\frac{\log j}{j-i+1}},$$

we can establish that

$$\mathbb{P}\big(\mathcal{A}(i,j)^{\mathrm{c}}\big) \leq j^{-6}. \tag{37b}$$

Moreover, for any stage-round pair (n, r), we define

$$\mathcal{A}_{n,r} \coloneqq \bigcap_{i=\tau_{n,r}}^{\tau_{n,r+1}-1} \bigcap_{j=i+1}^{\tau_{n,r+1}-1} \mathcal{A}(i,j), \tag{37c}$$

where we recall that $\tau_{n,r}$ indicates the time at which round r of stage n begins and $\tau_{n,r_n+1} = \tau_{n+1,1}$. For convenience, we also write

$$\tau_n \coloneqq \tau_{n,1}. \tag{38}$$

Moreover, we introduce several notations for the change-point setting. Recall that under the change-point model, the entire horizon $[T]$ is partitioned into $N^{\mathsf{cp}} + 1$ time segments, within each of which the score distributions remain fixed.

Definition B.1. *We define the following notation:*

- $\mathcal{I}_1, \cdots, \mathcal{I}_{N^{\mathsf{cp}}+1}$*: the* $N^{\mathsf{cp}} + 1$ *time segments over the entire horizon;*
- $K_{n,r}$*: the total number of time segments in round* r *of stage* n*;*
- $S_{n,r}$ *and* t_n*: the total number of iterations in round* r *of stage* n*;*
  in particular, let $t_n = S_{n,r_n}$*, i.e., the number of iterations in the last round of this stage;*

- $\mathcal{I}_{n,r,k}$ $(k = 1, \cdots, K_{n,r})$: *the k-th time segment in round r of stage n;*
- S_n: *the total number of iterations in stage n;*
- J_n: *the total number of time segments in stage n;*
- $\mathcal{I}_{n,j}$ $(j = 1, \cdots, J_n)$: *the j-th time segment in stage n.*

Also, for any time segment $\mathcal{I}$, we let $|\mathcal{I}|$ represent the length of this time segment. In addition, while the last round r_n of stage n contains $t_n \le T_{r_n}$ iterations, we generate—for convenience of presentation—a set of random variables $\{s_{n,r,l}\}$ for $l > t_n$ in a way that obeys

$$\mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) = \alpha \qquad \text{for all } l > t_n. \tag{39}$$

Moreover, for round r in stage n, we define the cumulative KS distance within this round as

$$\mathsf{KS}_{n,r}^{\mathsf{round}} \coloneqq \sum_{l=1}^{S_{n,r}-1} \mathsf{KS}(s_{n,r,l}, s_{n,r,l+1}). \tag{40a}$$

We also define the cumulative KS distance within stage n (which contains r_n rounds) as

$$\mathsf{KS}_{n}^{\mathsf{stage}} \coloneqq \sum_{r=1}^{r_n} \mathsf{KS}_{n,r}^{\mathsf{round}}. \tag{40b}$$

B.1.2 Decomposing and bounding the cumulative regret

In order to bound the cumulative regret, we first decompose it based on stages and rounds of DriftOCP as well as the typical events $\{\mathcal{A}_{n,r}\}$ introduced in Section B.1.1:

$$\begin{aligned}\sum_{t=1}^{T} \big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\big| = &\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r} \big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big| \mathbb{1}\{\mathcal{A}_{n,r}\} \\ &+ \sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r} \big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big| \mathbb{1}\{\mathcal{A}_{n,r}^{\mathrm{c}}\}.\end{aligned} \tag{41}$$

As it turns out, the first term on the right-hand side of (41) serves as the dominant term, as argued below.

Consider any time point t belonging to round r of stage n. According to the procedure of DriftOCP—particularly the fact that the rounds length grow geometrically with $T_r = 3^r$—it follows that neither $t/4$ nor $4t$ lies within the same round (n, r), and as a result,

$$\mathcal{E}_t \coloneqq \bigcap_{j=t/4}^{4t} \bigcap_{i=1}^{j} \mathcal{A}(i,j) \subseteq \mathcal{A}_{n,r}.$$

It can thus be seen from (37b) that

$$\mathbb{P}(\mathcal{E}_t^{\mathrm{c}}) \le \sum_{j=t/4}^{4t}\sum_{i=1}^{j} \mathbb{P}\big(\mathcal{A}(i,j)^{\mathrm{c}}\big) \le \sum_{j=t/4}^{4t}\sum_{i=1}^{j} \frac{1}{j^4} \le \sum_{j=t/4}^{4t} \frac{1}{j^3} \le \frac{32}{t^2},$$

which helps us control the second term on the right-hand side of (41) as

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r} \big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big| \, \mathbb{1}\{\mathcal{A}_{n,r}^{\mathrm{c}}\}\right] &\le \mathbb{E}\left[\sum_{t=1}^{T} \big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\big| \, \mathbb{1}\{\mathcal{E}_t^{\mathrm{c}}\}\right] \\ &\le \sum_{t=1}^{T} \mathbb{P}\big(\mathcal{E}_t^{\mathrm{c}}\big) \le \sum_{t=1}^{T} \frac{32}{t^2} = O(1).\end{aligned} \tag{42}$$

As a consequence, the remainder of this proof is devoted primarily to bounding the first term of (41). Towards this end, we begin by looking at the cumulative coverage gaps in round r of stage n. Informally, in the change-point setting, the cumulative coverage gap over this round on the typical events defined in Section B.1.1 is upper bounded by a sum of square-root terms in the lengths of the time segments. In contrast, under smooth drift, the bound contains a term that scales with a suitable KS distance raised to the $1/3$ power, in a manner that resembles the final regret bound in Theorem 3.1. This is stated in the following lemma, with the proof deferred to Section B.3.1.

Lemma B.1. *Consider any stage-round pair (n, r) in Algorithm 2. If no distribution shift has been detected by the subroutine* DriftDetect *by the end of this round, then it holds that*

$$\left(\sum_{l=1}^{T_r}\Big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\Big|\right)\mathbb{1}\{\mathcal{A}_{n,r}\} = \begin{cases} \widetilde{O}\left(\sum_{k=1}^{K_{n,r}}\sqrt{|\mathcal{I}_{n,r,k}|}\right), & \text{for the change-point setting;} \\ \widetilde{O}\left(\sqrt{T_r} + (\mathsf{KS}_{n,r}^{\mathsf{round}})^{\frac{1}{3}} T_r^{\frac{2}{3}}\right), & \text{for the smooth drift setting.} \end{cases}$$

Owing to the doubling trick employed in DriftOCP (i.e., the round lengths grow geometrically), we can lift the per-round cumulative gap bound in Lemma B.1 to a per-stage cumulative gap bound, again on the typical events defined in Section B.1.1. This result is formalized in the following lemma, whose proof is postponed to Section B.3.2.

Lemma B.2. *Consider any stage n in Algorithm 2, which comprises r_n rounds. Then we have*

$$\sum_{r=1}^{r_n}\left(\sum_{l=1}^{T_r}\Big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\Big|\right)\mathbb{1}\{\mathcal{A}_{n,r}\} = \begin{cases} \widetilde{O}\left(\sum_{j=1}^{J_n}\sqrt{|\mathcal{I}_{n,j}|}\right), & \text{for the change-point setting;} \\ \widetilde{O}\left(\sqrt{S_n} + (\mathsf{KS}_{n}^{\mathsf{stage}})^{\frac{1}{3}} S_n^{\frac{2}{3}}\right), & \text{for the smooth drift setting.} \end{cases}$$

It remains to see how the per-stage cumulative regret bounds derived above can be leveraged to establish Theorem 3.1. To this end, we cope with the change-point and smooth drift settings separately in what follows.

B.1.3 Controlling the dominant term in the change-point setting

Recall the definition of $\{\mathcal{I}_k\}_{k=1}^{N^{\mathsf{cp}}+1}$ and $\{\mathcal{I}_{n,j}\}_{j=1}^{J_n}$ in Section B.1.1, and denote

$$\mathcal{B}_n = \mathcal{A}_{n,r_n-1} \cap \mathcal{A}_{n,r_n}. \tag{43}$$

Suppose that DriftOCP contains N stages. Applying Lemma B.2 tells us that

$$\begin{aligned} \sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\Big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\Big|\,\mathbb{1}\{\mathcal{A}_{n,r}\} &\le \widetilde{O}\left(\sum_{n=1}^{N}\left(\sum_{j=1}^{J_n}\sqrt{|\mathcal{I}_{n,j}|}\right)\right) \\ &\le \widetilde{O}\left(\sum_{n=1}^{N}\left(\sum_{j=1}^{J_n}\sqrt{|\mathcal{I}_{n,j}|}\right)\mathbb{1}\{\mathcal{B}_n\} + \sum_{n=1}^{N}(\tau_{n+1}-\tau_n)\,\mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\}\right), \end{aligned} \tag{44}$$

where the second relation follows since, for any $n \in [N]$,

$$\sum_{j=1}^{J_n}\sqrt{|\mathcal{I}_{n,j}|} \le \sqrt{J_n\left(\sum_{j=1}^{J_n}|\mathcal{I}_{n,j}|\right)} = \sqrt{J_n(\tau_{n+1}-\tau_n)} \le \tau_{n+1}-\tau_n.$$

In the sequel, we bound the two terms on the right-hand side of (44) separately.

- To bound the first term on the right-hand side of (44), we first show that on the typical events, the final two rounds of any stage do not share exactly the same score distributions. This is formally stated in the lemma below, whose proof is provided in Section B.3.3.

  Lemma B.3. *Consider any stage n in the change-point setting. On the event $\mathcal{A}_{n,r_n-1} \cap \mathcal{A}_{n,r_n}$, the rounds $r_n - 1$ and r_n cannot both be entirely contained within the same time segment from $\{\mathcal{I}_k\}_{k=1}^{N^{\mathsf{cp}}+1}$.*

In view of Lemma B.3, each time segment $\mathcal{I}_k$ cannot overlap with more than two consecutive stages; in other words, each $\mathcal{I}_k$ can contain at most two time segments from $\{\mathcal{I}_{n,j} : n \in [N],\ j \in [J_n]\}$. As a consequence,

$$\sum_{n=1}^{N}\Big(\sum_{j=1}^{J_n}\sqrt{|\mathcal{I}_{n,j}|}\Big)\mathbb{1}\{\mathcal{B}_n\} \le 2\sum_{k=1}^{N^{\mathsf{cp}}+1}\sqrt{|\mathcal{I}_k|} \le 2\sqrt{(N^{\mathsf{cp}}+1)T},$$

where the last step follows from Cauchy-Schwarz as well as the fact that $\sum_{k=1}^{N^{\mathsf{cp}}+1}|\mathcal{I}_k| = T$.

- Turning to the second term on the right-hand side of (44), we make note of the decomposition:

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{N}(\tau_{n+1}-\tau_n)\mathbb{1}\left\{\mathcal{B}_n^{\mathrm{c}}\right\}\right] &\le \sum_{n=1}^{\infty}\mathbb{E}\left[(\tau_{n+1}-\tau_n)\mathbb{1}\left\{\mathcal{B}_n^{\mathrm{c}}\right\}\right] \\
&\le \sum_{n=1}^{\infty}\sum_{i<j}(j-i)\mathbb{P}\left(\tau_n=i,\tau_{n+1}=j,\mathcal{B}_n^{\mathrm{c}}\right) \\
&\le \sum_{n=1}^{\infty}\sum_{i<j}(j-i)\mathbb{P}(\tau_n=i)\mathbb{P}\left(\bigcup_{k=i\vee\frac{j}{16}}^{j}\bigcup_{l=i\vee\frac{j}{16}}^{k-1}\mathcal{A}(l,k)^{\mathrm{c}},\tau_{n+1}=j \;\middle|\; \tau_n=i\right),
\end{aligned} \tag{45}$$

where the last inequality uses a simple property of DriftOCP, namely, $16\tau_{n,r_n-1} \ge \tau_{n+1}$ due to the exponential growth of round lengths. Additionally, observe that under the independent data assumption, for any $l,k \in [i,j]$ the event $\mathcal{A}(l,k)^{\mathrm{c}}$ is independent of what has happened prior to time i, which taken together with (37b) and the union bound gives

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k=i\vee\frac{j}{16}}^{j}\bigcup_{l=i\vee\frac{j}{16}}^{k-1}\mathcal{A}(l,k)^{\mathrm{c}},\tau_{n+1}=j \;\middle|\; \tau_n=i\right) &\le \sum_{k=i\vee\frac{j}{16}}^{j}\sum_{l=i\vee\frac{j}{16}}^{k-1}\mathbb{P}\big(\mathcal{A}(l,k)^{\mathrm{c}} \mid \tau_n=i\big) \\
&= \sum_{k=i\vee\frac{j}{16}}^{j}\sum_{l=i\vee\frac{j}{16}}^{k-1}\mathbb{P}\big(\mathcal{A}(l,k)^{\mathrm{c}}\big) \le \sum_{k=i\vee\frac{j}{16}}^{j}\sum_{l=i\vee\frac{j}{16}}^{k-1}\frac{1}{k^6} \\
&\le \sum_{k=i\vee\frac{j}{16}}^{j}\frac{1}{k^5} < \frac{1}{4\left(i\vee\frac{j}{16}\right)^4}.
\end{aligned}$$

Substituting this bound into (45) results in

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{N}(\tau_{n+1}-\tau_n)\mathbb{1}\left\{\mathcal{B}_n^{\mathrm{c}}\right\}\right] &\le \sum_{n=1}^{\infty}\sum_{i<j}(j-i)\mathbb{P}(\tau_n=i)\frac{1}{4\left(i\vee\frac{j}{16}\right)^4} \\
&\lesssim \sum_{n=1}^{\infty}\sum_{i=1}^{\infty}\mathbb{P}(\tau_n=i)\left\{\sum_{j=i+1}^{16i}\frac{j-i}{i^4}+\sum_{j=16i+1}^{\infty}\frac{j-i}{j^4}\right\} \\
&\lesssim \sum_{n=1}^{\infty}\sum_{i=1}^{\infty}\mathbb{P}(\tau_n=i)\frac{1}{i^2} \overset{\text{(a)}}{\le} \sum_{n=1}^{\infty}\sum_{i=1}^{\infty}\mathbb{P}(\tau_n=i)\frac{1}{n^2} \\
&= \sum_{n=1}^{\infty}\frac{1}{n^2} = O(1),
\end{aligned}$$

where (a) relies on the elementary bound $\tau_n \ge n$.

Putting the preceding bounds together, we arrive at

$$\mathbb{E}\left[\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\left|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\right|\mathbb{1}\{\mathcal{A}_{n,r}\}\right] \le 2\sqrt{(N^{\mathsf{cp}}+1)T} + O(1).$$

The advertised bound in the change-point setting then follows by combining this with (42).

B.1.4 Controlling the dominant term in the smooth drift setting

From Lemma B.2, we have established how to bound the cumulative regret within a single stage. As before, suppose there are N stages in total. Summing over the stage index n then yields the following bound:

$$\begin{aligned}\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\left|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\right|\mathbb{1}\{\mathcal{A}_{n,r}\} &\le \sum_{n=1}^{N}\widetilde{O}\left(\sqrt{S_n} + \left(\mathsf{KS}_n^{\mathsf{stage}}\right)^{\frac{1}{3}}S_n^{\frac{2}{3}}\right)\\ &\overset{\text{(a)}}{\le} \widetilde{O}\left(\sum_{n=1}^{N-1}\sqrt{S_n} + \sqrt{T} + \sum_{n=1}^{N}\left(\mathsf{KS}_n^{\mathsf{stage}}\right)^{\frac{1}{3}}S_n^{\frac{2}{3}}\right)\\ &\overset{\text{(b)}}{\le} \widetilde{O}\left(\sum_{n=1}^{N-1}\sqrt{S_n} + \sqrt{T} + \left(\sum_{n=1}^{N}\mathsf{KS}_n^{\mathsf{stage}}\right)^{\frac{1}{3}}\left(\sum_{n=1}^{N}S_n\right)^{\frac{2}{3}}\right)\\ &= \widetilde{O}\left(\sum_{n=1}^{N-1}\sqrt{S_n} + \sqrt{T} + \left(\mathsf{KS}_T\right)^{\frac{1}{3}}T^{\frac{2}{3}}\right).\end{aligned} \tag{46}$$

Here, (a) holds since $\sqrt{S_N} \le \sqrt{T}$, whereas (b) results from Hölder's inequality.

With the above inequality in mind, a crucial task is to bound $\sum_{n=1}^{N-1}\sqrt{S_n}$, which can be decomposed into

$$\sum_{n=1}^{N-1}\sqrt{S_n} = \sum_{n=1}^{N-1}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n\} + \sum_{n=1}^{N-1}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\} \tag{47}$$

with $\mathcal{B}_n$ denoting the event $\mathcal{B}_n = \mathcal{A}_{n,r_{n-1}} \cap \mathcal{A}_{n,r_n}$ (see (43)). By applying an argument analogous to the one used to control the second term in (44), we can readily obtain

$$\mathbb{E}\left[\sum_{n=1}^{N-1}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\}\right] = O(1), \tag{48}$$

where we omit the details for brevity. Therefore, everything comes down to controlling $\sum_{n=1}^{N-1}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n\}$, which forms the main content of the remainder of this subsection.

For $n \in [N-1]$, note that the last round r_n of each of these stages ends with a restart; that is, a drift detection is declared in round r_n. Following the proof of Lemma B.2, let t_n denote the number of iterations in round r_n, then according to the drift detection subroutine, there exists some $j_n \in [t_n]$ such that

$$\left|\sum_{l=j_n}^{t_n}\left(\mathbb{1}\{s_{n,r_n,l} > q_{n,r_n}\} - \alpha\right)\right| > 24\sqrt{(t_n - j_n + 1)\log(4\tau_{n,r_n})}. \tag{49}$$

On the event $\mathcal{B}_n$, it is observed that

$$\begin{aligned}&\left|\sum_{l=j_n}^{t_n}\left(\mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha\right)\right|\\ &\ge \left|\sum_{l=j_n}^{t_n}\left(\mathbb{1}\{s_{n,r_n,l} > q_{n,r_n}\} - \alpha\right)\right| - \left|\sum_{l=j_n}^{t_n}\left(\mathbb{1}\{s_{n,r_n,l} > q_{n,r_n}\} - \mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n})\right)\right|\\ &> 24\sqrt{(t_n - j_n + 1)\log(4\tau_{n,r_n})} - 6\sqrt{(t_n - j_n + 1)\log\tau_{n+1}} > 18\sqrt{(t_n - j_n + 1)\log\tau_{n+1}},\end{aligned} \tag{50}$$

where the last line arises from (49) and the definition of $\mathcal{A}_{n,r_n}$. Further, let us introduce

$$B_n \coloneqq \frac{1}{t_n - j_n + 1}\sum_{l=j_n}^{t_n}\left(\mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha\right),$$

which, according to Eqn. (50), satisfies

$$\frac{\sqrt{t_n - j_n + 1}}{18\sqrt{\log \tau_{n+1}}} |B_n| \, \mathbb{1}\{\mathcal{B}_n\} \geq \mathbb{1}\{\mathcal{B}_n\}. \tag{51}$$

With this inequality in place, we can readily obtain

$$\begin{aligned}
\sum_{n=1}^{N-1} \sqrt{S_n} \mathbb{1}\{\mathcal{B}_n\} &\overset{(51)}{\leq} \sum_{n=1}^{N-1} \sqrt{S_n} \left(\frac{\sqrt{t_n - j_n + 1}}{18\sqrt{\log \tau_{n+1}}} |B_n| \right)^{\frac{1}{3}} \mathbb{1}\{\mathcal{B}_n\} \\
&\leq \sum_{n=1}^{N-1} S_n^{\frac{2}{3}} |B_n|^{\frac{1}{3}} \mathbb{1}\{\mathcal{B}_n\} \leq \left(\sum_{n=1}^{N-1} S_n \right)^{\frac{2}{3}} \left(\sum_{n=1}^{N-1} |B_n| \, \mathbb{1}\{\mathcal{B}_n\} \right)^{\frac{1}{3}} \leq T^{\frac{2}{3}} \left(\sum_{n=1}^{N-1} |B_n| \, \mathbb{1}\{\mathcal{B}_n\} \right)^{\frac{1}{3}},
\end{aligned} \tag{52}$$

where the last line follows from Hölder's inequality and Jensen's inequality. Thus, it amounts to bounding $\sum_{n=1}^{N-1} |B_n| \, \mathbb{1}\{\mathcal{B}_n\}$, which we accomplish next.

For every B_n, it follows from the triangle inequality that

$$\begin{aligned}
|B_n| &= \left| \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \left(\mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha \right) \right| \\
&\leq \underbrace{\left| \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \left(\mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \frac{1}{T_{r_n - 1}} \sum_{i=1}^{T_{r_n-1}} \mathbb{P}(s'_{n,r_n-1,i} > q_{n,r_n} \mid q_{n,r_n}) \right) \right|}_{=:\mathcal{T}_{n,1}} \\
&\quad + \underbrace{\left| \frac{1}{T_{r_n-1}} \sum_{l=1}^{T_{r_n-1}} \left(\mathbb{P}(s'_{n,r_n-1,l} > q_{n,r_n} \mid q_{n,r_n}) - \mathbb{1}\{s_{n,r_n-1,l} > q_{n,r_n}\} \right) \right|}_{=:\mathcal{T}_{n,2}} \\
&\quad + \underbrace{\left| \frac{1}{T_{r_n-1}} \sum_{l=1}^{T_{r_n-1}} \left(\mathbb{1}\{s_{n,r_n-1,l} > q_{n,r_n}\} - \alpha \right) \right|}_{=:\mathcal{T}_{n,3}},
\end{aligned}$$

where, for each $i = 1, \ldots, T_{r_n-1}$, $s'_{n,r_n-1,i}$ is an independent copy of $s_{n,r_n-1,i}$. The above inequality reduces the problem to controlling three terms.

- Regarding $\mathcal{T}_{n,1}$, one can insert $\mathbb{P}(s_{n,r_n,1} > q_{n,r_n} \mid q_{n,r_n})$ into each summand to derive

$$\begin{aligned}
\mathcal{T}_{n,1} \leq &\left| \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \left(\mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \mathbb{P}(s_{n,r_n,1} > q_{n,r_n} \mid q_{n,r_n}) \right) \right| \\
&+ \left| \frac{1}{T_{r_n-1}} \sum_{l=1}^{T_{r_n-1}} \left(\mathbb{P}(s_{n,r_n,1} > q_{n,r_n} \mid q_{n,r_n}) - \mathbb{P}(s'_{n,r_n-1,l} > q_{n,r_n} \mid q_{n,r_n}) \right) \right|.
\end{aligned} \tag{53}$$

For any $l \in [j_n, t_n]$, we can apply the telescoping technique and the triangle inequality to obtain

$$\begin{aligned}
\left| \mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \mathbb{P}(s_{n,r_n,1} > q_{n,r_n} \mid q_{n,r_n}) \right| &\leq \sup_{q \in \mathbb{R}} \left\{ \left| \mathbb{P}(s_{n,r_n,l} > q) - \mathbb{P}(s_{n,r_n,1} > q) \right| \right\} \\
&\leq \sum_{j=1}^{l-1} \sup_{q \in \mathbb{R}} \left\{ \left| \mathbb{P}(s_{n,r_n,j+1} > q) - \mathbb{P}(s_{n,r_n,j} > q) \right| \right\} \\
&\leq \sum_{j=1}^{l-1} \mathsf{KS}(s_{n,r_n,j}, s_{n,r_n,j+1}) \leq \sum_{j=1}^{t_n-1} \mathsf{KS}(s_{n,r_n,j}, s_{n,r_n,j+1});
\end{aligned}$$

Repeating the same arguments and adopting the notation $s_{n,r_n-1,T_{r_n-1}+1} \coloneqq s_{n,r_n,1}$ also give

$$\mathbb{P}(s_{n,r_n,1} > q_{n,r_n} \mid q_{n,r_n}) - \mathbb{P}(s'_{n,r_n-1,l} > q_{n,r_n} \mid q_{n,r_n}) \leq \sum_{j=1}^{T_{r_n-1}} \mathsf{KS}(s_{n,r_n-1,j}, s_{n,r_n-1,j+1}).$$

Substituting these two results into Eqn. (53) yields

$$\begin{aligned}\mathcal{T}_{n,1} &\leq \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \sum_{j=1}^{t_n-1} \mathsf{KS}(s_{n,r_n,j}, s_{n,r_n,j+1}) + \frac{1}{T_{r_n-1}} \sum_{l=1}^{T_{r_n-1}} \sum_{j=1}^{T_{r_n-1}} \mathsf{KS}(s_{n,r_n-1,j}, s_{n,r_n-1,j+1}) \\ &= \sum_{j=1}^{T_{r_n-1}} \mathsf{KS}(s_{n,r_n-1,j}, s_{n,r_n-1,j+1}) + \sum_{j=1}^{t_n-1} \mathsf{KS}(s_{n,r_n,j}, s_{n,r_n,j+1}),\end{aligned}$$

where we let $r_n - 1$ represent round r_{n-1} of stage $n-1$ if $r_n = 1$. Therefore, when summing over n, each KS distance term is counted at most twice, and as a result,

$$\sum_{n=1}^{N-1} \mathcal{T}_{n,1} \leq 2 \sum_{t=1}^{T-1} \mathsf{KS}(s_t, s_{t+1}) = 2\mathsf{KS}_T.$$

- With regards to $\mathcal{T}_{n,2}$, on the event $\mathcal{B}_n$ (cf. (43)), we can from the definition of $\mathcal{A}_{n,r_n-1}$ (cf. (37)) that

$$\mathcal{T}_{n,2} = \left| \frac{1}{T_{r_n-1}} \sum_{l=1}^{T_{r_n-1}} \left(\mathbb{P}(s'_{n,r_n-1,l} > q_{n,r_n} \mid q_{n,r_n}) - \mathbb{1}\{s_{n,r_n-1,l} > q_{n,r_n}\} \right) \right| \leq 6\sqrt{\frac{\log \tau_{n+1}}{T_{r_n-1}}},$$

which in turn implies that

$$\begin{aligned}\sum_{n=1}^{N-1} \mathcal{T}_{n,2} \mathbb{1}\{\mathcal{B}_n\} &\leq \sum_{n=1}^{N-1} 6\sqrt{\frac{\log \tau_{n+1}}{T_{r_n-1}}} \mathbb{1}\{\mathcal{B}_n\} \\ &\overset{(51)}{\leq} \sum_{n=1}^{N-1} 6\sqrt{\frac{\log \tau_{n+1}}{T_{r_n-1}}} \left(\frac{\sqrt{t_n}}{18\sqrt{\log \tau_{n+1}}} |B_n| \, \mathbb{1}\{\mathcal{B}_n\} \right) \leq \sum_{n=1}^{N-1} \frac{2}{3} |B_n| \, \mathbb{1}\{\mathcal{B}_n\}.\end{aligned}$$

Here, the last inequality follows from the fact that $t_n \leq T_{r_n} \leq 4T_{r_n-1}$ for $r_n > 1$ and $t_n \leq T_{r_n} = 1 \leq T_{r_n-1}$ for $r_n = 1$.

- When it comes to $\mathcal{T}_{n,3}$, it is seen from the definition of $q_{n,r}$—which is chosen to be the α-empirical-quantile of $\{s_{n,r_n-1,l}\}_{l=1}^{T_{r_n-1}}$—that

$$\begin{aligned}\mathcal{T}_{n,3} \mathbb{1}\{\mathcal{B}_n\} &= \left| \frac{1}{T_{r_n-1}} \sum_{l=1}^{T_{r_n-1}} \left(\mathbb{1}\{s_{n,r_n-1,l} > q_{n,r_n}\} - \alpha \right) \right| \mathbb{1}\{\mathcal{B}_n\} \\ &\leq \frac{1}{T_{r_n-1}} \mathbb{1}\{\mathcal{B}_n\} \leq \frac{3\sqrt{\log \tau_{n+1}}}{2\sqrt{t_n - j_n + 1}} \mathbb{1}\{\mathcal{B}_n\} \overset{(51)}{\leq} \frac{1}{12} |B_n| \mathbb{1}\{\mathcal{B}_n\}.\end{aligned}$$

Taking together the preceding bounds on $\mathcal{T}_{n,1}$, $\mathcal{T}_{n,2}$ and $\mathcal{T}_{n,3}$ results in

$$\begin{aligned}\sum_{n=1}^{N-1} |B_n| \, \mathbb{1}\{\mathcal{B}_n\} &\leq \sum_{n=1}^{N-1} (\mathcal{T}_{n,1} + \mathcal{T}_{n,2} + \mathcal{T}_{n,3}) \mathbb{1}\{\mathcal{B}_n\} \\ &\leq 2\mathsf{KS}_T + \frac{2}{3} \sum_{n=1}^{N-1} |B_n| \, \mathbb{1}\{\mathcal{B}_n\} + \frac{1}{12} \sum_{n=1}^{N-1} |B_n| \, \mathbb{1}\{\mathcal{B}_n\} = 2\mathsf{KS}_T + \frac{3}{4} \sum_{n=1}^{N-1} |B_n| \mathbb{1}\{\mathcal{B}_n\},\end{aligned}$$

from which it follows that $\sum_{n=1}^{N-1} |B_n|\,\mathbb{1}\{\mathcal{B}_n\} \le 8\mathsf{KS}_T$. Combine this bound with (52) to arrive at

$$\sum_{n=1}^{N-1} \sqrt{S_n}\mathbb{1}\{\mathcal{B}_n\} \le 2(\mathsf{KS}_T)^{\frac{1}{3}}T^{\frac{2}{3}}. \tag{54}$$

To finish up, taking (46), (48) and (54) collectively yields

$$\begin{aligned}
\mathbb{E}\Big[\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big|\mathbb{1}\{\mathcal{A}_{n,r}\}\Big] &\le \widetilde{O}\bigg(\mathbb{E}\Big[\sum_{n=1}^{N-1}\sqrt{S_n}\Big] + \sqrt{T} + \mathsf{KS}_T^{\frac{1}{3}}T^{\frac{2}{3}}\bigg) \\
&\le \widetilde{O}\bigg(\mathbb{E}\Big[\sum_{n=1}^{N-1}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n\}\Big] + \mathbb{E}\Big[\sum_{n=1}^{N-1}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\}\Big] + \sqrt{T} + (\mathsf{KS}_T)^{\frac{1}{3}}T^{\frac{2}{3}}\bigg) \\
&\overset{(54)}{\le} \widetilde{O}\big(\sqrt{T} + (\mathsf{KS}_T)^{\frac{1}{3}}T^{\frac{2}{3}}\big).
\end{aligned}$$

We can immediately finish the proof for the smooth drift setting by combining the above result with (42).

B.2 Proof of Theorem 3.2

Consider a given T, along with a change-point budget N^{cp} in the change-point setting and a cumulative variation budget KS_T in the smooth drift setting. In what follows, we intend to construct a subclass of distributions $\mathcal{L}'$ and use it to establish the claimed minimax lower bound.

Step 1: construction of a distribution subclass $\mathcal{L}'$. Partition the horizon $[T] \coloneqq \{1, \ldots, T\}$ into m consecutive time segments $\mathcal{I}_1, \ldots, \mathcal{I}_m$ of (nearly) equal size, where

$$\mathcal{I}_j \;\coloneqq\; \big\{(j-1)\lceil T/m\rceil + 1,\; \ldots,\; \min\{j\lceil T/m\rceil, T\}\big\}, \qquad j \in [m].$$

Define $\mathcal{L}'$ as the collection of distribution sequences whose corresponding score distributions $\{\mathcal{D}_t^{\mathsf{score}}\}_{t=1}^T$ obey

1. for each $t \in [T]$, $\mathcal{D}_t^{\mathsf{score}} \in \{\mathrm{Exp}(1), \mathrm{Exp}(1+\varepsilon)\}$, where $\mathrm{Exp}(\beta)$ denotes the exponential distribution with the rate parameter β, and the parameter $\varepsilon \in (0, 1]$ will be specified momentarily;
2. $\mathcal{D}_t^{\mathsf{score}}$ is blockwise constant, namely, for each $j \in [m]$, one has $\mathcal{D}_t^{\mathsf{score}} = \mathcal{D}_{t'}^{\mathsf{score}}$ for all $t, t' \in \mathcal{I}_j$.

Clearly, if $m = N^{\mathsf{cp}} + 1$, then $\mathcal{L}' \subseteq \mathcal{L}_1(N^{\mathsf{cp}})$. We also verify that in the smooth drift setting, $\mathcal{L}' \subseteq \mathcal{L}_2(\mathsf{KS}_T)$ for sufficiently small ε. To see this, we make the observation that

$$\begin{aligned}
\mathsf{KS}\big(\mathrm{Exp}(1), \mathrm{Exp}(1+\varepsilon)\big) = \sup_{x\ge 0}\big\{\big|e^{-x} - e^{-(1+\varepsilon)x}\big|\big\} &= \sup_{x\ge 0}\big\{e^{-x}\big(1 - e^{-\varepsilon x}\big)\big\} \\
&\le \sup_{x\ge 0}\big\{\varepsilon x e^{-x}\big\} \le 2\varepsilon,
\end{aligned}$$

where we have used the elementary inequalities $1 - e^{-u} \le u$ for $u \ge 0$ and $xe^{-x} \le \frac{x}{1+x} \le 1$ for $x \ge 0$. Consequently, for any $\{\mathcal{D}_t^{\mathsf{score}}\}_{t=1}^T \in \mathcal{L}'$, it is easily seen that

$$\sum_{t=1}^{T-1} \mathsf{KS}(\mathcal{D}_t^{\mathsf{score}}, \mathcal{D}_{t+1}^{\mathsf{score}}) \le \sum_{j=1}^{m-1} \mathsf{KS}\big(\mathrm{Exp}(1), \mathrm{Exp}(1+\varepsilon)\big) \le 2\varepsilon m.$$

Choosing $\varepsilon \le \mathsf{KS}_T/(2m)$ ensures that $\sum_{t=1}^{T-1} \mathsf{KS}(\mathcal{D}_t^{\mathsf{score}}, \mathcal{D}_{t+1}^{\mathsf{score}}) \le \mathsf{KS}_T$, and as a result, $\mathcal{L}' \subseteq \mathcal{L}_2(\mathsf{KS}_T)$.

Step 2: lower bound for a single time segment. Consider a fixed time segment and suppress the segment index here for notational simplicity. Within this time segment, the score random variables $\{s_t\}$ are i.i.d. drawn from either Exp(1) or Exp$(1+\varepsilon)$. Denote $s_{1:t} = \{s_1, \ldots, s_t\}$, let $q_t = q(s_{1:t})$ be an arbitrary estimator based on the past observations, and set $q_0^\star \coloneqq \log(1/\alpha)$ and $q_1^\star \coloneqq \frac{1}{1+\varepsilon}\log(1/\alpha)$.

Denote by $\mathcal{D}_{1:t}^{0,\mathsf{score}}$ and $\mathcal{D}_{1:t}^{1,\mathsf{score}}$ the joint distributions of $(s_1, \ldots, s_t)$ when $s_i \sim \text{Exp}(1)$ and $s_i \sim \text{Exp}(1+\varepsilon)$, respectively. Let the Bernoulli random variable $H \sim \text{Ber}(0.5)$ indicate which distribution generates the sequence $\{s_t\}$. Then, letting s be a random variable—independent of $s_{1:t}$ conditioned on H—such that $s \mid H = 0 \sim \text{Exp}(1)$ and $s \mid H = 1 \sim \text{Exp}(1+\varepsilon)$, and denoting by $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) the distribution when $H = 0$ (resp. $H = 1$), we can derive

$$\begin{aligned}
\mathbb{E}_{H, s_{1:t}\sim\mathcal{D}_{1:t}^{H,\mathsf{score}}}\left[\left|\mathbb{P}_H(s > q_t \,|\, q_t) - \alpha\right|\right] &= \frac{1}{2}\mathbb{E}[|\mathbb{P}_0(s > q_t \,|\, q_t) - \alpha|] + \frac{1}{2}\mathbb{E}_{s_{1:t}}[|\mathbb{P}_1(s > q_t \,|\, q_t) - \alpha|] \\
&= \frac{1}{2}\mathbb{E}\left[\left(\left|e^{-q_t} - \alpha\right| + \left|e^{-(1+\varepsilon)q_t} - \alpha\right|\right)\right] \\
&\geq \frac{1}{2}\mathbb{E}\left[\left(\left|e^{-q_t} - \alpha\right| + \left|e^{-(1+\varepsilon)q_t} - \alpha\right|\right)\mathbb{1}\left\{q_t \in \mathcal{K}\right\}\right] + \frac{\alpha}{4}\mathbb{P}\left(q_t \notin \mathcal{K}\right) \\
&= \mathbb{E}_{H, s_{1:t}\sim\mathcal{D}_{1:t}^{H,\mathsf{score}}}\left[\left|\mathbb{P}_H(s > q_t \,|\, q_t) - \alpha\right|\mathbb{1}\{q_t \in \mathcal{K}\}\right] + \frac{\alpha}{4}\mathbb{P}(q_t \notin \mathcal{K}),
\end{aligned} \tag{55}$$

where we take $\mathcal{K} \coloneqq \left[\frac{\log(2/3\alpha)}{1+\varepsilon}, \log\frac{2}{\alpha}\right]$. Here, the inequality above holds due to the elementary fact

$$\left|e^{-q} - \alpha\right| + \left|e^{-(1+\varepsilon)q} - \alpha\right| \geq \frac{\alpha}{2}$$

as long as $q > \log\frac{2}{\alpha}$ or $q < \frac{\log(2/3\alpha)}{1+\varepsilon}$.

Furthermore, for any $\lambda \in \{1, 1+\varepsilon\}$, the mean value theorem tells us the existence of some ξ between q_t and $q_\lambda^\star$ obeying

$$|\mathbb{P}_H(s > q_t \,|\, q_t) - \alpha| = |\mathbb{P}_H(s > q_t \,|\, q_t) - \mathbb{P}_H(s > q_H^\star)| = \lambda e^{-\lambda\xi}\,|q_t - q_H^\star|.$$

Note that when $\varepsilon \leq 1/2$, $1 \leq \lambda \leq 1+\varepsilon$ and $\xi \leq \log(2/\alpha)$, we have $\lambda e^{-\lambda\xi} \geq e^{-\lambda\log\frac{2}{\alpha}} \geq e^{-(1+\varepsilon)\log\frac{2}{\alpha}} = (\alpha/2)^{1+\varepsilon} \geq \alpha^{1+\varepsilon}/3$, and as a consequence,

$$|\mathbb{P}_H(s > q_t \,|\, q_t) - \alpha|\,\mathbb{1}\{q_t \in \mathcal{K}\} \;\geq\; \frac{\alpha^{1+\varepsilon}}{3}\,|q_t - q_H^\star|\,\mathbb{1}\{q_t \in \mathcal{K}\}. \tag{56}$$

Substituting (56) into (55) leads to

$$\begin{aligned}
\mathbb{E}_{H, s_{1:t}\sim\mathcal{D}_{1:t}^{H,\mathsf{score}}}\left[\left|\mathbb{P}_H(s > q_t \,|\, q_t) - \alpha\right|\right] &\geq \frac{\alpha}{4}\mathbb{P}(q_t \notin \mathcal{K}) + \mathbb{E}_{H, s_{1:t}\sim\mathcal{D}_{1:t}^{H,\mathsf{score}}}\left[\left|\mathbb{P}_H(s > q_t \,|\, q_t) - \alpha\right|\mathbb{1}\{q_t \in \mathcal{K}\}\right] \\
&\geq \frac{\alpha^{1+\varepsilon}}{3}\mathbb{E}\left[\left|q_t - q_H^\star\right|\mathbb{1}\{q_t \in \mathcal{K}\}\right] + \frac{\alpha}{4}\mathbb{P}(q_t \notin \mathcal{K}).
\end{aligned} \tag{57}$$

To further bound the second term on the right-hand side of (57), let us look at the following test:

$$\widehat{H} = 0 \;\text{ if } q_t \geq \frac{q_0^\star + q_1^\star}{2}; \quad \widehat{H} = 1 \;\text{ otherwise.}$$

Then given that $H \in \{0, 1\}$, it can be derived that

$$\mathbb{1}\{\widehat{H} \neq H\} \;\leq\; \mathbb{1}\big\{|q_t - q_H^\star| > (q_0^\star - q_1^\star)/2\big\},$$

and hence

$$\begin{aligned}
\mathbb{E}\big[|q_t - q_H^\star|\mathbb{1}\{q_t \in \mathcal{K}\}\big] &\geq \frac{q_0^\star - q_1^\star}{2}\,\mathbb{P}\Big(|q_t - q_H^\star| > \frac{q_0^\star - q_1^\star}{2};\; q_t \in \mathcal{K}\Big) \\
&\geq \frac{q_0^\star - q_1^\star}{2}\,\mathbb{P}(\widehat{H} \neq H;\; q_t \in \mathcal{K}).
\end{aligned} \tag{58}$$

In addition, if $q_0^\star - q_1^\star = \log\frac{1}{\alpha} - \frac{\log(1/\alpha)}{1+\alpha} = \frac{\varepsilon\log(1/\alpha)}{1+\varepsilon} < \frac{3}{4}$, then we have

$$\frac{\alpha}{4}\mathbb{P}(q_t \notin \mathcal{K}) \geq \frac{\alpha^{1+\varepsilon}}{4}\mathbb{P}(q_t \notin \mathcal{K}; \widehat{H} \neq H) \geq \frac{\alpha^{1+\varepsilon}}{3}(q_0^\star - q_1^\star)\mathbb{P}(q_t \notin \mathcal{K}; \widehat{H} \neq H). \tag{59}$$

Substituting (58) and (59) into (57) yields

$$\begin{aligned}\mathbb{E}_{H, s_{1:t}\sim\mathcal{D}_{1:t}^{H,\mathsf{score}}}\left[|\mathbb{P}_H(s > q_t \,|\, q_t) - \alpha|\right] &\geq \frac{\alpha^{1+\varepsilon}}{6}(q_0^\star - q_1^\star)\Big(\mathbb{P}(q_t \in \mathcal{K};\ \widehat{H} \neq H) + \mathbb{P}(q_t \notin \mathcal{K};\ \widehat{H} \neq H)\Big) \\ &= \frac{\alpha^{1+\varepsilon}\varepsilon\log(1/\alpha)}{6(1+\varepsilon)}\mathbb{P}(\widehat{H} \neq H).\end{aligned} \tag{60}$$

It remains to lower bound the probability $\mathbb{P}(\widehat{H} \neq H)$. Le Cam's two-point method (Tsybakov, 2009, Theorem 2.2) and Pinsker's inequality (Tsybakov, 2009, Lemma 2.5) imply that:

$$\begin{aligned}\mathbb{P}(\widehat{H} \neq H) &= \frac{1}{2}\mathbb{P}_0(\widehat{H} \neq 0) + \frac{1}{2}\mathbb{P}_1(\widehat{H} \neq 1) \\ &\geq \frac{1}{2}\Big(1 - \mathsf{TV}(\mathcal{D}_{1:t}^{0,\mathsf{score}}, \mathcal{D}_{1:t}^{1,\mathsf{score}})\Big) \ \geq\ \frac{1}{2}\Big(1 - \sqrt{\tfrac{1}{2}\,\mathsf{KL}\big(\mathcal{D}_{1:t}^{0,\mathsf{score}} \,\|\, \mathcal{D}_{1:t}^{1,\mathsf{score}}\big)}\Big).\end{aligned}$$

Since the observations are i.i.d. within this time segment, one has

$$\mathsf{KL}\big(\mathcal{D}_{1:t}^{0,\mathsf{score}} \,\|\, \mathcal{D}_{1:t}^{1,\mathsf{score}}\big) = t \cdot \mathsf{KL}(\mathrm{Exp}(1) \,\|\, \mathrm{Exp}(1+\varepsilon)).$$

Moreover, the KL divergence admits the following closed-form expression:

$$\mathsf{KL}(\mathrm{Exp}(1) \,\|\, \mathrm{Exp}(1+\varepsilon)) = \log\frac{1}{1+\varepsilon} + (1+\varepsilon) - 1 = \varepsilon - \log(1+\varepsilon),$$

and for $\varepsilon \in (0, 1]$ we have $\varepsilon - \log(1+\varepsilon) \leq \varepsilon^2/2$ since $\log(1+x) \geq x - x^2/2$ for $x \in [0, 1]$. Therefore,

$$\mathsf{KL}\big(\mathcal{D}_{1:t}^{0,\mathsf{score}} \,\|\, \mathcal{D}_{1:t}^{1,\mathsf{score}}\big) \leq t\,\varepsilon^2/2.$$

Choosing $t \leq 1/(4\varepsilon^2)$ yields $\mathsf{KL}(\mathcal{D}_{1:t}^{0,\mathsf{score}} \| \mathcal{D}_{1:t}^{1,\mathsf{score}}) \leq 1/8$, hence $\mathsf{TV}(\mathcal{D}_{1:t}^{0,\mathsf{score}}, \mathcal{D}_{1:t}^{1,\mathsf{score}}) \leq 1/4$. Consequently,

$$\mathbb{P}(\widehat{H} \neq H) = \frac{1}{2}\mathbb{P}_0(\widehat{H} \neq 0) + \frac{1}{2}\mathbb{P}_1(\widehat{H} \neq 1) \ \geq\ \frac{3}{8}. \tag{61}$$

Substituting (61) into (60) then yields, for all $\varepsilon \leq 1/(2\sqrt{t})$,

$$\mathop{\mathbb{E}}_{H\sim\mathrm{Ber}(0.5),\ s_{1:t}\sim\mathcal{D}_{1:t}^{H,\mathsf{score}}}\Big[\big|\mathbb{P}_H(s > q_t \mid q_t) - \alpha\big|\Big] \ \geq\ \frac{\alpha^{1+\varepsilon}\varepsilon\log(1/\alpha)}{6} \cdot \frac{3}{8} = \frac{\alpha^{1+\varepsilon}\varepsilon\log(1/\alpha)}{16}. \tag{62}$$

Step 3: extension from one time segment to entire horizon. We now demonstrate how the single-segment lower bound in Step 2 can be adapted to establish a lower bound for the entire horizon $[T]$. Recall that, at Step 1, $[T]$ is partitioned into m consecutive time segments $\mathcal{I}_1, \ldots, \mathcal{I}_m$ obeying $|\mathcal{I}_j| \asymp T/m$. Let $H_1, \ldots, H_m$ be i.i.d. Bernoulli random variables, and construct a random distribution sequence $\{\mathcal{D}_t^{\mathsf{score}}\}_{t=1}^T$ by setting, for each time segment j and each $t \in \mathcal{I}_j$,

$$\mathcal{D}_t^{\mathsf{score}} = \begin{cases}\mathrm{Exp}(1), & H_j = 0, \\ \mathrm{Exp}(1+\varepsilon), & H_j = 1.\end{cases}$$

Consider any online procedure producing $q_t = q(s_{1:t})$. Condition on the history up to the end of time segment $j-1$. Since H_j is independent of the past, the conditional prior on H_j remains uniform over $\{0, 1\}$, and the observations within time segment j are i.i.d. from $\mathrm{Exp}(1)$ or $\mathrm{Exp}(1+\varepsilon)$ accordingly. Therefore, the single-segment lower bound (62) applies for all times within time segment $\mathcal{I}_j$, with the proviso that $4t\varepsilon^2 \leq 1$.

To ensure that (62) holds throughout the entire time segment $\mathcal{I}_j$, we impose the condition $4\varepsilon^2 |\mathcal{I}_j| \le 1$. Let $\mathcal{F}_{j-1}$ denote the σ-field generated by the samples s_t observed prior to time segment j. We can then take the sum of (62) over $t \in \mathcal{I}_j$ to reach, for each $\mathcal{I}_j$,

$$\mathbb{E}\Big[\sum_{t\in\mathcal{I}_j} \big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\big| \;\Big|\; \mathcal{F}_{j-1}\Big] \;\ge\; |\mathcal{I}_j| \cdot \frac{\varepsilon\alpha^{1+\varepsilon}}{16} \cdot \log\frac{1}{\alpha},$$

and hence, summing over $j = 1, \ldots, m$ and taking expectation over $\{H_j\}_{j=1}^m$ gives

$$\begin{aligned}
\sup_{\{\mathcal{D}_k^{\mathsf{score}}\}_{k=1}^m} \mathbb{E}\left[\sum_{t=1}^T \big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\big|\right] &\ge \mathop{\mathbb{E}}_{H_{1:m}, s_{1:T}}\left[\sum_{t=1}^T \Big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\Big|\right] \\
&= \sum_{k=1}^m \sum_{i\in\mathcal{I}_k} \mathop{\mathbb{E}}_{H_k, s_{\mathcal{I}_k}}\Big[\Big|\mathbb{P}(s_i > q_i \mid q_i) - \alpha\Big|\Big] \\
&\ge \left(\frac{\alpha^{1+\varepsilon}\varepsilon}{16}\log\frac{1}{\alpha}\right)\sum_{k=1}^m |\mathcal{I}_k| \ge \frac{T\varepsilon\alpha^{1+\varepsilon}}{16}\log\frac{1}{\alpha},
\end{aligned} \tag{63}$$

provided that $\varepsilon \le \frac{\sqrt{m}}{2\sqrt{T}}$. To connect this inequality to the advertised minimax lower bounds, we look at the two distribution-shift settings separately.

- *Change-point setting.* In this case, taking $m = N^{\mathsf{cp}} + 1$ and $\varepsilon = \sqrt{(N^{\mathsf{cp}}+1)/(4T)}$ in (63) yields

$$\sup_{\{\mathcal{D}_k^{\mathsf{score}}\}_{k=1}^m} \mathbb{E}\left[\sum_{t=1}^T \big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\big|\right] = \Omega\left(T\varepsilon\,\alpha^{1+\varepsilon}\log\frac{1}{\alpha}\right) = \Omega\left(\alpha^2\log(1/\alpha)\sqrt{(N^{\mathsf{cp}}+1)T}\right).$$

- *Smooth drift setting.* In order for (63) to be applicable in this setting, it suffices to choose ε such that

$$\varepsilon \le \min\Big\{\frac{\mathsf{KS}_T}{2m},\ \sqrt{\frac{m}{4T}}\Big\}.$$

  Now, if $\mathsf{KS}_T\sqrt{T} \ge 1$, then we can choose

$$m = \mathsf{KS}_T^{2/3}T^{1/3}\,(\ge 1), \qquad \varepsilon = \frac{\mathsf{KS}_T^{1/3}}{2T^{1/3}},$$

  which satisfies the above requirement. Plugging this choice into (63) gives

$$\sup_{\{\mathcal{D}_k^{\mathsf{score}}\}_{k=1}^m} \mathbb{E}\left[\sum_{t=1}^T \big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\big|\right] = \Omega\left(T\varepsilon\,\alpha^{1+\varepsilon}\log\frac{1}{\alpha}\right) = \Omega\left(\alpha^2\log(1/\alpha)\,\mathsf{KS}_T^{1/3}T^{2/3}\right).$$

  On the other hand, if $\mathsf{KS}_T\sqrt{T} < 1$, then one can apply (62) to the entire horizon $[T]$ to arrive at

$$\begin{aligned}
&\mathbb{E}_{H\sim\mathrm{Ber}(0.5),\ s_{1:T}\sim\mathcal{D}_{1:T}^{H,\mathsf{score}}}\left[\sum_{t=1}^T \big|\mathbb{P}_H(s > q_t \mid q_t) - \alpha\big|\right] \\
&\qquad = \mathbb{E}_{H\sim\mathrm{Ber}(0.5)}\left[\sum_{t=1}^T \mathbb{E}_{s_{1:t}\sim\mathcal{D}_{1:t}^{H,\mathsf{score}}}\left[\big|\mathbb{P}(s > q_t \mid q_t) - \alpha\big|\right] \;\Big|\; H\right] \\
&\qquad \ge \frac{\alpha^{1+\varepsilon}T\varepsilon\log(1/\alpha)}{6}\cdot\frac{3}{8} = \frac{\alpha^{1+\varepsilon}T\varepsilon\log(1/\alpha)}{16},
\end{aligned}$$

  for all $\varepsilon \le \frac{1}{2\sqrt{T}}$. Thus, taking $\varepsilon = \frac{1}{2\sqrt{T}}$ leads to

$$\sup_{\{\mathcal{D}_k^{\mathsf{score}}\}_{k=1}^m} \mathbb{E}\left[\sum_{t=1}^T \big|\mathbb{P}(s_t > q_t \mid q_t) - \alpha\big|\right] = \Omega\left(\alpha^2\sqrt{T}\log(1/\alpha)\right).$$

  These two cases taken collectively conclude the proof for the smooth drift setting.

B.3 Proof of auxiliary lemmas

B.3.1 Proof of Lemma B.1

Since round r has not been terminated due to the detection of distribution shift, it follows that

$$\left|\sum_{l=i}^{j}\big(\mathbb{1}\{s_{n,r,l} > q_{n,r}\} - \alpha\big)\right| \le \sigma_{n,r}\sqrt{j-i+1} \le 24\sqrt{(j-i+1)\log(4\tau_{n,r})} \tag{64}$$

for every $1 \le i, j \le T_r$. On the event $\mathcal{A}_{n,r}$, combine the definition (37) of $\mathcal{A}_{n,r}$ and Eqn. (64) to reach

$$\left|\sum_{l=i}^{j}\big(\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big)\right| \le 30\sqrt{(j-i+1)\log(4\tau_{n,r})}, \quad \text{for all } 1 \le i < j \le T_r. \tag{65}$$

Below, we look at the two drift settings separately.

Change-point setting. In this setting, the score distribution remains fixed within each time segment $\mathcal{I}_{n,r,k}$. Consequently, for every $l \in \mathcal{I}_{n,r,k}$, the conditional exceedance probability $\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r})$ is identical (i.e., does not depend on l), which in turn implies that

$$\sum_{l\in\mathcal{I}_{n,r,k}}\big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big| = \left|\sum_{l\in\mathcal{I}_{n,r,k}}\big(\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big)\right|.$$

Combining this with Eqn. (65) yields

$$\left(\sum_{l\in\mathcal{I}_{n,r,k}}\big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big|\right)\mathbb{1}\{\mathcal{A}_{n,r}\} \le 30\sqrt{|\mathcal{I}_{n,r,k}|\log(4\tau_{n,r})}.$$

Summing this inequality over $k = 1, \ldots, K_{n,r}$ (i.e., summing over all segments in this round) yields the advertised bound for the change-point setting.

Smooth drift setting. Consider a given $q_{n,r}$. Partition the time indices $1, 2, \ldots, T_r$ into K consecutive time segments $\mathcal{I}_k := \{i_{k-1}+1, \ldots, i_k\}$, $k = 1, \ldots, K$, with $i_0 = 0$ and $i_K = T_r$. This partition is chosen so that, for $l \in \mathcal{I}_1, \mathcal{I}_3, \ldots, \mathcal{I}_{2[\frac{K-1}{2}]+1}$, the quantity $\mathbb{P}(s_{n,r,l} > q_{n,r}) - \alpha$ has the same sign; without loss of generality, assume that the signs are positive. Then for any $l \in \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_{2[\frac{K}{2}]}$, we have $\mathbb{P}(s_{n,r,l} > q_{n,r}) - \alpha < 0$. Consequently, the cumulative regret within this round can be expressed by grouping terms with positive and negative signs as follows:

$$\begin{aligned}
&\sum_{l=1}^{T_r}\big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big| \\
&\quad= \sum_{k:k \text{ is odd}}\sum_{l\in\mathcal{I}_k}\big(\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big) + \sum_{k:k \text{ is even}}\sum_{l\in\mathcal{I}_k}\big(\alpha - \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r})\big) \\
&\quad= \sum_{k=1}^{K}\sum_{l\in\mathcal{I}_k}\big(\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big) + 2\sum_{k:k \text{ is even}}\sum_{l\in\mathcal{I}_k}\big(\alpha - \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r})\big).
\end{aligned} \tag{66}$$

The first term on the right-hand side of (66) can be readily controlled on the event $\mathcal{A}_{n,r}$; more specifically, it is seen from (65) that, on the event $\mathcal{A}_{n,r}$,

$$\sum_{k=1}^{K}\sum_{l\in\mathcal{I}_k}\big(\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big) = \sum_{l=1}^{T_r}\big(\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\big) \le 30\sqrt{T_r\log(4\tau_{n,r})}. \tag{67}$$

We then turn to the second term on the right-hand side of (66). For every $k \in [K]$, set

$$A_k \coloneqq \frac{1}{|\mathcal{I}_k|} \sum_{l \in \mathcal{I}_k} \big(\alpha - \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r})\big).$$

On the event $\mathcal{A}_{n,r}$, it agains follows from (65) that

$$\frac{\sqrt{|\mathcal{I}_k|}}{30\sqrt{\log(4\tau_{n,r})}} A_k = \frac{\sqrt{|\mathcal{I}_k|}}{30\sqrt{\log(4\tau_{n,r})}} \left(\frac{1}{|\mathcal{I}_k|} \sum_{l \in \mathcal{I}_k} \big(\alpha - \mathbb{P}(s_{n,r,l} > q_{n,r}) \mid q_{n,r}\big) \right) \le 1. \tag{68}$$

This allows one to deduce that, on the event $\mathcal{A}_{n,r}$,

$$\begin{aligned} \sum_{k:k \text{ is even}} \sum_{l \in \mathcal{I}_k} \big(\alpha - \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r})\big) &= \sum_{k:k \text{ is even}} |\mathcal{I}_k| \cdot A_k = \sum_{k:k \text{ is even}} |\mathcal{I}_k|^{\frac{2}{3}} \left(|\mathcal{I}_k|^{\frac{1}{2}} A_k \right)^{\frac{2}{3}} A_k^{\frac{1}{3}} \\ &\overset{(68)}{\le} \sum_{k:k \text{ is even}} 30\, |\mathcal{I}_k|^{\frac{2}{3}} A_k^{\frac{1}{3}} \sqrt{\log T} \le 30\sqrt{\log T} \Big(\sum_{k:k \text{ is even}} |\mathcal{I}_k| \Big)^{\frac{2}{3}} \Big(\sum_{k:k \text{ is even}} A_k \Big)^{\frac{1}{3}}, \end{aligned} \tag{69}$$

where the last inequality follows from Hölder's inequality. This leaves us with two sums to control.

- Regarding the summation of $|\mathcal{I}_k|$, it is easily seen that

$$\sum_{k:k \text{ is even}} |\mathcal{I}_k| \le \sum_{k=1}^{K} |\mathcal{I}_k| = T_r. \tag{70}$$

- Let us now turn to the summation of A_k. Given the way we partition the sets $\mathcal{I}_k$, we see that for any even number k (≥ 2), $\mathbb{P}\big(s_{n,r,i_{k-1}} > q_{n,r}\big) > \alpha$. Consequently, for any even k, one can bound A_k as

$$\begin{aligned} A_k &= \frac{1}{|\mathcal{I}_k|} \sum_{l \in \mathcal{I}_k} \big(\alpha - \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r})\big) \\ &\le \frac{1}{|\mathcal{I}_k|} \sum_{l \in \mathcal{I}_k} \big(\mathbb{P}\big(s_{n,r,i_{k-1}} > q_{n,r} \mid q_{n,r}\big) - \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r})\big) = \frac{1}{|\mathcal{I}_k|} \sum_{l \in \mathcal{I}_k} \sum_{i=i_{k-1}}^{l-1} \Delta_{n,r,i}, \end{aligned} \tag{71}$$

  where we define

$$\Delta_{n,r,i} \coloneqq \mathbb{P}\big(s_{n,r,i} > q_{n,r} \mid q_{n,r}\big) - \mathbb{P}\big(s_{n,r,i+1} > q_{n,r} \mid q_{n,r}\big).$$

  The definition (6) of the KS distance tells us that

$$\Delta_{n,r,i} \le \mathsf{KS}\big(s_{n,r,i}, s_{n,r,i+1}\big) \quad \text{for all } i \in [T_r],$$

  which combined with Eqn. (71) leads to

$$\begin{aligned} A_k &\le \frac{1}{|\mathcal{I}_k|} \sum_{l \in \mathcal{I}_k} \sum_{i=i_{k-1}}^{l-1} \mathsf{KS}\big(s_{n,r,i}, s_{n,r,i+1}\big) \\ &\le \frac{1}{|\mathcal{I}_k|} \sum_{l \in \mathcal{I}_k} \sum_{i=i_{k-1}}^{i_k-1} \mathsf{KS}\big(s_{n,r,i}, s_{n,r,i+1}\big) = \sum_{i=i_{k-1}}^{i_k-1} \mathsf{KS}\big(s_{n,r,i}, s_{n,r,i+1}\big). \end{aligned} \tag{72}$$

  Let $s_{n,r,0} = s_{n,r,1}$. Summing over all even k yields

$$\sum_{k:k \text{ is even}} A_k \le \sum_{k=1}^{K} \sum_{i=i_{k-1}}^{i_k-1} \mathsf{KS}\big(s_{n,r,i}, s_{n,r,i+1}\big) = \sum_{i=0}^{T_r} \mathsf{KS}\big(s_{n,r,i}, s_{n,r,i+1}\big). \tag{73}$$

Putting Eqns. (70) and (73) together yields

$$\left(\sum_{k:k \text{ is even}} |\mathcal{I}_k|\right)^{\frac{2}{3}} \left(\sum_{k:k \text{ is even}} A_k\right)^{\frac{1}{3}} \leq T_r^{\frac{2}{3}} \left(\sum_{i=1}^{T_r} \mathsf{KS}\big(s_{n,r,i}, s_{n,r,i+1}\big)\right)^{\frac{1}{3}} = \big(\mathsf{KS}_{n,r}^{\mathsf{round}}\big)^{\frac{1}{3}} T_r^{\frac{2}{3}},$$

which taken together with Eqns. (66), (67) and (69) establishes that

$$\sum_{l=1}^{T_r} \Big|\mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha\Big| \cdot \mathbb{1}\{\mathcal{A}_{n,r}\} \leq 30\sqrt{\log T}\left(\sqrt{T_r} + \big(\mathsf{KS}_{n,r}^{\mathsf{round}}\big)^{\frac{1}{3}} T_r^{\frac{2}{3}}\right) = \widetilde{O}\left(\sqrt{T_r} + \big(\mathsf{KS}_{n,r}^{\mathsf{round}}\big)^{\frac{1}{3}} T_r^{\frac{2}{3}}\right).$$

B.3.2 Proof of Lemma B.2

Consider stage n, and denote by t_n the number of iterations in the r_n-th round (recall that r_n denotes the index of the last round of stage n). Observe that for any $r \in [r_n - 1]$ (resp. for $r = r_n$), no initiation of a new stage—i.e., no detection of distribution drift—is triggered within iterations $\{1, \ldots, T_r\}$ (resp. $\{1, \ldots, t_n - 1\}$). In what follows, we look at the two drift settings separately.

Change-point setting. Note that the collection of time segments $\{\mathcal{I}_{n,r,k}\}$ described in Section B.1.1 can be viewed as a refinement of $\{\mathcal{I}_{n,j}\}_{j=1}^{J_n}$; for instance, a given $\mathcal{I}_{n,j}$ might appear in more than one round, possibly due to imperfect drift detection. For each $j \in [J_n]$, denote by $r^{(j)}$ the index of the first round that overlaps with the segment $\mathcal{I}_{n,j}$; it is straightforward to see that the last round that overlaps with $\mathcal{I}_{n,j}$ cannot exceed $r^{(j+1)}$ (here, if this is already the last time segment in stage n, we can simply let $r^{(j+1)}$ be the last round of this stage). Note that $\mathcal{I}_{n,j}$ may share its first and/or last round with $\mathcal{I}_{n,j-1}$ or $\mathcal{I}_{n,j+1}$. Therefore, we have

$$\mathcal{I}_{n,j} \subseteq \left\{\bigcup_{r=r^{(j)}+1}^{r^{(j+1)}-1} \bigcup_{k=1}^{K_{n,r}} \mathcal{I}_{n,r,k}\right\} \cup \left\{\mathcal{I}_{n,r^{(j)},K_{n,r^{(j)}}}\right\} \cup \left\{\mathcal{I}_{n,r^{(j+1)},1}\right\}, \tag{74}$$

which in turn gives

$$\sum_{r=1}^{r_n} \sum_{k=1}^{K_{n,r}} \sqrt{|\mathcal{I}_{n,r,k}|} \leq \sum_{j=1}^{J_n} \left\{\sum_{r=r^{(j)}+1}^{r^{(j+1)}-1} \sum_{k=1}^{K_{n,r}} \sqrt{|\mathcal{I}_{n,r,k}|} + \sqrt{\big|\mathcal{I}_{n,r^{(j)},K_{n,r^{(j)}}}\big|} + \sqrt{\big|\mathcal{I}_{n,r^{(j+1)},1}\big|}\right\}. \tag{75}$$

Recognizing that each of the intermediate rounds $r^{(j)} + 1, \ldots, r^{(j+1)} - 1$ is fully contained within $\mathcal{I}_{n,j}$, we see that, by construction, each of these rounds also contains a single time segment from the collection $\{\mathcal{I}_{n,r,k}\}$. This means that for each $r \in \{r^{(j)} + 1, \ldots, r^{(j+1)} - 1\}$, we have

$$\sum_{k=1}^{K_{n,r}} \sqrt{|\mathcal{I}_{n,r,k}|} = \sqrt{|\mathcal{I}_{n,r,1}|} \leq \sqrt{T_r}.$$

Consequently, for each $j \in [J_n]$, we can bound

$$\sum_{r=r^{(j)}+1}^{r^{(j+1)}-1} \sum_{k=1}^{K_{n,r}} \sqrt{|\mathcal{I}_{n,r,k}|} + \sqrt{\big|\mathcal{I}_{n,r^{(j)},K_{n,r^{(j)}}}\big|} + \sqrt{\big|\mathcal{I}_{n,r^{(j+1)},1}\big|} \leq \sum_{r=r^{(j)}+1}^{r^{(j+1)}-1} \sqrt{T_r} + 2\sqrt{|\mathcal{I}_{n,j}|}, \tag{76}$$

where the last inequality holds since $\mathcal{I}_{n,r^{(j)},K_{n,r^{(j)}}} \cup \mathcal{I}_{n,r^{(j+1)},1} \subseteq \mathcal{I}_{n,j}$.

Next, we bound the summation of $\sqrt{T_r}$ on the right-hand side of (76). If $r^{(j)} + 1 > r^{(j+1)} - 1$, then this summation term is equal to 0; otherwise, it can be seen that (by construction)

$$T_{r^{(j+1)}-1} = \big|\mathcal{I}_{n,r^{(j+1)}-1,1}\big| \leq |\mathcal{I}_{n,j}|,$$

which implies that

$$\sum_{r=r^{(j)}+1}^{r^{(j+1)-1}} \sqrt{T_r} \leq \sum_{r=1}^{r^{(j+1)-1}} \sqrt{T_r} \overset{\text{(a)}}{=} \sum_{r=1}^{r^{(j+1)-1}} 3^{\frac{r}{2}} \leq 3 \cdot 3^{\frac{r^{(j+1)}-1}{2}} \overset{\text{(b)}}{=} 3\sqrt{T_{r^{(j+1)}-1}} \leq 3\sqrt{|\mathcal{I}_{n,j}|}. \tag{77}$$

Here, (a) and (b) are valid due to our choice $T_r = 3^r$ and the fact that, except for the last round of this stage, the r-th round has time length exactly equal to T_r.

Invoking Lemma B.1 as well as (39), we can demonstrate that

$$\begin{aligned}
&\sum_{r=1}^{r_n} \left\{ \sum_{l=1}^{T_r} \left| \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha \right| \right\} \mathbb{1}\{\mathcal{A}_{n,r}\} \\
&\leq \sum_{r=1}^{r_n-1} \left\{ \sum_{l=1}^{T_r} \left| \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha \right| \right\} \mathbb{1}\{\mathcal{A}_{n,r}\} + \left\{ \sum_{l=1}^{t_n-1} \left| \mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha \right| \right\} \mathbb{1}\{\mathcal{A}_{n,r_n}\} + 1 \\
&\leq \widetilde{O}\Big(\sum_{r=1}^{r_n} \sum_{k=1}^{K_{n,r}} \sqrt{|\mathcal{I}_{n,r,k}|} \Big) + 1 \leq \widetilde{O}\Big(\sum_{j=1}^{J_n} \sqrt{|\mathcal{I}_{n,j}|} \Big),
\end{aligned}$$

where the last line follows from Lemma B.1 and the fact that no distribution shift has been detected before the last iteration of stage n. This taken collectively with (75)-(77) then yields

$$\sum_{r=1}^{r_n} \left\{ \sum_{l=1}^{T_r} \left| \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha \right| \right\} \mathbb{1}\{\mathcal{A}_{n,r}\} \leq \widetilde{O}\Big(\sum_{r=1}^{r_n} \sum_{k=1}^{K_{n,r}} \sqrt{|\mathcal{I}_{n,r,k}|} \Big) \leq \widetilde{O}\Big(\sum_{j=1}^{J_n} \sqrt{|\mathcal{I}_{n,j}|} \Big)$$

as claimed.

Smooth drift setting. To begin with, Lemma B.1 tells us that

$$\left\{ \sum_{l=1}^{T_r} \left| \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha \right| \mathbb{1} \right\} \{\mathcal{A}_{n,r}\} = \widetilde{O}\left(\sqrt{T_r} + \left(\mathsf{KS}_{n,r}^{\mathsf{round}}\right)^{\frac{1}{3}} T_r^{\frac{2}{3}} \right), \quad r = 1, 2, \ldots, r_n - 1 \tag{78a}$$

and

$$\begin{aligned}
\sum_{l=1}^{t_n} \left| \mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha \right| \mathbb{1}\{\mathcal{A}_{n,r_n}\} &\leq \sum_{l=1}^{t_n-1} \left| \mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha \right| \mathbb{1}\{\mathcal{A}_{n,r_n}\} + 1 \\
&= \widetilde{O}\left(\sqrt{t_n} + \left(\mathsf{KS}_{n,r_n}^{\mathsf{round}}\right)^{\frac{1}{3}} t_n^{\frac{2}{3}} \right).
\end{aligned} \tag{78b}$$

Sum Eqn. (78) over all rounds in this stage to arrive at:

$$\begin{aligned}
&\sum_{r=1}^{r_n-1} \sum_{l=1}^{T_r} \left| \mathbb{P}(s_{n,r,l} > q_{n,r} \mid q_{n,r}) - \alpha \right| \mathbb{1}\{\mathcal{A}_{n,r}\} + \sum_{l=1}^{t_n} \left| \mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha \right| \mathbb{1}\{\mathcal{A}_{n,r_n}\} \\
&\quad = \sum_{r=1}^{r_n-1} \widetilde{O}\left(\sqrt{T_r} + \left(\mathsf{KS}_{n,r}^{\mathsf{round}}\right)^{\frac{1}{3}} T_r^{\frac{2}{3}} \right) + \widetilde{O}\left(\sqrt{t_n} + \left(\mathsf{KS}_{n,r_n}^{\mathsf{round}}\right)^{\frac{1}{3}} t_n^{\frac{2}{3}} \right) \\
&\quad \overset{\text{(a)}}{\leq} \widetilde{O}\left(\sum_{r=1}^{r_n} 3^{\frac{r}{2}} + \sum_{r=1}^{r_n-1} \left(\mathsf{KS}_{n,r}^{\mathsf{round}}\right)^{\frac{1}{3}} T_r^{\frac{2}{3}} + \left(\mathsf{KS}_{n,r_n}^{\mathsf{round}}\right)^{\frac{1}{3}} t_n^{\frac{2}{3}} \right) \\
&\quad \overset{\text{(b)}}{\leq} \widetilde{O}\left(3^{\frac{r_n}{2}} + \left(\sum_{r=1}^{r_n} \left(\mathsf{KS}_{n,r}^{\mathsf{round}}\right) \right)^{\frac{1}{3}} S_n^{\frac{2}{3}} \right) \overset{\text{(c)}}{\leq} \widetilde{O}\left(\sqrt{S_n} + \mathsf{KS}_{n,S_n}^{\frac{1}{3}} S_n^{\frac{2}{3}} \right).
\end{aligned}$$

Here, (a) holds since $T_r = 3^r$ and $t_n \le 3^{r_n}$, (b) follows from Hölder's inequality, and (c) holds since the total number S_n of time points within stage n satisfies

$$\sqrt{S_n} \ge \left(\sum_{r=1}^{r_n-1} 3^r\right)^{1/2} \asymp 3^{\frac{r_n}{2}}.$$

This taken together with (39) concludes the proof.

B.3.3 Proof of Lemma B.3

We establish this lemma by contradiction. Suppose that round $r_n - 1$ and round r_n are completely contained within the same time segment from the collection $\{\mathcal{I}_k\}_{k=1}^{N^{\mathsf{cp}}+1}$. According to the procedure of DRIFTOCP, on the event $\mathcal{A}_{n,r_n-1} \cap \mathcal{A}_{n,r_n}$ there exists an index j_n such that

$$\frac{1}{T_{r_n-1}}\left|\sum_{l=1}^{T_{r_n-1}}\Big(\mathbb{1}\{s_{n,r_n-1,l} > q_{n,r_n}\} - \alpha\Big)\right| \le \frac{1}{T_{r_n-1}}, \tag{79a}$$

$$\frac{1}{t_n - j_n + 1}\left|\sum_{l=j_n}^{t_n}\Big(\mathbb{1}\{s_{n,r_n,l} > q_{n,r_n}\} - \alpha\Big)\right| > \frac{24\sqrt{\log(4\tau_{n,r_n})}}{\sqrt{t_n - j_n + 1}}, \tag{79b}$$

where (79a) is valid since q_{n,r_n} is taken to be the α-empirical-quantile of the set $\{s_{n,r_n-1,l}\}_{l=1}^{T_{r_n-1}}$, and in (79b) we use the detection threshold $\sigma_{n,r} = 24\sqrt{\log(4\tau_{n,r})}$.

Since the two rounds lie within the same time segment from the collection $\{\mathcal{I}_k\}_{k=1}^{N^{\mathsf{cp}}+1}$, the scores in these two rounds are identically distributed. Let s denote an independent copy of the score from this segment. Then, on $\mathcal{A}_{n,r_n-1} \cap \mathcal{A}_{n,r_n}$ (cf. (37)), we can invoke (79a) and the triangle inequality to obtain

$$\begin{aligned}
\Big|\mathbb{P}(s > q_{n,r_n} \mid q_{n,r_n}) - \alpha\Big| &= \frac{1}{T_{r_n-1}}\left|\sum_{l=1}^{T_{r_n-1}}\Big(\mathbb{P}(s_{n,r_n-1,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha\Big)\right| \\
&\le \frac{1}{T_{r_n-1}}\left|\sum_{l=1}^{T_{r_n-1}}\Big(\mathbb{1}\{s > q_{n,r_n}\} - \alpha\Big)\right| + \frac{1}{T_{r_n-1}}\left|\sum_{l=1}^{T_{r_n-1}}\Big(\mathbb{1}\{s_{n,r_n-1,l} > q_{n,r_n}\} - \mathbb{P}(s > q_{n,r_n} \mid q_{n,r_n})\Big)\right| \\
&\le \frac{1}{T_{r_n-1}} + \frac{6\sqrt{\log\tau_{n+1}}}{\sqrt{T_{r_n-1}}} \le \frac{7\sqrt{\log\tau_{n+1}}}{\sqrt{T_{r_{n-1}}}} \le \frac{14\sqrt{\log\tau_{n+1}}}{\sqrt{T_{r_n}}}.
\end{aligned} \tag{80}$$

In the meantime, applying (79b) again on the same event and invoking the triangle inequality gives

$$\begin{aligned}
\Big|\mathbb{P}(s > q_{n,r_n} \mid q_{n,r_n}) - \alpha\Big| &= \frac{1}{t_n - j_n + 1}\left|\sum_{l=j_n}^{t_n}\Big(\mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n}) - \alpha\Big)\right| \\
&\ge \frac{1}{t_n - j_n + 1}\left|\sum_{l=j_n}^{t_n}\Big(\mathbb{1}\{s_{n,r_n,l} > q_{n,r_n}\} - \alpha\Big)\right| \\
&\quad - \frac{1}{t_n - j_n + 1}\left|\sum_{l=j_n}^{t_n}\Big(\mathbb{1}\{s_{n,r_n,l} > q_{n,r_n}\} - \mathbb{P}(s_{n,r_n,l} > q_{n,r_n} \mid q_{n,r_n})\Big)\right| \\
&\overset{\text{(a)}}{\ge} \frac{24\sqrt{\log(4\tau_{n,r_n})}}{\sqrt{t_n - j_n + 1}} - \frac{6\sqrt{\log\tau_{n+1}}}{\sqrt{t_n - j_n + 1}} \ge \frac{18\sqrt{\log\tau_{n+1}}}{\sqrt{T_{r_n}}},
\end{aligned} \tag{81}$$

where (a) follows by combining (79b) with the event $\mathcal{A}_{n,r_n}$. However, (81) contradicts (80), which in turn completes the proof.

C Detailed proofs in Section 4

This section is devoted to establishing the main results in Section 4. Throughout this section, we define, for any cumulative distribution function (CDF) F, the quantile function

$$\mathsf{Q}_{1-\alpha}(F) \coloneqq \inf\{x \in \mathbb{R} : F(x) \geq 1-\alpha\}. \tag{82}$$

Also, we denote by $\mathcal{C}(\cdot \mid \mathcal{S}^{\mathsf{cal}}, \mathcal{S}^{\mathsf{train}})$ the prediction-set mapping constructed via (22), where $\mathcal{S}^{\mathsf{train}}$ is used to fit the model and $\mathcal{S}^{\mathsf{cal}}$ is used to form the quantile.

C.1 Proof of Proposition 4.1

We first present the proof of Proposition 4.1, which concerns the training-conditional coverage guarantees for standard full conformal methods. Before embarking on the proof, we introduce the following convenient notation (when there is no ambiguity), which shall be used repeatedly throughout Section C.1.

Definition C.1 (Basic notation)**.** *We introduce the following notation, all conditioned on the realization $Z^{\mathsf{train}}_{m+1:n} = z^{\mathsf{train}}_{m+1:n}$ (i.e., the portion of the training set that is disjoint from the calibration set).*

- *For any dataset $\mathcal{S} \subseteq \mathcal{X} \times \mathbb{R}$, let $\widehat{\mu}_{\mathcal{S}}(\cdot) \coloneqq \mathcal{A}\big(\mathcal{S} \cup z^{\mathsf{train}}_{m+1:n}\big)$ be the fitted model trained obtained by algorithm $\mathcal{A}$ on $\mathcal{S} \cup z^{\mathsf{train}}_{m+1:n}$.*
- *For any dataset $\mathcal{S} \subseteq \mathcal{X} \times \mathbb{R}$ and any $Z = (X, Y)$, let $\widehat{\mu}^{Z}_{\mathcal{S}}(\cdot) \coloneqq \widehat{\mu}_{\mathcal{S} \cup \{Z\}}(\cdot)$ be the fitted model trained obtained by algorithm $\mathcal{A}$ on $\mathcal{S} \cup \{(X, Y)\} \cup z^{\mathsf{train}}_{m+1:n}$.*
- *For any dataset $\mathcal{S} \subseteq \mathcal{X} \times \mathbb{R}$ and any $Z = (X, Y)$, $Z' = (X', Y')$, define the scores*

$$s_{\mathcal{S}}(Z') \coloneqq \big|Y' - \widehat{\mu}_{\mathcal{S}}(X')\big| \qquad \text{and} \qquad s^{Z}_{\mathcal{S}}(Z') \coloneqq \big|Y' - \widehat{\mu}^{Z}_{\mathcal{S}}(X')\big|. \tag{83}$$

Remark C.1. *Note that we introduce $\widehat{\mu}^{Z}_{\mathcal{S}}(\cdot)$ in addition to $\widehat{\mu}_{\mathcal{S}}(\cdot)$. This is because, in the full conformal algorithm, the target sample Z is used for both model fitting and construction of the calibration quantile. To emphasize this role and distinguish it from the pretrained-score setting, we adopt a separate notation.*

C.1.1 Key lemmas

We first single out two key lemmas that play a pivotal role in the proof of Proposition 4.1. Here and throughout, we take $L = L_1 L_2$.

The first lemma characterizes the discrepancy between the tail distribution of the scores conditional on a random calibration set and the corresponding tail distribution obtained after averaging over the randomness of the calibration set, provided that a stable learning algorithm is used for model fitting. The proof is deferred to Section C.3.1.

Lemma C.1. *Consider the same setting as in Proposition 4.1. Let $\widehat{\mu}^{(X,Y)}(\cdot)$ denote a fitted model trained on the data $Z^{\mathsf{train}}_{1:n}$ together with the target sample $Z = (X, Y)$, and assume that $\widehat{\mu}^{(X,Y)}(\cdot)$ satisfies Assumption 4.3 with coefficient L_2. Further, suppose that for each $i \in [m]$, $Z^{\mathsf{cal}}_i = (X^{\mathsf{cal}}_i, Y^{\mathsf{cal}}_i)$ is independently drawn from $\mathcal{D}_i$, and let the target pair $Z = (X, Y) \sim \mathcal{D}$. Then, for any $\delta \in (0, 1)$ and conditional on any given realization $Z^{\mathsf{train}}_{m+1:n} = z^{\mathsf{train}}_{m+1:n}$, we have*

$$\begin{aligned}\sup_{x \in \mathbb{R}} \left\{ \left| \mathbb{P}_{\mathcal{D}} \left(\big|Y - \widehat{\mu}^{(X,Y)}(X)\big| > x \,\middle|\, Z^{\mathsf{cal}}_{1:m}, Z^{\mathsf{train}}_{m+1:n} = z^{\mathsf{train}}_{m+1:n} \right) - \mathbb{P}_{\mathcal{D}_{1:m} \times \mathcal{D}} \left(\big|Y - \widehat{\mu}^{(X,Y)}(X)\big| > x \right) \right| \right\} \\ \leq \frac{16L}{n} \sqrt{m \log \frac{1}{\delta}}\end{aligned} \tag{84}$$

with probability at least $1 - \delta$. Here, we adopt the notation

$$\mathbb{P}_{\mathcal{D}_{1:m} \times \mathcal{D}} \left(\big|Y - \widehat{\mu}^{(X,Y)}(X)\big| > x \right) \coloneqq \mathbb{E}_{Z^{\mathsf{cal}}_{1:m}} \left[\mathbb{P}_{\mathcal{D}} \left(\big|Y - \widehat{\mu}^{(X,Y)}(X)\big| > x \,\middle|\, Z^{\mathsf{cal}}_{1:m}, Z^{\mathsf{train}}_{m+1:n} = z^{\mathsf{train}}_{m+1:n} \right) \right].$$

Another key lemma establishes a high-probability upper bound on the deviation of the average empirical scores from their mean over a given time window. In contrast to the pretrained-score setting, full conformal methods induce complicated statistical dependency among the scores $\{s_i\}$, leading to additional technical difficulties. The proof of this lemma is provided in Section C.3.2.

Lemma C.2. *Consider the setting same in Proposition 4.1. Let $\widehat{\mu}(\cdot)$ represent a fitted model trained on $Z_{1:n}^{\mathsf{train}}$ satisfying Assumption 4.3 with coefficient L_2. Further, suppose that for each $i \in [m]$, $Z_i^{\mathsf{cal}} = (X_i^{\mathsf{cal}}, Y_i^{\mathsf{cal}})$ is independently drawn from $\mathcal{D}_i$. For every $i = 1, \ldots, m$, we let*

$$s_i := \left| Y_i^{\mathsf{cal}} - \widehat{\mu}(X_i^{\mathsf{cal}}) \right|, \qquad i = 1, \ldots, m.$$

Then, for any $\delta \in (0,1)$ and conditional on any realization $Z_{m+1:n}^{\mathsf{train}} = z_{m+1:n}^{\mathsf{train}}$, the following event

$$\sup_{x \in \mathbb{R}} \left\{ \frac{1}{m} \left| \sum_{i=1}^{m} \Big(\mathbb{1}\{s_i \le x\} - \mathbb{P}_{\mathcal{D}_{1:m}}(s_i \le x) \Big) \right| \right\} \le 24 \sqrt{\frac{\log(10/\delta)}{m}} + \frac{24L}{n} \sqrt{m \log\left(\frac{10m}{\delta} + n \right)}$$

happens with probability at least $1 - \delta$.

C.1.2 Proof of Proposition 4.1

For notational convenience, we write $Z_{1:m}$ in place of $Z_{1:m}^{\mathsf{cal}}$ throughout this proof when it is clear from the context. Fix an auxiliary sample $z_0 = (x_0, y_0)$, which shall be treated as deterministic in the following. We introduce several addition notation:

- $s_i := s_{Z_{1:m} \cup \{z_0\}}(Z_i) = \left| Y_i - \widehat{\mu}_{Z_{1:m} \cup \{z_0\}}(X_i) \right|$ for $i = 1, \ldots, m$, and $s_{\mathsf{test}} := \left| Y - \widehat{\mu}_{Z_{1:m} \cup \{z_0\}}(X) \right|$, where (X, Y) is not used for model fitting;
- $s_i^{(X,Y)} := s_{Z_{1:m}}^{Z}(Z_i) = \left| Y_i - \widehat{\mu}_{Z_{1:m}}^{(X,Y)}(X_i) \right|$ for $i = 1, \ldots, m$, and $s_{\mathsf{test}}^{(X,Y)} := \left| Y - \widehat{\mu}_{Z_{1:m}}^{(X,Y)}(X) \right|$, where (X, Y) is used for for model fitting;
- $\widehat{\mathsf{Q}}_{1-\alpha} := \mathsf{Q}_{1-\alpha}\left(\frac{1}{m+1} \Big\{ \delta\{s_{\mathsf{test}}^{(X,Y)}\} + \sum_{i=1}^{m} \delta\{s_i^{(X,Y)}\} \Big\} \right)$, which indicates the quantile when (X, Y) is also used for model fitting;
- $\widetilde{\mathsf{Q}}_{1-\alpha} := \mathsf{Q}_{1-\alpha}\left(\frac{1}{m+1} \Big\{ \delta\{s_{\mathsf{test}}^{(X,Y)}\} + \sum_{i=1}^{m} \delta\{s_i\} \Big\} \right)$; note that except for $s_{\mathsf{test}}^{(X,Y)}$, the remaining scores $\{s_i\}_{i=1}^m$ are computed when (X, Y) is not used for training;
- $F_{\mathsf{test}}(u; z_{1:m}) := \mathbb{P}_{(X,Y) \sim \mathcal{D}}\Big(\left| Y - \widehat{\mu}_{z_{1:m}}^{(X,Y)}(X) \right| \le u \Big)$, $F_{\mathsf{test}}(u) := \mathbb{E}_{Z_{1:m} \sim \mathcal{D}_{1:m}}[F_{\mathsf{test}}(u; Z_{1:m})]$;
- $F_i(u) := \mathbb{P}_{\mathcal{D} \times \mathcal{D}_{1:m}}\big(s_i^{(X,Y)} \le u\big)$ and $F_i^0(u) := \mathbb{P}_{\mathcal{D}_{1:m}}\big(s_i \le u\big)$ for $i = 1, \ldots, m$.

Step 1: eliminating the dependence of the fitted model on (X, Y). The first step of the proof is to examine the effect of removing the dependence of the fitted model $\widehat{\mu}^{(X,Y)}(\cdot)$ on the target sample (X, Y). To be precise, consider the discrepancy between

$$\mathbb{P}_{\mathcal{D}}(Y \in \mathcal{C}(X) \mid Z_{1:m}) = \mathbb{P}_{\mathcal{D}}(s_{\mathsf{test}}^{(X,Y)} \le \widehat{\mathsf{Q}}_{1-\alpha} \mid Z_{1:m}) \qquad \text{and} \qquad \mathbb{P}_{\mathcal{D}}(s_{\mathsf{test}}^{(X,Y)} \le \widetilde{\mathsf{Q}}_{1-\alpha}).$$

For the two score sets $\{s_{\mathsf{test}}^{(X,Y)}\} \cup \{s_i^{(X,Y)}\}_{i=1}^m$ and $\{s_{\mathsf{test}}^{(X,Y)}\} \cup \{s_i\}_{i=1}^m$, Assumption 4.3 tells us that

$$\max_{i \in [m]} \big\{ \left| s_i^{(X,Y)} - s_i \right| \big\} \le \max_{i \in [m]} \big\{ \left| \widehat{\mu}_{Z_{1:m}}^{(X,Y)}(X_i) - \widehat{\mu}_{Z_{1:m}}(X_i) \right| \big\} \le \frac{L_2}{n}.$$

Then by virtue of Han et al. (2024a, Lemma B.1), we obtain

$$\left| \widehat{\mathsf{Q}}_{1-\alpha} - \widetilde{\mathsf{Q}}_{1-\alpha} \right| \le \frac{L_2}{n} \tag{85}$$

for any (X, Y) and $Z_{1:m}$. Combining this with Assumption 4.2 yields

$$\begin{aligned}
&\Big|\mathbb{P}_{\mathcal{D}}(Y \in \mathcal{C}(X) \mid Z_{1:m}) - \mathbb{P}_{\mathcal{D}}\big(s_{\mathsf{test}}^{(X,Y)} \le \widetilde{\mathsf{Q}}_{1-\alpha} \mid Z_{1:m}\big)\Big| \\
&\qquad = \Big|\mathbb{P}_{\mathcal{D}}\big(s_{\mathsf{test}}^{(X,Y)} \le \widehat{\mathsf{Q}}_{1-\alpha} \mid Z_{1:m}\big) - \mathbb{P}_{\mathcal{D}}\big(s_{\mathsf{test}}^{(X,Y)} \le \widetilde{\mathsf{Q}}_{1-\alpha} \mid Z_{1:m}\big)\Big| \\
&\qquad \le \mathbb{P}_{\mathcal{D}}\Big(s_{\mathsf{test}}^{(X,Y)} \in \big[\widetilde{\mathsf{Q}}_{1-\alpha} - \big|\widehat{\mathsf{Q}}_{1-\alpha} - \widetilde{\mathsf{Q}}_{1-\alpha}\big|, \widetilde{\mathsf{Q}}_{1-\alpha} + \big|\widehat{\mathsf{Q}}_{1-\alpha} - \widetilde{\mathsf{Q}}_{1-\alpha}\big|\big] \mid Z_{1:m}\Big) \\
&\qquad \le \mathbb{P}_{\mathcal{D}}\Big(s_{\mathsf{test}}^{(X,Y)} \in \Big[\widetilde{\mathsf{Q}}_{1-\alpha} - \frac{L_2}{n}, \widetilde{\mathsf{Q}}_{1-\alpha} + \frac{L_2}{n}\Big] \Big| Z_{1:m}\Big) \\
&\qquad \le \frac{4L_1L_2}{n} = \frac{4L}{n},
\end{aligned} \tag{86}$$

where the penultimate line results from (85), and the last line is due to Assumption 4.2. This inequality allows us to switch attention to $\mathbb{P}_{\mathcal{D}}(s_{\mathsf{test}}^{(X,Y)} \le \widetilde{\mathsf{Q}}_{1-\alpha} \mid Z_{1:m})$.

Step 2: replacing $\widetilde{\mathsf{Q}}_{1-\alpha}$ with an adjusted quantile independent of (X, Y). Note that by definition, $\widetilde{\mathsf{Q}}_{1-\alpha}$ still depends on the test score $s_{\mathsf{test}}^{(X,Y)}$, which motivates us to consider replacing $\widetilde{\mathsf{Q}}_{1-\alpha}$ with an alternative quantile independent of $s_{\mathsf{test}}^{(X,Y)}$. More precisely, define the following adjusted quantile

$$\check{\mathsf{Q}}_{1-\alpha} \;\coloneqq\; \inf\Big\{x \in \mathbb{R} : \sum_{n=1}^{m} \mathbb{1}\{s_i \le x\} \ge \lceil (1-\alpha)m - \alpha \rceil\Big\}., \tag{87}$$

which satisfies the following property.

Claim C.1. *The adjusted quantile defined in (87) satisfies*

$$\big\{s_{\mathsf{test}}^{(x,y)} \le \widetilde{\mathsf{Q}}_{1-\alpha}\big\} \quad\Longleftrightarrow\quad \big\{s_{\mathsf{test}}^{(x,y)} \le \check{\mathsf{Q}}_{1-\alpha}\big\}.$$

It then follows immediately from Claim C.1 that

$$\mathbb{P}_{\mathcal{D}}\big(s_{\mathsf{test}}^{(X,Y)} \le \widetilde{\mathsf{Q}}_{1-\alpha} \mid Z_{1:m}\big) = \mathbb{P}_{\mathcal{D}}\big(s_{\mathsf{test}}^{(X,Y)} \le \check{\mathsf{Q}}_{1-\alpha} \mid Z_{1:m}\big) = F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha}; Z_{1:m}), \tag{88}$$

where $F_{\mathsf{test}}(\cdot;\cdot)$ is defined at the beginning of this subsection. It then boils down to controlling $F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha}; Z_{1:m})$.

Proof of Claim C.1. From the definition of the quantile functional, one has

$$\begin{aligned}
\big\{s_{\mathsf{test}}^{(x,y)} \le \widetilde{\mathsf{Q}}_{1-\alpha}\big\} \quad &\Longleftrightarrow \quad s_{\mathsf{test}}^{(x,y)} \le \inf\Big\{q \in \mathbb{R} : \frac{1}{m+1}\mathbb{1}\{s_{\mathsf{test}}^{(x,y)} \le q\} + \frac{1}{m+1}\sum_{i=1}^{m}\mathbb{1}\{s_i \le q\} \ge 1-\alpha\Big\} \\
&\Longleftrightarrow \quad \mathbb{1}\{s_{\mathsf{test}}^{(x,y)} \le s_{\mathsf{test}}^{(x,y)}\} + \sum_{i=1}^{m}\mathbb{1}\{s_i \le s_{\mathsf{test}}^{(x,y)}\} \le \lceil (1-\alpha)(m+1)\rceil.
\end{aligned}$$

Given the trivial fact $\mathbb{1}\{s_{\mathsf{test}}^{(x,y)} \le s_{\mathsf{test}}^{(x,y)}\} = 1$, the last display is equivalent to

$$\sum_{i=1}^{m}\mathbb{1}\{s_i \le s_{\mathsf{test}}^{(x,y)}\} \le \lceil (1-\alpha)(m+1) - 1\rceil = \lceil (1-\alpha)m - \alpha\rceil.$$

By the definition (87) of $\check{\mathsf{Q}}_{1-\alpha}$, this inequality holds if and only if $s_{\mathsf{test}}^{(x,y)} \le \check{\mathsf{Q}}_{1-\alpha}$. This proves the claim. □

Step 3: controlling $F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha}; Z_{1:m})$. In order to control $F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha}; Z_{1:m})$, we begin with the following decomposition:

$$\Big|F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha}; Z_{1:m}) - (1-\alpha)\Big| \le \underbrace{\big|F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha}; Z_{1:m}) - F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha})\big|}_{=:\mathcal{T}_1} + \underbrace{\Big|\frac{1}{m}\sum_{i=1}^{m}\big[F_i^0(\check{\mathsf{Q}}_{1-\alpha}) - \mathbb{1}\big\{s_i \le \check{\mathsf{Q}}_{1-\alpha}\big\}\big]\Big|}_{=:\mathcal{T}_2}$$

$$+\underbrace{\left|F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha})-\frac{1}{m}\sum_{i=1}^{m}F_i^0(\check{\mathsf{Q}}_{1-\alpha})\right|}_{=:\mathcal{T}_3}+\underbrace{\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{s_i\le\check{\mathsf{Q}}_{1-\alpha}\right\}-(1-\alpha)\right|}_{=:\mathcal{T}_4}, \quad (89)$$

where both $F_{\mathsf{test}}(\cdot)$ and $F_i^0(\cdot)$ are defined at the beginning of this subsection. This decomposition leaves us with four terms to cope with.

- *Bounding $\mathcal{T}_1$ and $\mathcal{T}_2$.* Define the typical events $\mathcal{E}_1$ and $\mathcal{E}_2$ as:

$$\mathcal{E}_1 := \left\{\sup_{u\in\mathbb{R}}\left\{\left|F_{\mathsf{test}}(u;Z_{1:m})-F_{\mathsf{test}}(u)\right|\right\}\le\frac{16L}{n}\sqrt{m\log\frac{2}{\delta}}\right\};$$

$$\mathcal{E}_2 := \left\{\sup_{u\in\mathbb{R}}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{1}\left\{s_i\le u\right\}-F_i^0(u)\right)\right|\right\}\le 24\sqrt{\frac{\log(40/\delta)}{m}}+\frac{24L}{n}\sqrt{m\log\left(\frac{40m}{\delta}+n\right)}\right\}.$$

  Lemmas C.1 and C.2 imply that, with probability at least $1-\delta$, these two events occur simultaneously. On the event $\mathcal{E}_1\cap\mathcal{E}_2$, the terms $\mathcal{T}_1$ and $\mathcal{T}_2$ satisfy

$$\mathcal{T}_1 = \left|F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha};Z_{1:m})-F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha})\right|\le\sup_{u\in\mathbb{R}}\left|F_{\mathsf{test}}(u;Z_{1:m})-F_{\mathsf{test}}(u)\right|\le\frac{16L\sqrt{m\log(2/\delta)}}{n}; \quad (90)$$

$$\begin{aligned}\mathcal{T}_2 &= \left|\frac{1}{m}\sum_{i=1}^{m}\left[F_i^0(\check{\mathsf{Q}}_{1-\alpha})-\mathbb{1}\left\{s_i\le\check{\mathsf{Q}}_{1-\alpha}\right\}\right]\right|\le\sup_{u\in\mathbb{R}}\left|\frac{1}{m}\sum_{i=1}^{m}\left[F_i^0(u)-\mathbb{1}\left\{s_i\le u\right\}\right]\right|\\ &\le 24\sqrt{\frac{\log(40/\delta)}{m}}+\frac{24L}{n}\sqrt{m\log\left(\frac{40m}{\delta}+n\right)}. \end{aligned} \quad (91)$$

- *Bounding $\mathcal{T}_3$.* Regarding $\mathcal{T}_3$, it follows from the triangle inequality and the definition of the total-variation distance that

$$\begin{aligned}\mathcal{T}_3 &= \left|F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha})-\frac{1}{m}\sum_{i=1}^{m}F_i^0(\check{\mathsf{Q}}_{1-\alpha})\right|\le\frac{1}{m}\sum_{i=1}^{m}\left|F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha})-F_i^0(\check{\mathsf{Q}}_{1-\alpha})\right|\\ &\le\frac{1}{m}\sum_{i=1}^{m}\left(\sup_{u\in\mathbb{R}}\left\{\left|F_{\mathsf{test}}(u)-F_i(u)\right|\right\}+\sup_{u\in\mathbb{R}}\left\{\left|F_i^0(u)-F_i(u)\right|\right\}\right)\\ &\overset{\text{(a)}}{\le}\frac{1}{m}\sum_{i=1}^{m}\left(\mathsf{KS}\big(s_{\mathsf{test}}^{(X,Y)},s_i^{(X,Y)}\big)+\frac{4L}{n}\right)\overset{\text{(b)}}{\le}\frac{2}{m}\sum_{i=0}^{m}\mathsf{TV}(Z,Z_i)+\frac{8L}{n}, \end{aligned} \quad (92)$$

  where $F_i(\cdot)$ is defined at the beginning of this subsection. Inequalities (a) and (b) are justified below.

  – To validate inequality (a) in (92), observe that

$$\begin{aligned}F_i^0(u)-F_i(u) &= \mathbb{P}\big(s_i^{(X,Y)}\le u\big)-\mathbb{P}(s_i\le u)\\ &\overset{\text{(i)}}{\le}\mathbb{P}\Big(u-\left|\widehat{\mu}_{Z_{1:m}}^{(X,Y)}(X_i)-\widehat{\mu}_{Z_{1:m}\cup\{z_0\}}(X_i)\right|<s_i\le u+\left|\widehat{\mu}_{Z_{1:m}}^{(X,Y)}(X_i)-\widehat{\mu}_{Z_{1:m}\cup\{z_0\}}(X_i)\right|\Big)\\ &\overset{\text{(ii)}}{\le}\mathbb{P}\Big(u-\frac{L_2}{n}<s_i\le u+\frac{L_2}{n}\Big)\\ &\le\mathbb{P}\Big(\big(\widehat{\mu}_{Z_{1:m}\cup\{z_0\}}(X_i)-u\big)-\frac{L_2}{n}<Y_i\le\big(\widehat{\mu}_{Z_{1:m}\cup\{z_0\}}(X_i)-u\big)+\frac{L_2}{n}\Big)\\ &\quad+\mathbb{P}\Big(\big(\widehat{\mu}_{Z_{1:m}\cup\{z_0\}}(X_i)+u\big)-\frac{L_2}{n}<Y_i\le\big(\widehat{\mu}_{Z_{1:m}\cup\{z_0\}}(X_i)+u\big)+\frac{L_2}{n}\Big)\\ &\overset{\text{(iii)}}{\le}\frac{4L_1L_2}{n}=\frac{4L}{n}, \end{aligned}$$

where (ii) arises from Assumption 4.3, (iii) follows from Assumption 4.2, respectively, and (i) is a direct consequence of the definition of $s_i^{(X,Y)}$ and s_i (see the beginning of the subsection) and the following elementary fact.

Fact C.1. $\big|\mathbb{P}(|a|>u)-\mathbb{P}(|a+\delta|>u)\big| \le \mathbb{P}(u-|\delta|\le|a|\le u+|\delta|)$.

Proof of Fact C.1. We observe that

$$\begin{aligned}\big|\mathbb{P}(|a|>u)-\mathbb{P}(|a+\delta|>u)\big| &= \max\Big\{\mathbb{P}(|a|>u)-\mathbb{P}(|a+\delta|>u),\, \mathbb{P}(|a+\delta|>u)-\mathbb{P}(|a|>u)\Big\}\\ &\le \mathbb{P}(|a|+|\delta|>u)-\mathbb{P}(|a|-|\delta|>u) \le \mathbb{P}(u-|\delta|\le|a|\le u+|\delta|)\end{aligned}$$

as claimed. □

– We now justify inequality (b) in (92). Consider any index i, and define $Z^i_{1:m}$ as the dataset obtained from $Z_{1:m}$ by replacing the i-th sample Z_i with the target sample Z. Based on $Z^i_{1:m}$, introduce the auxiliary score

$$s_i' \coloneqq \big|\, Y_i - \widehat{\mu}^{(X_i,Y_i)}_{Z^i_{1:m}}(X_i)\,\big|.$$

In view of Assumption 4.3, s_i' differs from $s_i^{(X,Y)}$ by at most $2L_2/n$. Combining this with Assumption 4.2 immediately yields

$$\mathsf{KS}\big(s_i', s_i^{(X,Y)}\big) \le \frac{4L_1L_2}{n} = \frac{4L}{n}.$$

As a consequence, for each $i=1,\dots,m$ we have

$$\begin{aligned}\mathsf{KS}\big(s_{\mathsf{test}}^{(X,Y)}, s_i^{(X,Y)}\big) &\le \mathsf{KS}\big(s_{\mathsf{test}}^{(X,Y)}, s_i'\big) + \mathsf{KS}\big(s_i', s_i^{(X,Y)}\big)\\ &\overset{\text{(iv)}}{\le} \mathsf{TV}\Big(\big(Z_{1:m}, Z\big), \big(Z^i_{1:m}, Z_i\big)\Big) + \frac{4L}{n}\\ &\overset{\text{(v)}}{\le} 2\mathsf{TV}(Z, Z_i) + \frac{4L}{n}.\end{aligned}$$

Here, (iv) is valid since $s_{\mathsf{test}}^{(X,Y)}$ and s_i' are outputs of the same measurable mapping, evaluated at $(Z_{1:m}, Z)$ and $(Z^i_{1:m}, Z_i)$, respectively, whereas (v) invokes Barber et al. (2023, Lemma 1).

- *Bounding $\mathcal{T}_4$.* According to the definition of $\check{\mathsf{Q}}_{1-\alpha}$, the term $\mathcal{T}_4$ can be bounded by

$$\mathcal{T}_4 = \left|\frac{1}{m}\sum_{i=1}^m \mathbb{1}\left\{s_i \le \check{\mathsf{Q}}_{1-\alpha}\right\} - (1-\alpha)\right| = \left|\frac{\lceil (1-\alpha)m-\alpha\rceil}{m} - (1-\alpha)\right| < \frac{1}{m}, \tag{93}$$

where the last inequality follows since

$$\begin{aligned}-1 < (1-\alpha)m-\alpha-(1-\alpha)m \le \lceil (1-\alpha)m-\alpha\rceil - (1-\alpha)m\\ \le (1-\alpha)m-\alpha+1-(1-\alpha)m < 1.\end{aligned}$$

Substituting the preceding bounds on $\mathcal{T}_1,\dots,\mathcal{T}_4$ into (89), we arrive at

$$\begin{aligned}\Big|F_{\mathsf{test}}(\check{\mathsf{Q}}_{1-\alpha}; Z_{1:m}) - (1-\alpha)\Big| &\le \frac{16L\sqrt{m\log(2/\delta)}}{n} + 24\sqrt{\frac{\log(40/\delta)}{m}} + \frac{24L}{n}\sqrt{m\log\left(\frac{40m}{\delta}+n\right)}\\ &\quad + \frac{2}{m}\sum_{i=0}^m \mathsf{TV}(Z,Z_i) + \frac{8L}{n} + \frac{1}{m}\\ &\le \frac{40L\sqrt{m\log(45n/\delta)}}{n} + 24\sqrt{\frac{\log(40/\delta)}{m}} + \frac{2}{m}\sum_{i=0}^m \mathsf{TV}(Z,Z_i) + \frac{8L}{n} + \frac{1}{m}\\ &\le \frac{48L\sqrt{m\log(45n/\delta)}}{n} + 25\sqrt{\frac{\log(40/\delta)}{m}} + \frac{2}{m}\sum_{i=0}^m \mathsf{TV}(Z,Z_i).\end{aligned} \tag{94}$$

Step 4: putting all this together. Finally, taking the above results (86), (88) and (94) collectively, we can readily finish the proof of Proposition 4.1.

C.2 Proof of Theorem 4.1

This section is dedicated to establishing Theorem 4.1.

Notation. For ease of presentation, we adopt the notation introduced in Definition B.1 as well as in Section 4.1. In addition, we shall adopt the following notation:

- For any $k \le m < t \in [T]$, let

$$\mathsf{Q}_{1-\alpha}^{k,m,t} := \mathsf{Q}_{1-\alpha}\left(\frac{1}{m-k+2}\left(\delta\{s_{Z_{1:m}}^{Z_t}(Z_t)\} + \sum_{l=k}^{m}\delta\{s_{Z_{1:m}}^{Z_t}(Z_l)\}\right)\right), \tag{95}$$

  representing the quantile when (i) the data $Z_{1:m}\cup\{Z_t\}$ are used for training; and (ii) the data $Z_{k:m}\cup\{Z_t\}$ are used for calibration.

- For any $k \le m < i \le j$, define the event

$$\mathcal{A}(k,m;i,j) := \left\{\left|\sum_{t=i}^{j}\Big(\mathbb{1}\big\{s_{Z_{1:m}}^{Z_t}(Z_t) \le \mathsf{Q}_{1-\alpha}^{k,m,t}\big\} - \mathbb{P}_{\mathcal{D}_t}\big(s_{Z_{1:m}}^{Z_t}(Z_t) \le \mathsf{Q}_{1-\alpha}^{k,m,t} \mid Z_{1:m}\big)\Big)\right| \le 2\sqrt{(j-i+1)\log(2j)}\right\}, \tag{96a}$$

  which is concerned with the deviation of the empirical coverage from the training-conditional mean coverage over the time window $[i,j]$. As we shall see momentarily, this is a high-probability event.

- For any stage-round pair (n,r), take

$$\mathcal{A}_{n,r} := \bigcap_{i=\tau_{n,r}}^{\tau_{n,r+1}-1}\ \bigcap_{j=i+1}^{\tau_{n,r+1}-1}\mathcal{A}(\tau_{n,r-1},\tau_{n,r}-1;i,j). \tag{96b}$$

Step 1: regret decomposition. Following the proof structure of Theorem 3.1, we decompose the cumulative regret of interest as

$$\begin{aligned}\sum_{t=1}^{T}\big|\mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{C}_t) - (1-\alpha)\big| &= \sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\Big|\mathbb{P}(Y_{n,r,l} \in \mathcal{C}_{n,r}(X_{n,r,l}) \mid \mathcal{C}_{n,r}) - (1-\alpha)\Big|\,\mathbb{1}\{\mathcal{A}_{n,r}\} \\ &\quad + \sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\Big|\mathbb{P}(Y_{n,r,l} \in \mathcal{C}_{n,r}(X_{n,r,l}) \mid \mathcal{C}_{n,r}) - (1-\alpha)\Big|\,\mathbb{1}\{\mathcal{A}_{n,r}^{\mathrm{c}}\}.\end{aligned} \tag{97}$$

Here we write $\mathbb{P}(\cdot \mid \mathcal{C}_{n,r})$ (or $\mathbb{P}(\cdot \mid \mathcal{C}_t)$) for the conditional probability given the set-valued mapping $\mathcal{C}_{n,r}(\cdot)$ (or $\mathcal{C}_t(\cdot)$). Further, similar to (39), we augment the data to simplify notation: although the last round of stage n contains only $t_n \le T_r$ time instances, we still generate $Z_{n,r,l} = (X_{n,r,l}, Y_{n,r,l})$ for every $l > t_n$ in an i.i.d. manner obeying

$$\mathbb{P}(Y_{n,r,l} \in \mathcal{C}_{n,r}(X_{n,r,l}) \mid \mathcal{C}_{n,r}) = 1-\alpha \qquad \text{for all } l > t_n. \tag{98}$$

Step 2: bounding the second term on the right-hand side of (97). Akin to the pretrained-score setting, the first term on the right-hand side of (97) is the dominant term in the above regret decomposition. To justify this, let us look at the second term on the right-hand side of (97). Fix a realization $Z_{1:m} = z_{1:m}$. Then for each $t \in [i,j]$ with $m < i$, the indicator $\mathbb{1}\big\{s_{z_{1:m}}^{Z_t}(Z_t) \le Q_{1-\alpha}^{k,m,t}\big\}$ is a function of Z_t only, and hence

this collection of indicator variables over the time window $[i,j]$ are statistically independent. Moreover, the conditional mean of this indicator variable at time t is $\mathbb{P}_{\mathcal{D}_t}\big(s^{Z_t}_{z_{1:m}}(Z_t) \le Q^{k,m,t}_{1-\alpha} \mid Z_{1:m} = z_{1:m}\big)$. Therefore, Hoeffding's inequality readily yields

$$\mathbb{P}\big(\mathcal{A}(k,m;i,j)^{\mathrm{c}} \mid Z_{1:m} = z_{1:m}\big) \le j^{-8} \quad \text{for all } z_{1:m} \tag{99a}$$

$$\Longrightarrow \quad \mathbb{P}\big(\mathcal{A}(k,m;i,j)^{\mathrm{c}}\big) \le \mathbb{E}_{Z_{1:m}}\big[\mathbb{P}\big(\mathcal{A}(k,m;i,j)^{\mathrm{c}} \mid Z_{1:m}\big)\big] \le j^{-8}. \tag{99b}$$

Now consider any time point $t \ge 4$ that resides within round r of stage n. By the construction of DRIFTOCP-FULL, we have

$$\frac{t}{16} \le \frac{\tau_{n,r}}{4} \le \tau_{n,r-1} < \tau_{n,r} \le t \le \tau_{n,r+1} \le 4t.$$

Consequently, defining

$$\mathcal{E}_t := \bigcap_{m=\frac{t}{4}}^{t} \bigcap_{k=\frac{t}{16}}^{m} \bigcap_{j=\frac{t}{4}}^{4t} \bigcap_{i=\frac{t}{4}}^{t} \mathcal{A}(k,m;i,j)$$

in which the index pair (k,m) ranges over all values that $(\tau_{n,r-1}, \tau_{n,r})$ may take, we have

$$\mathcal{E}_t \subseteq \mathcal{A}_{n,r}.$$

Regarding this event $\mathcal{E}_t$, it follows from (99) that

$$\mathbb{P}(\mathcal{E}_t^{\mathrm{c}}) \le \sum_{m=\frac{t}{4}}^{t} \sum_{k=\frac{t}{16}}^{m} \sum_{j=\frac{t}{4}}^{4t} \sum_{i=\frac{t}{4}}^{j} \mathbb{P}\big(\mathcal{A}(k,m;i,j)^{\mathrm{c}}\big) \le t^2 \sum_{j=\frac{t}{4}}^{4t} j\mathbb{P}\big(\mathcal{A}(k,m;i,j)^{\mathrm{c}}\big) \le t^2 \sum_{j=\frac{t}{4}}^{4t} \frac{1}{j^7} = O\big(t^{-4}\big).$$

Therefore, the second term on the right-hand side of (97) can be bounded above by

$$\begin{aligned} &\mathbb{E}\bigg[\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\Big|\mathbb{P}(Y_{n,r,l} \in \mathcal{C}_{n,r}(X_{n,r,l}) \mid \mathcal{C}_{n,r}) - (1-\alpha)\Big|\, \mathbb{1}\{\mathcal{A}_{n,r}^{\mathrm{c}}\}\bigg] \\ &\qquad \le \mathbb{E}\left[\sum_{t=1}^{T}\Big|\mathbb{P}(Y_t \in \mathcal{C}_t(X_t) \mid \mathcal{C}_{n,r}) - \alpha\Big|\, \mathbb{1}\{\mathcal{E}_t^{\mathrm{c}}\}\right] \le \sum_{t=1}^{T}\mathbb{P}(\mathcal{E}_t^{\mathrm{c}}) = O\left(\sum_{t=1}^{T}\frac{1}{t^4}\right) = O(1). \end{aligned} \tag{100}$$

Step 3: bounding the first term on the right-hand side of (97). It remains to bound the first term on the right-hand side of (97). The overall proof follows a similar strategy to that used in the pretrained-score setting. To avoid unnecessary repetition, we shall focus primarily on the steps that differ nontrivially from the pretrained-score case.

To begin with, by adapting the arguments in the proof of Lemma B.2, we obtain the following result, whose proof of Lemma C.3 is deferred to Section C.3.3.

Lemma C.3. *Consider any stage n in Algorithm 4, which comprises r_n rounds and S_n time points. Reusing the notation introduced in Definition B.1, we have*

$$\begin{aligned} &\sum_{r=1}^{r_n}\left(\sum_{l=1}^{T_r}\Big|\mathbb{P}(Y_{n,r,l} \in \mathcal{C}_{n,r} \mid \mathcal{C}_{n,r}) - (1-\alpha)\Big|\right)\mathbb{1}\{\mathcal{A}_{n,r}\} \\ &\qquad \le \begin{cases} \widetilde{O}\Big(\sum_{j=1}^{J_n}\sqrt{|\mathcal{I}_{n,j}|}\Big), & \textit{for the change-point setting,} \\ \widetilde{O}\left(\sqrt{S_n} + (\mathsf{TV}_n^{\mathsf{stage}})^{\frac{1}{3}} S_n^{\frac{2}{3}}\right), & \textit{for the smooth drift setting,} \end{cases} \end{aligned} \tag{101}$$

where we set

$$\mathsf{TV}_n^{\mathsf{stage}} := \sum_{r=1}^{r_n}\sum_{l=1}^{S_{n,r}-1} \mathsf{TV}(Z_{n,r,l}, Z_{n,r,l+1}). \tag{102}$$

Given that Lemma C.3 controls the cumulative regret within a single stage, it remains to extend this bound to the entire time horizon, which we handle separately for the two drift settings in Steps 4 and 5.

Before proceeding, let us introduce a collection of typical events that will be used in both settings. For any two time points $1 \le k < m \le T$, define the event

$$\mathcal{G}(k,m) := \Big\{ \Big| \mathbb{P}\Big(Y_m \notin \mathcal{C}(X_m \mid Z_{k:m-1}, Z_{1:m-1}) \mid Z_{1:m-1}\Big) - \alpha \Big| \le 2^6 \sqrt{\frac{\log(40m)}{m-k}} + \frac{2^7 L \sqrt{(m-k)\log(40m)}}{m} + \frac{1}{m-k} \sum_{l=k}^{m-1} \mathsf{TV}\big(Z_m, Z_l\big) \Big\}, \tag{103}$$

as motivated by Proposition 4.1. Here the notation $\mathcal{C}(\cdot \mid \cdot, \cdot)$ is defined at the beginning of Section C. For any stage n, define the typical event $\mathcal{B}_n$ as follows

$$\mathcal{B}_n := \mathcal{A}_{n,r_n} \cap \mathcal{G}(\tau_{n,r_n-1}, \tau_{n,r_n}). \tag{104}$$

In addition, the following two lemmas will be used in the analysis for both drift settings, and we therefore state them here for subsequent use. The first lemma shows that $\mathcal{B}_n$ is an event with sufficiently high probability (even when suitably weighted by τ_{n+1}); the proof can be found in Section C.3.4.

Lemma C.4. *For any $n \ge 1$, recall that τ_n denotes the starting time of stage n. Then we have*

$$\mathbb{E}\big[\tau_{n+1} \mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\}\big] \le O\big(n^{-2}\big).$$

Another useful lemma shows that the aggregate total variation over the last two rounds of each stage is sufficiently large (at least on some high-probability event). The proof can be found in Section C.3.5.

Lemma C.5. *For any stage $n \le N-1$, define*

$$\mathsf{TV}_n^{\mathsf{tail}} := \sum_{j=1}^{T_{r_n}-1} \mathsf{TV}(Z_{n,r_n,j}, Z_{n,r_n,j+1}) + \sum_{i=1}^{T_{r_n-1}} \mathsf{TV}(Z_{n,r_n-1,i}, Z_{n,r_n-1,i+1}), \tag{105}$$

and introduce the event

$$\mathcal{H}_n := \Big\{ T_{r_n-1} \sqrt{\log(40\tau_{n,r_n})} \le \frac{\tau_{n,r_n}}{256} \Big\}. \tag{106}$$

Recalling that $t_n \le T_{r_n}$ is the number of iterations in round r_n of stage n, one has

$$\sqrt{t_n}\, \mathsf{TV}_n^{\mathsf{tail}} \mathbb{1}\big\{\mathcal{B}_n \cap \mathcal{H}_n\big\} \ge 3 \cdot \mathbb{1}\big\{\mathcal{B}_n \cap \mathcal{H}_n\big\}.$$

Step 4: analysis for the change-point setting. In this setting, Lemma C.3 allows one to decompose

$$\begin{aligned} \sum_{n=1}^{N} \sum_{r=1}^{r_n} \sum_{l=1}^{T_r} \big|\mathbb{P}(Y_{n,r,l} \notin \mathcal{C}_{n,r}(X_{n,r,l}) \,|\, \mathcal{C}_{n,r}) - \alpha\big| \, \mathbb{1}\{\mathcal{A}_{n,r}\} &\le \widetilde{O}\bigg(\sum_{n=1}^{N} \Big(\sum_{j=1}^{J_n} \sqrt{|\mathcal{I}_{n,j}|} \Big) \bigg) \\ &\le \widetilde{O}\bigg(\sum_{n=1}^{N} \Big(\sum_{j=1}^{J_n} \sqrt{|\mathcal{I}_{n,j}|} \Big) \mathbb{1}\{\mathcal{B}_n\} + \sum_{n=1}^{N} (\tau_{n+1} - \tau_n) \mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\} \bigg), \end{aligned} \tag{107}$$

where the last line follows since, by Cauchy-Schwarz,

$$\sum_{j=1}^{J_n} \sqrt{|\mathcal{I}_{n,j}|} \le \sqrt{J_n \sum_{j=1}^{J_n} |\mathcal{I}_{n,j}|} = \sqrt{J_n(\tau_{n+1} - \tau_n)} \le \tau_{n+1} - \tau_n.$$

- We start with the last term on the right-hand side of (107), for which Lemma C.4 indicates that

$$\mathbb{E}\Big[\sum_{n=1}^{N} (\tau_{n+1} - \tau_n) \mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\} \Big] \le \sum_{n=1}^{\infty} \mathbb{E}\big[\tau_{n+1} \mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\}\big] \le O\bigg(\sum_{n=1}^{\infty} \frac{1}{n^2} \bigg) = O(1). \tag{108}$$

- When it comes to the first term on the right-hand side of (107), we first divide it into

$$\sum_{n=1}^{N}\Big(\sum_{j=1}^{J_n}\sqrt{|\mathcal{I}_{n,j}|}\Big)\mathbb{1}\{\mathcal{B}_n\}\le\sum_{n=1}^{N}\Big(\sum_{j=1}^{J_n-1}\sqrt{|\mathcal{I}_{n,j}|}\Big)+\sum_{n=1}^{N}\sqrt{|\mathcal{I}_{n,J_n}|}\mathbb{1}\{\mathcal{B}_n\cap\mathcal{H}_n\}+\sum_{n=1}^{N}\sqrt{|\mathcal{I}_{n,J_n}|}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\},\tag{109}$$

where $\mathcal{H}_n$ is defined in (106). We shall bound the three terms on the right-hand side of (109) separately.

  - As for the first term on the right-hand side of (109), we first make the observation that: the time segments $\mathcal{I}_{n,j}$ ($n=1,\ldots,N$ and $j=1,\ldots,J_n-1$) belong to distinct time segments in $\{\mathcal{I}_k\}_{k=1}^{N^{\mathsf{cp}}+1}$. As a result, we can derive

$$\sum_{n=1}^{N}\sum_{j=1}^{J_n-1}\sqrt{|\mathcal{I}_{n,j}|}\le\sum_{k=1}^{N^{\mathsf{cp}}+1}\sqrt{|\mathcal{I}_k|}\le\sqrt{(N^{\mathsf{cp}}+1)T},\tag{110}$$

    where the last relation arises from the Cauchy-Schwarz inequality.

  - With regards to the second term on the right-hand side of (109), by the construction of $\mathcal{I}_{n,j}$ we know that, for each terminal interval $\mathcal{I}_{n,J_n}$, there exists a unique time segment $\mathcal{I}_{k_n}$ defined in Definition B.1 such that $\mathcal{I}_{n,J_n}\subseteq\mathcal{I}_{k_n}$. Moreover, the indices are nondecreasing, namely $k_n\le k_{n+1}$ for $n=1,\ldots,N-1$. In particular, when $J_n\ge 2$, stage n must contain a distribution change, which implies $k_n>k_{n-1}$. Using these properties, we can obtain

$$\begin{aligned}\sum_{n=1}^{N}\sqrt{|\mathcal{I}_{n,J_n}|}\mathbb{1}\{\mathcal{B}_n\cap\mathcal{H}_n\}&\overset{\text{(a)}}{\le}\sum_{n=1}^{N}\sqrt{|\mathcal{I}_{n,J_n}|}\mathbb{1}\{J_n\ge 2\}\\&\le\sum_{n=1}^{N}\sqrt{|\mathcal{I}_{k_n}|}\mathbb{1}\{k_n>k_{n-1}\}\le\sum_{k=1}^{N^{\mathsf{cp}}+1}\sqrt{|\mathcal{I}_k|}\le\sqrt{(N^{\mathsf{cp}}+1)T}.\end{aligned}\tag{111}$$

    Here, (a) follows since, on the event $\mathcal{B}_n\cap\mathcal{H}_n$, we have $\sqrt{t_n}\mathsf{TV}_n^{\mathsf{tail}}\ge 3>0$ (according to Lemma C.5), which implies that stage n must contain a distribution shift and hence necessarily requires $J_n\ge 2$.

  - It remains to bound the third term on the right-hand side of (109). To this end, we make note of the following relations between the two sequences $\{\tau_{n,r}\}_{n,r}$ and $\{T_r\}_r$:

$$\tau_{n-1,r_{n-1}}\le\tau_{n,r_n-1};\quad \tau_{n,r_n}-\tau_{n,r_n-1}=T_{r_n-1}.$$

    In particular, the intervals $\{[\tau_{n,r_n-1}+1,\tau_{n,r_n}]\}_{n=1}^{N}$ are pairwise disjoint, and as a result,

$$\begin{aligned}\sum_{n=1}^{N}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}&=\sum_{n=1}^{N}\mathbb{1}\left\{T_{r_n-1}\sqrt{\log(40\tau_{n,r_n})}>\frac{\tau_{n,r_n}}{256L}\right\}\le\sum_{n=1}^{N}\frac{256L\sqrt{\log(40T)}(\tau_{n,r_n}-\tau_{n,r_n-1})}{\tau_{n,r_n}}\\&\le 256L\sqrt{\log(40T)}\sum_{n=1}^{N}\sum_{i=\tau_{n,r_n-1}+1}^{\tau_{n,r_n}}\frac{1}{\tau_{n,r_n}}\le 256L\sqrt{\log(40T)}\sum_{n=1}^{N}\sum_{i=\tau_{n,r_n-1}+1}^{\tau_{n,r_n}}\frac{1}{i}\\&\le 256L\sqrt{\log(40T)}\sum_{i=1}^{T}\frac{1}{i}\le 256L\big(\log(40T)\big)^{\frac{3}{2}}.\end{aligned}\tag{112}$$

    Consequently, taking this together with the Cauchy-Schwarz inequality yields

$$\begin{aligned}\sum_{n=1}^{N}\sqrt{|\mathcal{I}_{n,J_n}|}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}&\le\sum_{n=1}^{N}\sqrt{S_n}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}\le\sqrt{\Big(\sum_{n=1}^{N}S_n\Big)\Big(\sum_{n=1}^{N}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}\Big)}\\&\le\sqrt{T\Big(\sum_{n=1}^{N}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}\Big)}\overset{(112)}{\le}\widetilde{O}\big(\sqrt{LT}\big).\end{aligned}\tag{113}$$

Combining (107)–(111) and (113) reveals that

$$\mathbb{E}\left[\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\left|\mathbb{P}(Y_{n,r,l}\notin\mathcal{C}_{n,r}(X_{n,r,l})\,|\,\mathcal{C}_{n,r})-\alpha\right|\mathbb{1}\{\mathcal{A}_{n,r}\}\right]\leq\widetilde{O}(\sqrt{(N^{\mathsf{cp}}+L+1)T}),\tag{114}$$

which together with (97) and (100) establishes the advertised regret bound for the change-point setting.

Step 5: analysis for the smooth drift setting. With Lemma C.3 in mind, we first analyze $\sum_{n=1}^{N}\sqrt{S_n}$. Recalling the definition of $\mathsf{TV}_n^{\mathsf{tail}}$ in (105), we make the observation that

$$\begin{aligned}
\sum_{n=1}^{N}\sqrt{S_n}\mathbb{1}\left\{\mathcal{B}_n\right\} &\leq \sqrt{S_N}+\sum_{n=1}^{N-1}\sqrt{S_n}\mathbb{1}\left\{\mathcal{H}_n\right\}\mathbb{1}\left\{\mathcal{B}_n\right\}+\sum_{n=1}^{N}\sqrt{S_n}\mathbb{1}\left\{\mathcal{H}_n^{\mathrm{c}}\right\}\mathbb{1}\left\{\mathcal{B}_n\right\}\\
&\leq \sqrt{T}+\sum_{n=1}^{N-1}\sqrt{S_n}\left(\sqrt{t_n}\mathsf{TV}_n^{\mathsf{tail}}\right)^{\frac{1}{3}}+\sqrt{\left(\sum_{n=1}^{N}S_n\right)\left(\sum_{n=1}^{N}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}\right)}\\
&\leq \sqrt{T}+\sum_{n=1}^{N-1}S_n^{\frac{2}{3}}\left(\mathsf{TV}_n^{\mathsf{tail}}\right)^{\frac{1}{3}}+\sqrt{T\left(\sum_{n=1}^{N}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}\right)}\\
&\leq \sqrt{T}+\left(\sum_{n=1}^{N-1}S_n\right)^{\frac{2}{3}}\left(\sum_{n=1}^{N-1}\mathsf{TV}_n^{\mathsf{tail}}\right)^{\frac{1}{3}}+\sqrt{T\left(\sum_{n=1}^{N}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}\right)}\\
&\leq \sqrt{T}+2T^{\frac{2}{3}}\mathsf{TV}_T^{\frac{1}{3}}+\sqrt{T\left(\sum_{n=1}^{N}\mathbb{1}\{\mathcal{H}_n^{\mathrm{c}}\}\right)}\overset{(112)}{=}\widetilde{O}\Big(T^{\frac{2}{3}}\mathsf{TV}_T^{\frac{1}{3}}+\sqrt{(L+1)T}\Big),
\end{aligned}\tag{115}$$

where the second line arises from Lemma C.5 and the Cauchy-Schwarz inequality, the penultimate line results from Hölder's inequality, and the last inequality holds because for any n, $\mathsf{TV}_n^{\mathsf{tail}}$ is counted at most twice in the summation from 1 to N. Taking this collectively with Lemma C.4, we can demonstrate that

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{N}\sqrt{S_n}\right] &\leq \mathbb{E}\left[\sum_{n=1}^{N}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n\}\right]+\mathbb{E}\left[\sum_{n=1}^{N}\sqrt{S_n}\mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\}\right]\\
&\leq \widetilde{O}\left(\sqrt{(L+1)T}+T^{\frac{3}{2}}\mathsf{TV}_T^{\frac{1}{3}}\right)+\mathbb{E}\left[\sum_{n=1}^{N}\tau_{n+1}\mathbb{1}\{\mathcal{B}_n^{c}\}\right]\\
&\leq \widetilde{O}\left(\sqrt{(L+1)T}+T^{\frac{3}{2}}\mathsf{TV}_T^{\frac{1}{3}}\right)+\sum_{n=1}^{\infty}\mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{B}_n^{c}\}\big]\\
&= \widetilde{O}\left(\sqrt{(L+1)T}+T^{\frac{3}{2}}\mathsf{TV}_T^{\frac{1}{3}}+\sum_{n=1}^{\infty}\frac{1}{n^2}\right)=\widetilde{O}\left(\sqrt{(L+1)T}+\mathsf{TV}_T^{\frac{1}{3}}T^{\frac{2}{3}}\right).
\end{aligned}$$

Armed with the above bound, we can readily invoke Lemma C.3 and apply Hölder's inequality to yield

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^{N}\sum_{r=1}^{r_n}\sum_{l=1}^{T_r}\Big|\mathbb{P}(Y_{n,r,l}\in\mathcal{C}_{n,r}(X_{n,r,l})\,|\,\mathcal{C}_{n,r})-(1-\alpha)\Big|\mathbb{1}\{\mathcal{A}_{n,r}\}\right] &= \widetilde{O}\Big(\mathbb{E}\Big[\sum_{n=1}^{N}\sqrt{S_n}\Big]+\sum_{n=1}^{N}\left(\mathsf{TV}_n^{\mathsf{stage}}\right)^{\frac{1}{3}}S_n^{\frac{2}{3}}\Big)\\
&\leq \widetilde{O}\Bigg(\sqrt{(L+1)T}+\mathsf{TV}_T^{\frac{1}{3}}T^{\frac{2}{3}}+\left(\sum_{n=1}^{N}\mathsf{TV}_n^{\mathsf{stage}}\right)^{\frac{1}{3}}\left(\sum_{n=1}^{N}S_n\right)^{\frac{2}{3}}\Bigg)\\
&\leq \widetilde{O}\Big(\sqrt{(L+1)T}+\mathsf{TV}_T^{\frac{1}{3}}T^{\frac{2}{3}}\Big).
\end{aligned}$$

Taking this together with (97) and (100) establishes the claimed regret bound for the smooth drift setting.

C.3 Proof of auxiliary lemmas

C.3.1 Proof of Lemma C.1

For notational convenience, we write $Z_{1:m}$ here in place of $Z^{\mathsf{cal}}_{1:m}$ as long as it is clear from the context.

To apply McDiarmid's inequality, consider two given calibration datasets, $z_{1:m}$ and $z'_{1:m}$, which differ in exactly one sample. For any given point $(x, y) \in \mathcal{Z}$, define the corresponding scores as

$$s^{(X,Y)}_{z_{1:m}}(x,y) \coloneqq \left|y - \widehat{\mu}^{(X,Y)}_{z_{1:m}}(x)\right|, \qquad s^{(X,Y)}_{z'_{1:m}}(x,y) \coloneqq \left|y - \widehat{\mu}^{(X,Y)}_{z'_{1:m}}(x)\right|.$$

Note that both $\widehat{\mu}^{(X,Y)}_{z_{1:m}}(\cdot)$ and $\widehat{\mu}^{(X,Y)}_{z'_{1:m}}(\cdot)$ are trained on $n+1$ data points. Then, by Assumption 4.3, for any $x \in \mathcal{X}$ we have

$$\left|\widehat{\mu}^{(X,Y)}_{z_{1:m}}(x) - \widehat{\mu}^{(X,Y)}_{z'_{1:m}}(x)\right| \leq \frac{L_2}{n}.$$

Combining this with Assumption 4.2, which assumes the Lipschitz continuity of the distribution function, we see that: for $(X, Y) \sim \mathcal{D}$,

$$\begin{aligned}
\left|\mathbb{P}_{\mathcal{D}}\left(s^{(X,Y)}_{z_{1:m}}(X,Y) > u\right) - \mathbb{P}_{\mathcal{D}}\left(s^{(X,Y)}_{z'_{1:m}}(X,Y) > u\right)\right| &\overset{\text{(a)}}{\leq} \mathbb{P}_{\mathcal{D}}\left(u - \Delta \leq s^{(X,Y)}_{z_{1:m}}(X,Y) \leq u + \Delta\right) \\
&\overset{\text{(b)}}{\leq} 4L_1 \sup_{X,Y}\left\{\left|\widehat{\mu}^{(X,Y)}_{z_{1:m}}(X) - \widehat{\mu}^{(X,Y)}_{z'_{1:m}}(X)\right|\right\} \\
&\leq \frac{4L_1L_2}{n} = \frac{4L}{n},
\end{aligned} \tag{116}$$

where $\Delta \coloneqq \left|\widehat{\mu}^{(X,Y)}_{z_{1:m}}(X) - \widehat{\mu}^{(X,Y)}_{z'_{1:m}}(X)\right|$. Here, (a) is a result of Fact C.1 whereas (b) follows since

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}\left(u - \Delta \leq s^{(X,Y)}_{z_{1:m}}(X,Y) \leq u + \Delta\right) &\leq \mathbb{P}_{\mathcal{D}}\left(-u - \Delta \leq Y - \widehat{\mu}^{(X,Y)}_{z_{1:m}} \leq -u + \Delta\right) \\
&\quad + \mathbb{P}_{\mathcal{D}}\left(u - \Delta \leq Y - \widehat{\mu}^{(X,Y)}_{z_{1:m}} \leq u + \Delta\right) \\
&\leq \sup_{\mu \in \mathbb{R}} \mathbb{P}_{\mathcal{D}}\Big((\mu - u) - \Delta' \leq Y \leq \Delta' + (\mu - u)\Big) \\
&\quad + \sup_{\mu \in \mathbb{R}} \mathbb{P}_{\mathcal{D}}\Big((\mu + u) - \Delta' \leq Y \leq \Delta' + (\mu + u)\Big) \leq 4L_1\Delta'
\end{aligned} \tag{117}$$

with $\Delta' \coloneqq \sup_{X,Y}\left\{\left|\widehat{\mu}^{(X,Y)}_{z_{1:m}}(X) - \widehat{\mu}^{(X,Y)}_{z'_{1:m}}(X)\right|\right\}$.

From (116), we observe that $\mathbb{P}_{\mathcal{D}}\big(s^{(X,Y)}_{z_{1:m}}(X,Y) > u\big)$—when viewed as a function of $z_{1:m}$—satisfies the bounded difference property with coefficient $4L/n$. Now, we make the following definition:

$$\begin{aligned}
Q(z_{1:m}, u) &\coloneqq \left|\mathbb{P}_{\mathcal{D}}\left(s^{(X,Y)}_{z_{1:m}}(X,Y) > u\right) - \mathbb{P}_{\mathcal{D}_{1:m}\times\mathcal{D}}\left(s^{(X,Y)}_{Z_{1:m}}(X,Y) > u\right)\right|, \\
Q(z_{1:m}) &\coloneqq \sup_{u \in \mathbb{R}}\left\{Q(z_{1:m}, u)\right\},
\end{aligned}$$

where

$$\mathbb{P}_{\mathcal{D}_{1:m}\times\mathcal{D}}\left(s^{(X,Y)}_{Z_{1:m}}(X,Y) > u\right) \coloneqq \mathbb{E}_{Z_{1:m}\sim\mathcal{D}_{1:m}}\left[\mathbb{P}_{(X,Y)\sim\mathcal{D}}\left(s^{(X,Y)}_{Z_{1:m}}(X,Y) > u \,\middle|\, Z_{1:m}\right)\right]. \tag{118}$$

Then, basic calculation yields

$$\begin{aligned}
Q(z_{1:m}) - Q(z'_{1:m}) &= \sup_{u \in \mathbb{R}}\left\{Q(z_{1:m}, u)\right\} - \sup_{u \in \mathbb{R}}\left\{Q(z'_{1:m}, u)\right\} \\
&\leq \sup_{u \in \mathbb{R}}\left\{Q(z_{1:m}, u) - Q(z'_{1:m}, u)\right\} \leq \frac{4L}{n}.
\end{aligned}$$

Hence, by applying McDiarmid's inequality (Lemma E.1), we can demonstrate that

$$\begin{aligned}\sup_{u\in\mathbb{R}}\left\{\left|\mathbb{P}_{\mathcal{D}}\left(s^{(X,Y)}_{Z_{1:m}}(X,Y)>u\;\middle|\;Z_{1:m}\right)-\mathbb{P}_{\mathcal{D}_{1:m}\times\mathcal{D}}\left(s^{(X,Y)}_{Z_{1:m}}(X,Y)>u\right)\right|\right\}&=Q(Z_{1:m})\\ &\le \mathbb{E}_{Z_{1:m}}\left[Q(Z_{1:m})\right]+4L\frac{\sqrt{m\log\frac{1}{\delta}}}{n}\end{aligned}\tag{119}$$

holds with probability exceeding $1-\delta$.

Now consider any given $z_0=(x_0,y_0)$, and denote $\widehat{\mu}^0_{z_{1:m}}(\cdot)\coloneqq\widehat{\mu}^{(x_0,y_0)}_{z_{1:m}}(\cdot)$ and $s^0_{z_{1:m}}(x,y)=\left|y-\widehat{\mu}^0_{z_{1:m}}(x)\right|$. Let $Z^*_{1:m}$ be an independent copy of $Z_{1:m}$. Then one can show that

$$\begin{aligned}\mathbb{E}_{Z_{1:m}}[Q(Z_{1:m})]&=\mathbb{E}_{Z_{1:m}}\left[\sup_{u\in\mathbb{R}}\left|\mathbb{P}_{\mathcal{D}}\big(s^{(X,Y)}_{Z_{1:m}}(X,Y)>u\mid Z_{1:m}\big)-\mathbb{P}_{\mathcal{D}_{1:m}\times\mathcal{D}}\big(s^{(X,Y)}_{Z_{1:m}}(X,Y)>u\big)\right|\right]\\ &\le \mathop{\mathbb{E}}_{Z_{1:m},Z^*_{1:m}}\left[\sup_{u\in\mathbb{R}}\left|\mathbb{P}_{\mathcal{D}}\big(s^{(X,Y)}_{Z_{1:m}}(X,Y)>u\mid Z_{1:m}\big)-\mathbb{P}_{\mathcal{D}}\big(s^{(X,Y)}_{Z^*_{1:m}}(X,Y)>u\mid Z^*_{1:m}\big)\right|\right]\\ &\le 4L_1\mathop{\mathbb{E}}_{Z_{1:m},Z^*_{1:m}}\left[\sup_{u\in\mathbb{R}}\left\{\mathbb{E}_X\left[\left|\widehat{\mu}^0_{Z_{1:m}}(X)-\widehat{\mu}^0_{Z^*_{1:m}}(X)\right|+\frac{L_2}{n}\right]\right\}\right]\\ &= 4L_1\mathop{\mathbb{E}}_{Z_{1:m},Z^*_{1:m},X}\left[\left|\widehat{\mu}^0_{Z_{1:m}}(X)-\widehat{\mu}^0_{Z^*_{1:m}}(X)\right|\right]+\frac{8L}{n}.\end{aligned}\tag{120}$$

Here, the second line follows from Jensen's inequality; the penultimate line follows since, for any given two arrays $z_{1:m}$ and $z^*_{1:m}$, one has

$$\begin{aligned}&\left|\mathbb{P}_{\mathcal{D}}\Big(s^{(X,Y)}_{z_{1:m}}(X,Y)>u\Big)-\mathbb{P}_{\mathcal{D}}\Big(s^{(X,Y)}_{z^*_{1:m}}(X,Y)>u\Big)\right|\\ &\le\left|\mathbb{P}_{\mathcal{D}}\Big(s^{(X,Y)}_{z_{1:m}}(X,Y)>u\Big)-\mathbb{P}_{\mathcal{D}}\Big(s^0_{z_{1:m}}(X,Y)>u\Big)\right|+\left|\mathbb{P}_{\mathcal{D}}\Big(s^{(X,Y)}_{z^*_{1:m}}(X,Y)>u\Big)-\mathbb{P}_{\mathcal{D}}\Big(s^0_{z^*_{1:m}}(X,Y)>u\Big)\right|\\ &\quad+\left|\mathbb{P}_{\mathcal{D}}\Big(s^0_{z_{1:m}}(X,Y)>u\Big)-\mathbb{P}_{\mathcal{D}}\Big(s^0_{z^*_{1:m}}(X,Y)>u\Big)\right|\\ &\overset{\text{(c)}}{\le}\frac{8L}{n}+\left|\mathbb{P}_{\mathcal{D}}\Big(s^0_{z_{1:m}}(X,Y)>u\Big)-\mathbb{P}_{\mathcal{D}}\Big(s^0_{z^*_{1:m}}(X,Y)>u\Big)\right|\\ &\overset{\text{Fact C.1}}{\le}\frac{8L}{n}+\mathbb{E}_X\left[\mathbb{P}\left(u-\left|\widehat{\mu}^0_{z_{1:m}}(X)-\widehat{\mu}^0_{z_{1:m}*}(X)\right|\le s^0_{z_{1:m}}(X,Y)\le u+\left|\widehat{\mu}^0_{z_{1:m}}(X)-\widehat{\mu}^0_{z_{1:m}*}(X)\right|\;\middle|\;X\right)\right]\\ &\overset{\text{(d)}}{\le}\frac{8L}{n}+4L_1\mathbb{E}_X\left[\left|\widehat{\mu}^0_{z_{1:m}}(X)-\widehat{\mu}^0_{z^*_{1:m}}(X)\right|\right],\end{aligned}$$

where (c) holds by Fact C.1, Assumptions 4.2 and 4.3 (similar to the arguments for (116)), and (d) makes use of Assumptions 4.2 (similar to the arguments for (117)).

To control the first term on the right-hand side of (120), we introduce the quantity below for any given X:

$$\nu_i(X)\coloneqq\mathbb{E}\left[\widehat{\mu}^0_{Z_{1:m}}(X)\,\middle|\,Z_{1:i+1}\right]-\mathbb{E}\left[\widehat{\mu}^0_{Z_{1:m}}(X)\,\middle|\,Z_{1:i}\right],\quad i=0,\ldots,m-1.$$

It is readily seen that $\{\nu_i(X)\}_{i=0}^{m-1}$ forms a martingale difference sequence. Moreover, Assumption 4.2 tells us that, for any $i=0,\ldots,m-1$ and $X\in\mathcal{X}$,

$$|\nu_i(X)|\le\sup_{z_{1:i+1}}\mathbb{E}_{Z_{i+2:m}}\left[\left|\widehat{\mu}^0_{z_{1:i}\cup z_{i+1}\cup Z_{i+2:m}}(X)-\mathbb{E}_{Z_{i+1}}\left[\widehat{\mu}^0_{z_{1:i}\cup Z_{i+1}\cup Z_{i+2:m}}(X)\right]\right|\right]\le\frac{L_2}{n}.$$

Therefore, it holds that

$$\begin{aligned}\mathop{\mathbb{E}}_{Z_{1:m},Z^*_{1:m},X}\left[\left|\widehat{\mu}^0_{Z_{1:m}}(X)-\widehat{\mu}^0_{Z^*_{1:m}}(X)\right|\right] &\le 2\mathbb{E}_{\mathcal{D}_{1:m}}\left[\left|\widehat{\mu}^0_{Z_{1:m}}(X)-\mathbb{E}_{\widetilde{Z}_{1:m}\sim\mathcal{D}_{1:m}}\left[\widehat{\mu}^0_{\widetilde{Z}_{1:m}}(X)\right]\right|\right]\\ &\le 2\mathbb{E}_{\mathcal{D}_{1:m}}\left[\left|\sum_{i=0}^{m-1}\nu_i(X)\right|\right] \le 2\left(\mathbb{E}_{\mathcal{D}_{1:m}}\left[\left(\sum_{i=0}^{m-1}\nu_i(X)\right)^2\right]\right)^{\frac{1}{2}} \qquad (121)\\ &= 2\left(\sum_{i=0}^{m-1}\mathbb{E}_{\mathcal{D}_{1:m}}\left[\nu_i(X)^2\right]\right)^{\frac{1}{2}} \le \frac{2L_2\sqrt{m}}{n},\end{aligned}$$

where the last equality holds since $\{\nu_i(X)\}$ is a martingale difference sequence. Taking (119), (120) and (121) together yields that, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{u\in\mathbb{R}}\left\{\left|\mathbb{P}_{\mathcal{D}}\left(s^{(X,Y)}_{Z_{1:m}}(X,Y)>u\;\middle|\;Z_{1:m}\right)-\mathbb{P}_{\mathcal{D}_{1:m}\times\mathcal{D}}\left(s^{(X,Y)}_{Z_{1:m}}(X,Y)>u\right)\right|\right\}\\ &\le 4L\frac{\sqrt{m\log(1/\delta)}}{n}+\frac{8L}{n}+8L\frac{\sqrt{m\log(1/\delta)}}{n}\le 16L\frac{\sqrt{m\log(1/\delta)}}{n},\end{aligned}$$

thereby concluding the proof of Lemma C.1.

C.3.2 Proof of Lemma C.2

Before proceeding, we introduce several additional convenient notations below.

- $Z_{1:m}$: we often use it in place of $Z^{\mathsf{cal}}_{1:m}$ when there is no ambiguity.
- $\widehat{\mu}_{Z_{1:m}}(\cdot)$: we remind the reader that this indicates the fitted model trained on $Z_{1:m}\cup z^{\mathsf{train}}_{m+1:n}$ (see Definition C.1).
- $Z^x_{1:m}$: this refers to $\{(x_i,Y_i)\}_{i=1}^m$, where the features are frozen to be $\{x_1,\ldots,x_m\}$; with this notation one clearly has $Z^X_{1:m}=Z_{1:m}$.
- $\widehat{\mu}_{Z^x_{1:m}}(\cdot)$: the fitted model trained on $\{(x_i,Y_i)\}_{i=1}^m$, with fixed features $\{x_i\}_{i=1}^m$ and random responses $\{Y_i\}_{i=1}^m$.
- $\widetilde{\mu}_{x_{1:m}}(x_i)$: the expected prediction of $\widehat{\mu}_{Z^x_{1:m}}(\cdot)$ w.r.t. x_i, taken over the randomness of $Y_{1:m}$, i.e.,

$$\widetilde{\mu}_{x_{1:m}}(x_i)\coloneqq\mathbb{E}_{Y_{1:m}|X_{1:m}}\left[\widehat{\mu}_{Z^x_{1:m}}(x_i)\mid X_{1:m}=x_{1:m}\right],\qquad i=1,\ldots,m, \qquad (122)$$

The proof of this lemma is organized into several steps below.

Step 1: proximity of $\widehat{\mu}_{Z^x_{1:m}}$ and its conditional expectation. Equipped with the above set of notation, we immediately note that:

- $\widehat{\mu}_{Z_{1:m}}(\cdot)$ is a function jointly dependent on the random objects $X_{1:m}$ and $Y_{1:m}$;
- For any fixed realization $x_{1:m}$, the collection $\{\widehat{\mu}_{Z^x_{1:m}}(x_i)\}_{i=1}^m$ can be viewed as a family of functions dependent on the random variables $Y_{1:m}$;
- Conditional on $X_{1:m}=x_{1:m}$, the random variables $Y_1,\ldots,Y_m$ are mutually independent.

We now make note of the following basic fact.

Claim C.2. *Recall the definition of $\widehat{\mu}_{Z^x_{1:m}}(x_i)$ and $\widetilde{\mu}_{x_{1:m}}(x_i)$ defined at the beginning of this subsection. Then for any fixed $x_{1:m}$ and any $0<\delta<1$, the event*

$$\mathcal{E}_1(x_{1:m})\coloneqq\left\{\sup_{i\in\{1,\ldots,m\}}\left\{\left|\widehat{\mu}_{Z^x_{1:m}}(x_i)-\widetilde{\mu}_{x_{1:m}}(x_i)\right|\right\}\le\frac{L_2}{n}\sqrt{m\log\frac{10m}{\delta}}\right\} \qquad (123)$$

occurs with probability at least $1-\delta/5$.

Proof. This claim follows directly by invoking McDiarmid's inequality (Lemma E.1) along with Assumption 4.3 and a union bound over $i\in\{1,\ldots,m\}$. We omit the details for brevity. □

Step 2: a surrogate empirical distribution. With Claim C.2 in place, our next step is to approximate the target empirical distribution

$$\widehat{F}_{Z_{1:m}}(u) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{s_i \le u\right\} \tag{124}$$

using a surrogate empirical distribution

$$\widetilde{F}_{Z_{1:m}}(u) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{\widetilde{s}_i \le u\right\}, \quad \text{where } \widetilde{s}_i \coloneqq \left|Y_i - \widetilde{\mu}_{X_{1:m}}(X_i)\right|, \quad i = 1, \dots, m. \tag{125}$$

In words, the fitted outcome $\widehat{\mu}_{Z_{1:m}}(X_i)$ is now replaced with $\widetilde{\mu}_{X_{1:m}}(X_i)$, the latter of which averages out the randomness over $Y_{1:m}$ (cf. (122)). On the event $\mathcal{E}_1(X_{1:m})$ defined in (123), we can bound the difference between these two quantities as follows

$$\begin{aligned}
\left|\widetilde{F}_{Z_{1:m}}(u) - \widehat{F}_{Z_{1:m}}(u)\right| &\le \frac{1}{m}\sum_{i=1}^{m}\left|\mathbb{1}\{s_i > u\} - \mathbb{1}\{\widetilde{s}_i > u\}\right| \\
&\overset{\text{(a)}}{\le} \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{u - |s_i - \widetilde{s}_i| < \widetilde{s}_i \le u + |s_i - \widetilde{s}_i|\right\} \\
&\le \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{u - |\widehat{\mu}_{Z_{1:m}}(X_i) - \widetilde{\mu}_{X_{1:m}}(X_i)| < \widetilde{s}_i \le u + |\widehat{\mu}_{Z_{1:m}}(X_i) - \widetilde{\mu}_{X_{1:m}}(X_i)|\right\} \\
&\overset{\text{(b)}}{\le} \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{u - \frac{L_2}{n}\sqrt{m\log\frac{10m}{\delta}} < \widetilde{s}_i \le u + \frac{L_2}{n}\sqrt{m\log\frac{10m}{\delta}}\right\},
\end{aligned} \tag{126}$$

where (a) holds because of Fact C.1 and (b) arises from the definition of the event $\mathcal{E}_1(X_{1:m})$.

For simplicity of notation, denote

$$\Delta_{n,m} \coloneqq \frac{L_2}{n}\sqrt{m\log\frac{10m}{\delta}} \qquad \text{and} \qquad \mathcal{B}(u,\varepsilon) \coloneqq (u-\varepsilon, u+\varepsilon], \tag{127}$$

and consider the event

$$\mathcal{E}_2(x_{1:m}) \coloneqq \left\{\sup_{u\in\mathbb{R}}\left\{\frac{1}{m}\left|\sum_{i=1}^{m}\Big(\mathbb{1}\{\widetilde{s}_i \in \mathcal{B}(u,\Delta_{n,m})\} - \mathbb{P}\big(\widetilde{s}_i \in \mathcal{B}(u,\Delta_{n,m}) \,|\, X_{1:m} = x_{1:m}\big)\Big)\right|\right\} \le 10\sqrt{\frac{\log(10/\delta)}{m}}\right\}.$$

We would like to prove that this event occurs with high probability conditional on $X_{1:m} = x_{1:m}$. Towards this end, we first observe that, for any $i = 1, \dots, m$ and any $x \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{1}\left\{\widetilde{s}_i \in \mathcal{B}(u,\Delta_{n,m})\right\} - \mathbb{P}\big(\widetilde{s}_i \in \mathcal{B}(u,\Delta_{n,m}) \,|\, X_{1:m} = x_{1:m}\big) \\
&\qquad = \Big(\mathbb{1}\{\widetilde{s}_i \le u + \Delta_{n,m}\} - \mathbb{P}\big(\widetilde{s}_i \le u + \Delta_{n,m} \,|\, X_{1:m} = x_{1:m}\big)\Big) \\
&\qquad\quad - \Big(\mathbb{1}\{\widetilde{s}_i \le u - \Delta_{n,m}\} - \mathbb{P}\big(\widetilde{s}_i \le u - \Delta_{n,m} \,|\, X_{1:m} = x_{1:m}\big)\Big),
\end{aligned}$$

which in turn implies that

$$\begin{aligned}
&\frac{1}{m}\left|\sum_{i=1}^{m}\Big(\mathbb{1}\{\widetilde{s}_i \in \mathcal{B}(u,\Delta_{n,m})\} - \mathbb{P}\big(\widetilde{s}_i \in \mathcal{B}(u,\Delta_{n,m}) \,|\, X_{1:m} = x_{1:m}\big)\Big)\right| \\
&\qquad \le \frac{1}{m}\left|\sum_{i=1}^{m}\Big(\mathbb{1}\{\widetilde{s}_i \le u + \Delta_{n,m})\} - \mathbb{P}\big(\widetilde{s}_i \le u + \Delta_{n,m} \,|\, X_{1:m} = x_{1:m}\big)\Big)\right| \\
&\qquad\quad + \frac{1}{m}\left|\sum_{i=1}^{m}\Big(\mathbb{1}\{\widetilde{s}_i \le u - \Delta_{n,m})\} - \mathbb{P}\big(\widetilde{s}_i \le u - \Delta_{n,m} \,|\, X_{1:m} = x_{1:m}\big)\Big)\right|.
\end{aligned} \tag{128}$$

In addition, it is straightforward to verify that: conditional on $X_{1:m} = x_{1:m}$, the quantity $\widetilde{s}_i$ (see (125) and (122)) is independent of $Y_{1:m} \setminus \{Y_i\}$, so that the collection $\{\widetilde{s}_i\}_{i=1}^m$ forms a set of mutually independent random variables. Hence, Lemma E.4 readily tells us that, conditional on $X_{1:m} = x_{1:m}$,

$$\sup_{u\in\mathbb{R}} \left\{ \frac{1}{m} \left| \sum_{i=1}^{m} \Big(\mathbb{1}\{\widetilde{s}_i \le u + \Delta_{m,n}\} - \mathbb{P}\big(\widetilde{s}_i \le u + \Delta_{n,m} \,|\, X_{1:m} = x_{1:m}\big) \Big) \right| \right\} \le 5\sqrt{\frac{\log(10/\delta)}{m}}$$
$$\sup_{u\in\mathbb{R}} \left\{ \frac{1}{m} \left| \sum_{i=1}^{m} \Big(\mathbb{1}\{\widetilde{s}_i \le u - \Delta_{m,n}\} - \mathbb{P}\big(\widetilde{s}_i \le u - \Delta_{n,m} \,|\, X_{1:m} = x_{1:m}\big) \Big) \right| \right\} \le 5\sqrt{\frac{\log(10/\delta)}{m}}$$

hold with probability exceeding $1 - \delta/5$, which taken together with (128) shows that for any realization $x_{1:m}$,

$$\mathbb{P}\left(\mathcal{E}_2(x_{1:m}) \mid X_{1:m} = x_{1:m}\right) \ge 1 - \frac{\delta}{5}. \tag{129}$$

Continuing from the derivation in Eqn. (126), we can now see that: on the event $\mathcal{E}_1(x_{1:m}) \cap \mathcal{E}_2(x_{1:m})$, for any $u \in \mathbb{R}$ we have

$$\begin{aligned}
\left|\widetilde{F}_{Z_{1:m}}(u) - \widehat{F}_{Z_{1:m}}(u)\right| &\le \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{\widetilde{s}_i \in \mathcal{B}(u, \Delta_{n,m})\right\} \\
&\le \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\big(\widetilde{s}_i \in \mathcal{B}(u, \Delta_{n,m}) \,|\, X_{1:m} = x_{1:m}\big) + 10\sqrt{\frac{\log(10/\delta)}{m}} \\
&\le \frac{1}{m}\sum_{i=1}^{m} 2L_1\Delta_{n,m} + 10\sqrt{\frac{\log(10/\delta)}{m}} \\
&= \frac{2L}{n}\sqrt{m\log\frac{10m}{\delta}} + 10\sqrt{\frac{\log(10/\delta)}{m}},
\end{aligned} \tag{130}$$

where the second line is valid on $\mathcal{E}_2(x_{1:m})$, and the third line makes use of the assumption stated in Assumption 4.2, along with the fact that the interval $\mathcal{B}(u, \varepsilon)$ has length 2ε. As a remark, in view of Claim C.2 and (129), we know that (130) holds with high probability when conditioned on $X_{1:m} = x_{1:m}$.

Step 3: the surrogate empirical distribution vs. the marginal coverage rate. Next, we would like to bound the discrepancy between the surrogate empirical distribution $\widetilde{F}_{Z_{1:m}}(u)$ (cf. (125)) and the marginal coverage rate

$$\overline{F}_m(u) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_{\mathcal{D}_{1:m}}\big(\widetilde{s}_i \le u\big). \tag{131}$$

Towards this end, we find it convenient to introduce two auxiliary coverage rates conditional on $X_{1:m}$ as intermediary quantities, namely,

$$\widetilde{F}_{X_{1:m}}(u) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_{Y_i|X_i}\left(\widetilde{s}_i \le u \;\middle|\; X_{1:m}\right), \quad \overline{F}_{X_{1:m}}(u) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_{Y_i|X_i}(\widetilde{s}_i \le u \mid X_i), \tag{132}$$

where for any $i = 1, \ldots, m$, we define

$$\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i \le u \mid X_i) \coloneqq \mathbb{E}_{X_{1:m}\setminus\{X_i\}}\big[\mathbb{P}(\widetilde{s}_i \le u \mid X_{1:m})\big].$$

It then follows from the triangle inequality that

$$\begin{aligned}\left|\widetilde{F}_{Z_{1:m}}(u)-\overline{F}_m(u)\right| &\le \left|\widetilde{F}_{Z_{1:m}}(u)-\widetilde{F}_{X_{1:m}}(u)\right|+\left|\widetilde{F}_{X_{1:m}}(u)-\overline{F}_{X_{1:m}}(u)\right|+\left|\overline{F}_{X_{1:m}}(u)-\overline{F}_m(u)\right| \\ &= \underbrace{\frac{1}{m}\left|\sum_{i=1}^m\Big(\mathbb{1}\{\widetilde{s}_i\le u\}-\mathbb{P}_{Y_i|X_{1:m}}\left(\widetilde{s}_i\le u\mid X_{1:m}\right)\Big)\right|}_{=:\mathcal{T}_1(u,Z_{1:m})} \\ &\quad+\underbrace{\frac{1}{m}\left|\sum_{i=1}^m\Big(\mathbb{P}_{Y_i|X_{1:m}}\left(\widetilde{s}_i\le u\,\middle|\,X_{1:m}\right)-\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_i)\Big)\right|}_{=:\mathcal{T}_2(u,X_{1:m})} \\ &\quad+\underbrace{\frac{1}{m}\left|\sum_{i=1}^m\left(\mathbb{P}_{Y_i|X_i}\left(\widetilde{s}_i\le u\,\middle|\,X_i\right)-\mathbb{P}_{\mathcal{D}_{1:m}}\left(\widetilde{s}_i\le u\right)\right)\right|}_{=:\mathcal{T}_3(u,X_{1:m})},\end{aligned} \tag{133}$$

thus leaving us with three terms to cope with.

Step 4: a bound on the first term in (133). Regarding the first term $\mathcal{T}_1(u,Z_{1:m})$ on the right-hand side of (133), consider the following event w.r.t. a given realization $x_{1:m}$:

$$\mathcal{E}_3(x_{1:m}) \coloneqq \left\{\sup_{u\in\mathbb{R}}\mathcal{T}_1(u,Z_{1:m})\le 5\sqrt{\frac{\log(5/\delta)}{m}}\right\}. \tag{134}$$

As mentioned earlier, when conditioned on $X_{1:m}=x_{1:m}$, the random variables $\{\widetilde{s}_i\}_{i=1}^m$ are mutually independent, which allows us to invoke the generalized DKW inequality (Lemma E.4) to show that

$$\mathbb{P}\left(\mathcal{E}_3(x_{1:m})\mid X_{1:m}=x_{1:m}\right)\ge 1-\delta/5. \tag{135}$$

Step 5: a bound on the second term in (133). Regarding the second term $\mathcal{T}_2(u,X_{1:m})$ on the right-hand side of (133), we first single out a few properties about $\overline{F}_{X_{1:m}}(u)$ (cf. (132)). Without loss of generality, consider two realizations $x_{1:m}$ and $x'_{1:m}$ that differ only in the first sample (i.e., $x_1\neq x'_1$). We would like to show that, for any $i\ge 1$, the function $\mathbb{P}_{Y_i|x_{1:m}}\big(\widetilde{s}_i\le u\mid X_{1:m}=x_{1:m}\big)$, viewed as a function of $x_{1:m}$, satisfies the bounded difference property. In fact, in view of the definition of $\widetilde{\mu}_{x_{1:m}}(x_i)$ (cf. (122)), we have

$$\begin{aligned}\left|\widetilde{\mu}_{x_{1:m}}(x_i)-\widetilde{\mu}_{x'_{1:m}}(x_i)\right| &= \left|\mathbb{E}_{Y_{1:m}|x_{1:m}}\left[\widehat{\mu}_{Z^x_{1:m}}(x_i)\right]-\mathbb{E}_{Y'_{1:m}|x'_{1:m}}\left[\widehat{\mu}_{Z^{x'}_{1,m}}(x_i)\right]\right| \\ &\overset{\text{(a)}}{\le} \left|\mathbb{E}_{Y_{2:m}|x_{2:m}}\left[\mathbb{E}_{Y_1|x_1}[\widehat{\mu}_{Z^x_{1:m}}(x_i)\mid Y_{2:m}]-\mathbb{E}_{Y'_1|x'_1}[\widehat{\mu}_{Z^{x'}_{1:m}}(x_i)\mid Y_{2:m}]\right]\right| \\ &\le \mathbb{E}_{Y_{2:m}|x_{2:m}}\left[\mathbb{E}_{Y_1\times Y'_1|(x_1,x'_1)}\left[\left|\widehat{\mu}_{Z^x_1\cup Z^x_{2:m}}(x_i)-\widehat{\mu}_{Z^{x'}_1\cup Z^x_{2:m}}(x_i)\right|\mid Y_{2:m}\right]\right] \\ &\overset{\text{(b)}}{\le} \mathbb{E}_{Y_{2:m}|x_{2:m}}\left[\mathbb{E}_{Y_1\times Y'_1|(x_1,x'_1)}\left[\frac{L_2}{n}\;\middle|\;Y_{2:m}\right]\right]=\frac{L_2}{n},\end{aligned} \tag{136}$$

where we denote $Z^{x'}_{1:m}\coloneqq\{(x'_i,Y'_i)\}_{i=1}^m$, $Z^x_1\coloneqq(x_1,Y_1)$, and $Z^{x'}_1\coloneqq(x'_1,Y'_1)$. Here, (b) results from Assumption 4.3. Regarding (a), it follows from the fact that $x_{2:m}=x'_{2:m}$; in particular, $Y_{2:m}$ and $Y'_{2:m}$ have the same joint distribution, and are both independent of (Y_1,Y'_1), which allow us to couple $Y_{1:m}$ and $Y'_{1:m}$, so that the two samples differ only at the first data point, i.e., $(x_1,Y_1)\neq(x'_1,Y'_1)$. Consequently, combining (136) with Assumption 4.2 reveals that, for any $i=2,\ldots,m$,

$$\begin{aligned}&\Big|\,\mathbb{P}_{Y_i|x_i}\left(|Y_i-\widetilde{\mu}_{x_{1:m}}(x_i)|>u\right)-\mathbb{P}_{Y_i|x_i}\left(\left|Y_i-\widetilde{\mu}_{x'_{1:m}}(x_i)\right|>u\right)\,\Big| \\ &\le \mathbb{P}\Big(u-\left|\widetilde{\mu}_{x_{1:m}}(x_i)-\widetilde{\mu}_{x'_{1:m}}(x_i)\right|\le|Y_i-\widetilde{\mu}_{x_{1:m}}(x_i)|\le u+\left|\widetilde{\mu}_{x_{1:m}}(x_i)-\widetilde{\mu}_{x'_{1:m}}(x_i)\right|\Big) \\ &\le 2L_1\left|\widetilde{\mu}_{x_{1:m}}(x_i)-\widetilde{\mu}_{x'_{1:m}}(x_i)\right|\le\frac{2L_1L_2}{n}=\frac{2L}{n}.\end{aligned} \tag{137}$$

Now, let us return to $\overline{F}_{x_{1:m}}(u)$. Applying the above bound (137) along with straightforward calculations reveals that, for any x,

$$\begin{aligned}\left|\widetilde{F}_{x_{1:m}}(u)-\widetilde{F}_{x'_{1:m}}(u)\right| &\le \frac{1}{m}\left|\mathbb{P}_{Y_1|x_1}\big(\widetilde{s}_1\le u\mid x_{1:m}\big)-\mathbb{P}_{Y'_1|x'_1}\big(\widetilde{s}'_1\le u\mid x'_{1:m}\big)\right| \\ &\quad+\frac{1}{m}\left|\sum_{i=2}^{m}\Big(\mathbb{P}_{Y_i|x_i}\big(\widetilde{s}_i\le u\mid x'_{1:m}\big)-\mathbb{P}_{Y_i|x_i}\big(\widetilde{s}'_i\le u\mid x'_{1:m}\big)\Big)\right| \le \frac{1}{m}+\frac{2L}{n},\end{aligned}$$

where $\widetilde{s}'_i \coloneqq \left|Y'_i-\widetilde{\mu}_{x'_{1:m}}(x'_i)\right|$. Also, note that $\mathbb{P}_{Y_i|x_i}(\widetilde{s}_i\le u\mid x_i)$ is a function of x_i only, and hence

$$\left|\,\overline{F}_{x_{1:m}}(u)-\overline{F}_{x'_{1:m}}(u)\,\right| = \frac{1}{m}\left|\,\mathbb{P}_{Y_1|x_1}(\widetilde{s}_1\le u\mid x_1)-\mathbb{P}_{Y_1|x'_1}(\widetilde{s}'_1\le u\mid x'_1)\,\right| \le \frac{1}{m}.$$

As a consequence, for any x, the function $\left|\,\widetilde{F}_{x_{1:m}}(u)-\overline{F}_{x_{1:m}}(u)\,\right|$ satisfies the bounded difference property with coefficient $\frac{2}{m}+\frac{2L}{n}$. If we define

$$\mathcal{T}_2(x_{1:m}) \coloneqq \sup_{u\in\mathbb{R}}\left\{\mathcal{T}_2(u,x_{1:m})\right\},$$

then simple computation yields

$$\begin{aligned}\mathcal{T}_2(x_{1:m})-\mathcal{T}_2(x'_{1:m}) &= \sup_{u\in\mathbb{R}}\left\{\mathcal{T}_2(u,x_{1:m})\right\}-\sup_{u'\in\mathbb{R}}\left\{\mathcal{T}_2(u',x'_{1:m})\right\} \\ &\le \sup_{u\in\mathbb{R}}\left\{\mathcal{T}_2(u,x_{1:m})-\mathcal{T}_2(u,x'_{1:m})\right\} \le \frac{2}{m}+\frac{2L}{n}.\end{aligned}$$

Thus, we can apply McDiarmid's inequality (Lemma E.1) to derive

$$\begin{aligned}\sup_{u\in\mathbb{R}}\left\{\left|\widetilde{F}_{X_{1:m}}(u)-\overline{F}_{X_{1:m}}(u)\right|\right\} &= \mathcal{T}_2(X_{1:m}) \\ &\le \mathbb{E}_{X_{1:m}}\left[\mathcal{T}_2(X_{1:m})\right]+\left(\frac{2}{m}+\frac{2L}{n}\right)\sqrt{m\log\frac{5}{\delta}}\end{aligned} \tag{138}$$

holds with probability at least $1-\delta/5$.

It then comes down to bounding $\mathbb{E}[\mathcal{T}_2(X_{1:m})]$. From the definition of $\mathcal{T}_2(X_{1:m})$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{T}_2(X_{1:m})] &= \mathbb{E}_{X_{1:m}}\left[\sup_{u\in\mathbb{R}}\left|\frac{1}{m}\sum_{i=1}^{m}\Big(\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_{1:m})-\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_i)\Big)\right|\right] \\ &\le \mathbb{E}_{X_{1:m}}\left[\sup_{u\in\mathbb{R}}\frac{1}{m}\sum_{i=1}^{m}\left|\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_{1:m})-\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_i)\right|\right] \\ &\le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{X_{1:m}}\left[\sup_{u\in\mathbb{R}}\left|\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_{1:m})-\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_i)\right|\right] \\ &\le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{X_i}\left[\mathop{\mathbb{E}}_{X_{1:m}\setminus\{X_i\}}\Big[\underbrace{\sup_{u\in\mathbb{R}}\left|\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_{1:m})-\mathbb{P}_{Y_i|X_i}(\widetilde{s}_i\le u\mid X_i)\right|}_{=:\,\mathcal{K}_i(X_{1:m})}\,\Big|\,X_i\Big]\right],\end{aligned} \tag{139}$$

which motivates us to control $\mathcal{K}_i(X_{1:m})$, $i=1,\ldots,m$. Without loss of generality, it suffices to analyze $\mathcal{K}_1(X_{1:m})$, since the same argument applies to the remaining i. Fix $X_1=x_1$, it is seen that $\widetilde{\mu}_{X_{1:m}}(\cdot)$ satisfies Assumption 4.3 with parameter L_2/n with respect to the remaining samples $X_{2:m}$. Applying Lemma C.1 reveals that, for any x_1 (it can be regarded as the target sample is (X_1,Y_1) and $X_1\sim\delta_{\{x_1\}}$ at this time),

$$\mathcal{K}_1(\{x_1\}\cup X_{2:m}) \le \frac{16L}{n}\sqrt{m\log n}$$

holds with probability at least $1 - 1/n$, which combined with (139) gives

$$\mathbb{E}[\mathcal{T}_2(X_{1:m})] \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{X_i}\left[\mathop{\mathbb{E}}_{X_{1:m}\setminus\{X_i\}}[\mathcal{K}_i(X_{1:m}) \mid X_i]\right] \le \frac{16L}{n}\sqrt{m\log n} + \frac{1}{n}.$$

Plugging this into (138) yields that the following event

$$\mathcal{E}_4 \coloneqq \left\{\sup_{u\in\mathbb{R}}\{\mathcal{T}_2(u, X_{1:m})\} \le 3\sqrt{\frac{\log(5/\delta)}{m}} + \frac{18L}{n}\sqrt{m\log\left(\frac{5}{\delta}+n\right)}\right\} \tag{140}$$

happens with probability at least $1 - \delta/5$.

Step 6: a bound on the last term in (133). We now turn attention to the last term $\mathcal{T}_3(u, X_{1:m})$ on the right-hand side of (133). Define

$$H(u, X_i) \coloneqq \mathbb{P}_{Y_i|X_i}(\widetilde{s}_i \le u \mid X_i), \qquad i = 1, \dots, m.$$

It is easily seen that the random variables $\{H(u, X_i)\}_{1\le i\le m}$ are mutually independent. Moreover, for any i and any fixed X_i, $H(u, X_i)$ is a non-decreasing function in u. Applying Lemma E.4 reveals that the event

$$\mathcal{E}_5 \coloneqq \left\{\sup_{u\in\mathbb{R}}\left\{\left|\overline{F}_{X_{1:m}}(u) - \overline{F}_m(u)\right|\right\} \le 5\sqrt{\frac{\log(5/\delta)}{m}}\right\} \tag{141}$$

happens with probability at least $1 - \delta/5$, where we remind the reader of the definitions of $\overline{F}_{X_{1:m}}$ and $\overline{F}_m$ in (132) and (131), respectively.

Step 7: putting all pieces together. To finish up, let us put together the preceding results. First, define

$$F_m(u) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_{\mathcal{D}_{1:m}}(s_i \le u). \tag{142}$$

On the event $\left(\bigcap_{i=1}^{3}\mathcal{E}_i(x_{1:m})\right)\cap\mathcal{E}_4\cap\mathcal{E}_5$, we see that for any $x\in\mathbb{R}$, it always holds that

$$\begin{aligned}
\sup_{u\in\mathbb{R}}\left|\widehat{F}_{Z_{1:m}}(u) - F_m(u)\right| &\le \left|\widehat{F}_{Z_{1:m}}(u) - \widetilde{F}_{Z_{1:m}}(u)\right| + \left|\widetilde{F}_{Z_{1:m}}(u) - F_m(u)\right| \\
&\overset{(133)}{\le} \sup_{u\in\mathbb{R}}\left\{\left|\widehat{F}_{Z_{1:m}}(u) - \widetilde{F}_{Z_{1:m}}(u)\right|\right\} + \sup_{u\in\mathbb{R}}\{\mathcal{T}_1(u, Z_{1:m})\} \\
&\quad + \sup_{u\in\mathbb{R}}\{\mathcal{T}_2(u, Z_{1:m})\} + \sup_{u\in\mathbb{R}}\{\mathcal{T}_3(u, Z_{1:m})\} + \sup_{u\in\mathbb{R}}\left\{\left|\overline{F}_m(u) - F_m(u)\right|\right\} \\
&\le \frac{2L}{n}\sqrt{m\log\frac{10m}{\delta}} + 10\sqrt{\frac{\log(10/\delta)}{m}} + 5\sqrt{\frac{\log(5/\delta)}{m}} \\
&\quad + 3\sqrt{\frac{\log(5/\delta)}{m}} + \frac{18L}{n}\sqrt{m\log\left(\frac{5}{\delta}+n\right)} + 5\sqrt{\frac{\log(5/\delta)}{m}} + \sup_{u\in\mathbb{R}}\left\{\left|\overline{F}_m(u) - F_m(u)\right|\right\} \\
&\le 24\sqrt{\frac{\log(10/\delta)}{m}} + \frac{24L}{n}\sqrt{m\log\left(\frac{10m}{\delta}+n\right)}.
\end{aligned}$$

To justify the last inequality, we observe that, for any u,

$$\left|\overline{F}_m(u) - F_m(u)\right| \le \frac{1}{m}\sum_{i=1}^{m}\left|\mathbb{P}(\widetilde{s}_i \le u) - \mathbb{P}(s_i \le u)\right|$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\left\{\mathbb{P}\Big(\widetilde{s}_i - |\widetilde{\mu}_{X_{1:m}}(X_i) - \widehat{\mu}_{Z_{1:m}}(X_i)| \leq u\Big) - \mathbb{P}\Big(\widetilde{s}_i + |\widetilde{\mu}_{X_{1:m}}(X_i) - \widehat{\mu}_{Z_{1:m}}(X_i)| \leq u\Big)\right\}$$
$$\overset{\text{(a)}}{\leq} \frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\left(\widetilde{s}_i \in \mathcal{B}\left(u, \Delta\right)\right) + \frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\big(|\widetilde{\mu}_{X_{1:m}}(X_i) - \widehat{\mu}_{Z_{1:m}}(X_i)| > \Delta\big) \leq \frac{4L}{n}\sqrt{m\log n} + \frac{1}{n},$$

where $\Delta \coloneqq \frac{2L_2}{n}\sqrt{m\log n}$, and (a) follows by invoking the same argument as in the analysis of $\mathcal{E}_1(X_{1:m})$. By combining our uniform high-probability bounds on $\mathcal{E}_i(x_{1:m})$ for $i = 1, 2, 3$ given any $X_{1:m} = x_{1:m}$, and applying the high-probability bound of $\mathcal{E}_4$ and $\mathcal{E}_5$ as well as the union bound, we arrive at

$$\mathbb{P}_{\mathcal{D}_{1:m}}\left(\left\{\left(\bigcap_{i=1}^{3}\mathcal{E}_i(X_{1:m})\right)\cap\mathcal{E}_4\cap\mathcal{E}_5\right\}^{\mathrm{c}}\right) = \mathbb{P}_{\mathcal{D}_{1:m}}\left(\left(\bigcup\mathcal{E}_i(X_{1:m})^{\mathrm{c}}\right)\cup\mathcal{E}_4^{\mathrm{c}}\cup\mathcal{E}_5^{\mathrm{c}}\right)$$
$$\leq \sum_{i=1}^{3}\mathbb{P}_{\mathcal{D}_{1:m}}\left(\mathcal{E}_i(X_{1:m})^{\mathrm{c}}\right) + \frac{\delta}{5} + \frac{\delta}{5} = \sum_{i=1}^{3}\mathbb{E}_{X_{1:m}}\left[\mathbb{P}_{Y_{1:m}|X_{1:m}}\left(\mathcal{E}_i(X_{1:m})^{\mathrm{c}} \mid X_{1:m}\right)\right] + \frac{2\delta}{5}$$
$$\leq \frac{3\delta}{5} + \frac{2\delta}{5} = \delta.$$

This completes the proof of Lemma C.2.

C.3.3 Proof of Lemma C.3

The proof closely follows the arguments in the proof of Lemmas B.1 and B.2. The only difference lies in that, for the smooth drift setting, inequality (71) in the pretrained-score setting needs to be modified as follows:

$$A_k = \frac{1}{|\mathcal{I}_k|}\sum_{l\in\mathcal{I}_k}\left(\alpha - \mathbb{P}(Y_{n,r,l}\notin\mathcal{C}_{n,r}(X_{n,r,l})\mid\mathcal{C}_{n,r})\right)$$
$$\leq \frac{1}{|\mathcal{I}_k|}\sum_{l\in\mathcal{I}_k}\left(\mathbb{P}\big(Y_{n,r,i_{k-1}}\notin\mathcal{C}_{n,r}(X_{n,r,i_{k-1}})\mid\mathcal{C}_{n,r}\big) - \mathbb{P}(Y_{n,r,l}\notin\mathcal{C}_{n,r}(X_{n,r,l}\mid\mathcal{C}_{n,r}))\right)$$
$$= \frac{1}{|\mathcal{I}_k|}\sum_{l\in\mathcal{I}_k}\sum_{i=i_{k-1}}^{l-1}\left\{\mathbb{P}\big(Y_{n,r,i}\notin\mathcal{C}_{n,r}(X_{n,r,i})\mid\mathcal{C}_{n,r}\big) - \mathbb{P}\big(Y_{n,r,i+1}\notin\mathcal{C}_{n,r}(X_{n,r,i+1})\mid\mathcal{C}_{n,r}\big)\right\} \tag{143}$$

for any even k, where we invoke the same arguments as in (71), albeit with notation adjusted to the full conformal setting. We observe that, for any $i \in i_{k-1}, \ldots, l-1$,

$$\mathbb{P}\big(Y_{n,r,i}\notin\mathcal{C}_{n,r}(X_{n,r,i})\mid\mathcal{C}_{n,r}\big) - \mathbb{P}\big(Y_{n,r,i+1}\notin\mathcal{C}_{n,r}(X_{n,r,i+1})\mid\mathcal{C}_{n,r}\big)$$
$$= \mathbb{E}\big[\mathbb{1}\{Y_{n,r,i}\notin\mathcal{C}_{n,r}(X_{n,r,i})\}\mid\mathcal{C}_{n,r}\big] - \mathbb{E}\big[\mathbb{1}\{Y_{n,r,i+1}\notin\mathcal{C}_{n,r}(X_{n,r,i+1})\}\mid\mathcal{C}_{n,r}\big]$$
$$\leq \sup_{h\in\mathcal{M}([0,1])}\left\{\mathbb{E}\big[h(Z_{n,r,i})\big] - \mathbb{E}\big[h(Z_{n,r,i+1})\big]\right\} = \mathsf{TV}\big(Z_{n,r,i}, Z_{n,r,i+1}\big),$$

where $\mathcal{M}([0,1])$ denotes all measurable functions of $Z \in \mathcal{X}\times\mathbb{R}$ that are bounded in $[0,1]$. Accordingly, the bound (72) on A_k in the pretrained-score case can now be replaced by

$$A_k \leq \frac{1}{|\mathcal{I}_k|}\sum_{l\in\mathcal{I}_k}\sum_{i=i_{k-1}}^{l-1}\mathsf{TV}(Z_{n,r,i}, Z_{n,r,i+1}) \leq \sum_{i=i_{k-1}}^{i_k-1}\mathsf{TV}(Z_{n,r,i}, Z_{n,r,i+1})$$

for each even k. As a result, under smooth drift, the complexity measure used to control the cumulative regret in stage n should now be $\mathsf{TV}_n^{\mathsf{stage}}$ (cf. (102)) rather than $\mathsf{KS}_n^{\mathsf{stage}}$ (cf. (40)).

The remaining arguments are the same as for Lemmas B.1 and B.2, and are hence omitted for brevity.

C.3.4 Proof of Lemma C.4

For any n, given that $\mathcal{B}_n = \mathcal{A}_{n,r_n} \cap \mathcal{G}(\tau_{n,r_n-1}, \tau_{n,r_n})$, one has

$$\mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{B}_n^{\mathrm{c}}\}\big] \le \mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{A}_{n,r_n}^{\mathrm{c}}\}\big] + \mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{G}(\tau_{n,r_n-1}, \tau_{n,r_n})^{\mathrm{c}}\}\big], \tag{144}$$

leaving us with two terms to control.

We first bound $\mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{A}_{n,r_n}^{\mathrm{c}}\}\big]$. Recognizing that $n \le \tau_{n,r_n-1} < \tau_{n,r_n}$ and $\tau_{n+1} \le 4\tau_{n,r_n}$ (since the round lengths grow geometrically), we can deduce that

$$\begin{aligned}
\mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{A}_{n,r_n}^{\mathrm{c}}\}\big] &\le \sum_{1\le k<m\le T} \mathbb{E}\Big[4m\,\mathbb{1}\big\{\mathcal{A}_{n,r_n}^{\mathrm{c}}\big\}\mathbb{1}\{\tau_{n,r_n-1}=k;\ \tau_{n,r_n}=m\}\Big] \\
&= \sum_{1\le k<m\le T} 4m\mathbb{E}\Big[\mathbb{1}\{\tau_{n,r_n-1}=k;\ \tau_{n,r_n}=m\}\mathbb{E}\big[\mathbb{1}\big\{\mathcal{A}_{n,r_n}^{\mathrm{c}}\big\} \mid Z_{1:m-1}\big]\Big] \\
&\le \sum_{1\le k<m\le T} 4m\mathbb{E}\Big[\mathbb{1}\{\tau_{n,r_n-1}=k;\tau_{n,r_n}=m\}\Big(\sum_{m\le i<j<\infty}\mathbb{P}\big(\mathcal{A}(k,m;i,j)^{\mathrm{c}} \mid Z_{1:m-1}\big)\Big)\Big] \\
&\overset{(99)}{\le} \sum_{1\le k<m\le T} 4m\mathbb{E}\Big[\mathbb{1}\{\tau_{n,r_n-1}=k;\ \tau_{n,r_n}=m\}\Big(\sum_{m\le i<j<\infty} j^{-8}\Big)\Big] \\
&\le \sum_{1\le k<m\le T} 4m\mathbb{E}\Big[\mathbb{1}\{\tau_{n,r_n-1}=k;\ \tau_{n,r_n}=m\}\big(m^{-6}/42\big)\Big] = \frac{2}{21}\mathbb{E}\big[\tau_{n,r_n}^{-5}\big] \le \frac{2}{21n^5}.
\end{aligned} \tag{145}$$

Next, we turn attention to the term $\mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{G}(\tau_{n,r_n-1}, \tau_{n,r_n})^{\mathrm{c}}\}\big]$. Using $\tau_{n+1} \le 16\tau_{n,r_n-1}$ (which is again due to the geometric growth of the round lengths), we obtain

$$\begin{aligned}
\mathbb{E}\big[\tau_{n+1}\mathbb{1}\{\mathcal{G}(\tau_{n,r_n-1}, \tau_{n,r_n})^{\mathrm{c}}\}\big] &\le \sum_{k=1}^{T}\mathbb{E}\Big[16k\,\mathbb{1}\{\tau_{n,r_n-1}=k\}\mathbb{1}\Big\{\mathcal{G}(k,\tau_{n,r_n})^{\mathrm{c}}\Big\}\Big] \\
&\le \sum_{k=1}^{T}\mathbb{E}\Big[16k\,\mathbb{1}\{\tau_{n,r_n-1}=k\}\Big(\sum_{m=k+1}^{\infty}\mathbb{P}\big(\mathcal{G}(k,m)^{\mathrm{c}} \mid Z_{1:k-1}\big)\Big)\Big] \\
&\overset{\text{(a)}}{\le} \sum_{k=1}^{T}\mathbb{E}\Big[16k\,\mathbb{1}\{\tau_{n,r_n-1}=k\}\Big(\sum_{m=k+1}^{\infty} m^{-4}\Big)\Big] \\
&\le \sum_{k=1}^{T} 6\mathbb{E}\Big[k^{-2}\mathbb{1}\{\tau_{n,r_n-1}=k\}\Big] \le 6\mathbb{E}[\tau_{n,r_n-1}^{-2}] \le \frac{6}{n^2},
\end{aligned} \tag{146}$$

where (a) follows from Proposition 4.1 with $\delta = m^{-4}$.

Taking together (144)–(146) thus completes the proof.

C.3.5 Proof of Lemma C.5

Recall the algorithm procedure of DriftOCP-full: in stage n ($\le N-1$), the distribution shift is detected in round r_n, and this round contains t_n iterations. Then in light of our drift detection subroutine (see Algorithm 3), there exists some $j_n \in [t_n]$ such that:

$$\left|\sum_{l=j_n}^{t_n}\big(\mathbb{1}\{Y_{n,r_n,l} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,l})\} - \alpha\big)\right| > 10\sqrt{t_n - j_n + 1}\log^3(40\tau_{n,r_n}).$$

Then on the event $\mathcal{A}_{n,r_n}$ (cf. (96)), it holds that

$$\left|\sum_{l=j_n}^{t_n}\big(\mathbb{P}(Y_{n,r_n,l} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,l}) \,|\, \mathcal{C}_{n,r_n}) - \alpha\big)\right| \ge \left|\sum_{l=j_n}^{t_n}\big(\mathbb{1}\{Y_{n,r_n,l} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,l})\} - \alpha\big)\right|$$

$$
\begin{aligned}
&- \left| \sum_{l=j_n}^{t_n} \Big(\mathbb{1}\big\{ Y_{n,r_n,l} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,l}) \big\} - \mathbb{P}\big(Y_{n,r_n,l} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,l}) \,|\, \mathcal{C}_{n,r_n} \big) \Big) \right| \\
&> 10\sqrt{t_n - j_n + 1}\log^3(40\tau_{n,r_n}) - 2\sqrt{(t_n - j_n + 1)\log(2\tau_{n+1,1})} \\
&\geq 8\sqrt{t_n - j_n + 1}\log^3(40\tau_{n,r_n}). \qquad (147)
\end{aligned}
$$

As a consequence, if we define

$$
B_n := \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \Big(\mathbb{P}\big(Y_{n,r_n,l} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,l}) \,|\, \mathcal{C}_{n,r_n} \big) - \alpha \Big), \qquad (148)
$$

then (147) implies that

$$
\frac{\sqrt{t_n - j_n + 1}}{8\log^3(40\tau_{n,r_n})} |B_n| \geq 1. \qquad (149)
$$

The next step is to analyze the quantity B_n defined in (148). Note that $\tau_{n,r_n} - \tau_{n,r_n-1} = T_{r_n-1}$. Then on the event $\mathcal{B}_n$ defined in (104)—more precisely, on the event $\mathcal{G}(\tau_{n,r_n-1}, \tau_{n,r_n})$ defined in (103)—we have

$$
\begin{aligned}
|B_n| &\leq \Big| \mathbb{P}\big(Y_{n,r_n,1} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,1}) \,|\, \mathcal{C}_{n,r_n} \big) - \alpha \Big| \\
&\quad + \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \Big| \mathbb{P}\big(Y_{n,r_n,l} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,l}) \,|\, \mathcal{C}_{n,r_n} \big) - \mathbb{P}\big(Y_{n,r_n,1} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,1}) \,|\, \mathcal{C}_{n,r_n} \big) \Big| \\
&\overset{\text{(a)}}{\leq} \Big| \mathbb{P}\big(Y_{n,r_n,1} \notin \mathcal{C}_{n,r_n}(X_{n,r_n,1}) \,|\, \mathcal{C}_{n,r_n} \big) - \alpha \Big| + \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \mathsf{TV}(Z_{n,r_n,l}, Z_{n,r_n,1}) \\
&\overset{\text{(b)}}{\leq} \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \left(2^6 \sqrt{\frac{\log(40\tau_{n,r_n})}{T_{r_n-1}}} + \frac{2^7 L}{\tau_{n,r_n}} \sqrt{T_{r_n-1}\log(40\tau_{n,r_n})} \right) \\
&\quad + \frac{1}{t_n - j_n + 1} \sum_{l=j_n}^{t_n} \sum_{j=1}^{l-1} \mathsf{TV}\big(Z_{n,r_n,j}, Z_{n,r_n,j+1} \big) + \frac{1}{T_{r_n-1}} \sum_{l=1}^{T_{r_n-1}} \sum_{i=l}^{T_{r_n-1}} \mathsf{TV}\big(Z_{n,r_n-1,i}, Z_{n,r_n-1,i+1} \big) \\
&\leq 2^6 \sqrt{\frac{\log(40\tau_{n,r_n})}{T_{r_n-1}}} + \frac{2^7 L}{\tau_{n,r_n}} \sqrt{T_{r_n-1}\log(40\tau_{n,r_n})} + \mathsf{TV}_n^{\mathsf{tail}}, \qquad (150)
\end{aligned}
$$

where (b) is valid on $\mathcal{B}_n$, and (a) results from the fact that, for any $l = j_n, \ldots, t_n$,

$$
\begin{aligned}
&\Big| \mathbb{P}\big(Y_{n,r,l} \notin \mathcal{C}_{n,r}(X_{n,r,l}) \mid \mathcal{C}_{n,r} \big) - \mathbb{P}\big(Y_{n,r,1} \notin \mathcal{C}_{n,r}(X_{n,r,1}) \mid \mathcal{C}_{n,r} \big) \Big| \\
&\qquad = \Big| \mathbb{E}\big[\mathbb{1}\{ Y_{n,r,l} \notin \mathcal{C}_{n,r}(X_{n,r,l}) \} \mid \mathcal{C}_{n,r} \big] - \mathbb{E}\big[\mathbb{1}\{ Y_{n,r,1} \notin \mathcal{C}_{n,r}(X_{n,r,1}) \} \mid \mathcal{C}_{n,r} \big] \Big| \\
&\qquad \leq \sup_{h \in \mathcal{M}([0,1])} \Big\{ \mathbb{E}\big[h(Z_{n,r,l}) \big] - \mathbb{E}\big[h(Z_{n,r,1}) \big] \Big\} = \mathsf{TV}\big(Z_{n,r,l}, Z_{n,r,1} \big),
\end{aligned}
$$

where $\mathcal{M}([0,1])$ denotes the set of all measurable functions $h : \mathcal{X} \times \mathbb{R} \to [0,1]$.

Regarding the first term on the right-hand side of (150), one can apply (149) together with a little algebra to show that: when $\tau_{n,r_n-1} \geq 2$:

$$
2^6 \sqrt{\frac{\log(40\tau_{n,r_n})}{T_{r_n-1}}} \overset{(149)}{\leq} \frac{2^6}{8\log^{5/2}(40\tau_{n,r_n})} \sqrt{\frac{t_n - j_n + 1}{T_{r_n-1}}} |B_n| \leq \frac{1}{2} |B_n| . \qquad (151)
$$

Combine this with (150) and rearrange terms to reach

$$
|B_n| \leq \frac{2^8 L}{\tau_{n,r_n}} \sqrt{T_{r_n-1}\log(40\tau_{n,r_n})} + 2\mathsf{TV}_n^{\mathsf{tail}}.
$$

To finish up, recall that $\mathcal{H}_n \coloneqq \left\{T_{r_n-1}\sqrt{\log(40\tau_{n,r_n})} \le \frac{\tau_{n,r_n}}{256L}\right\}$. On the event $\mathcal{B}_n \cap \mathcal{H}_n$, combining the above expression with (149) allows us to establish the following inequality:

$$\begin{aligned} 2 + 2\sqrt{t_n - j_n + 1}\,\mathsf{TV}_n^{\mathsf{tail}} &\overset{\mathcal{H}_n}{\ge} \sqrt{t_n - j_n + 1}\left(\frac{2^8 L}{\tau_{n,r_n}}\sqrt{T_{r_n-1}\log(40\tau_{n,r_n})} + 2\mathsf{TV}_n^{\mathsf{tail}}\right) \\ &\ge \sqrt{t_n - j_n + 1}\,|B_n| \overset{(149)}{\ge} 8, \end{aligned}$$

which in turn implies that

$$\sqrt{t_n - j_n + 1}\,\mathsf{TV}_n^{\mathsf{tail}}\mathbb{1}\{\mathcal{B}_n \cap \mathcal{H}_n\} \ge 3 \cdot \mathbb{1}\{\mathcal{B}_n \cap \mathcal{H}_n\}. \tag{152}$$

This immediately concludes the proof of this lemma.

C.4 Proof of Theorem 4.2

Throughout this subsection, we consider the case where no features $\{X_t\}_{t=1}^T$ are observed; instead, only the response Y_t is available at time t. Under this simplification, the set-valued functions $\{\mathcal{C}_t(\cdot)\}_{t\ge1}$ induced by algorithm $\pi = \{\pi_t\}_{t\ge1}$ in (32) admit the simpler representation

$$\mathcal{C}_t = \begin{cases} \pi_1(U), & \text{if } t = 1, \\ \pi_t(Y_{t-1}, \ldots, Y_1, U), & \text{if } t \ge 2, \end{cases} \tag{153}$$

where $\pi_t : \mathbb{R}^{t-1} \times [0,1] \to \mathcal{B}(\mathbb{R})$ for $t \ge 2$. In this setting, $\mathcal{C}_t$ is uniquely determined by $Y_{1:t-1}$ and U; accordingly, we often write it as $\mathcal{C}(Y_{1:t-1}, U)$ as long as it is clear from the context.

Our proof of Theorem 4.2 is organized into several steps, presented below.

Step 1: constructing a class of distributions with piecewise flat density. We begin by introducing the distribution class $\mathcal{I}$, constructed as follows.

- First, divide the interval $[0,1]$ into k subintervals:

$$I_j \coloneqq \left[\frac{j-1}{k}, \frac{j}{k}\right), \quad j = 1, \ldots, k-1; \quad I_k \coloneqq \left[\frac{k-1}{k}, 1\right]. \tag{154}$$

- For any given sequence $V_1, \ldots, V_k \in \{-1, 1\}$, generate a distribution with probability density function

$$f(y \mid V_{1:k}) \coloneqq \sum_{j=1}^k f_j(y), \tag{155a}$$

where $f_j(\cdot)$ is nonzero only within the subinterval I_j as follows:

$$f_j(y) \propto (1 + \epsilon V_j)\,\mathbb{1}_{I_j}(y),$$

with ϵ a small positive constant to be specified shortly, and $\mathbb{1}_{I_j}(y) \coloneqq \mathbb{1}\{y \in I_j\}$ the indicator function of I_j. The normalization constant can then be computed as

$$\sum_{j=1}^k \int_{I_j} (1 + \epsilon V_j)\,\mathrm{d}y = \sum_{j=1}^k \frac{1 + \epsilon V_j}{k} =: 1 + \epsilon\overline{V},$$

thereby allowing us to express

$$f_j(y) = \frac{(1 + \epsilon V_j)\mathbb{1}_{I_j}(y)}{1 + \epsilon\overline{V}}, \quad j = 1, 2, \ldots, k. \tag{155b}$$

- Accordingly, we construct a distribution class as follows

$$\mathcal{I} \coloneqq \left\{ f(y \mid V_{1:k}) \,\middle|\, V_{1:k} \in \{-1, 1\}^k \right\}. \tag{156}$$

Step 2: constructing a family of distribution sequences contained in $\mathcal{L}_3(N^{\mathsf{cp}})$ and $\mathcal{L}_4(\mathsf{TV}_T)$. With $\mathcal{I}$ in place, we would like to construct a family $\mathcal{L}'$ of distribution sequences such that it is composed of all $\{\mathcal{D}_t\}_{t=1}^T$ satisfying the following two conditions:

1. $\mathcal{D}_t \in \mathcal{I}$ for every $t = 1, \ldots, T$;
2. For every $l = 1, \ldots, m+1$, it holds that $\mathcal{D}_t = \mathcal{D}_{t+1}$ for any t obeying $(l-1)\lfloor T/m \rfloor + 1 \le t < \min\{l\lfloor T/m \rfloor, T\}$. In other words, the distributions are identical within each segment

$$\mathcal{T}_l \coloneqq \big[(l-1)\lfloor T/m \rfloor + 1, \min\{l\lfloor T/m \rfloor, T\}\big], \tag{157}$$

where each batch $\mathcal{T}_l$ (except for the $(m+1)$-th batch) contains $\lfloor T/m \rfloor$ time instances.

We now verify that $\mathcal{L}'$ belongs to both $\mathcal{L}_3(N^{\mathsf{cp}})$ and $\mathcal{L}_4(\mathsf{TV}_T)$ under an appropriate choice of parameters.

- Regarding the change-point setting, it is clearly seen that $\mathcal{L}' \subset \mathcal{L}_3(N^{\mathsf{cp}})$ when $m = N^{\mathsf{cp}}$.
- Turning to the smooth drift setting, we claim that $\mathcal{L}' \subset \mathcal{L}_4(\mathsf{TV}_T)$ for sufficiently small ϵ. To justify this, consider any two distributions $\mathcal{D}_1, \mathcal{D}_2 \in \mathcal{I}$, and suppose that their densities can be written as $p_{\mathcal{D}_i}(y) = f(y \mid V_{1:k}^{(i)})$ for $i = 1, 2$. Assuming that $\epsilon \le 1/2$, we can calculate

$$\begin{aligned}
\mathsf{TV}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2}\int_0^1 \left| f_1(y \mid V_{1:k}^{(1)}) - f_1(y \mid V_{1:k}^{(2)}) \right| \mathrm{d}y = \frac{1}{2}\sum_{j=1}^k \int_{I_j} \left| \frac{1+\epsilon V_j^{(1)}}{1+\epsilon \overline{V}^{(1)}} - \frac{1+\epsilon V_j^{(2)}}{1+\epsilon \overline{V}^{(2)}} \right| \mathrm{d}y \\
&\le \sum_{j=1}^k \frac{2}{k} \left| \epsilon\big(V_j^{(1)} + \overline{V}^{(2)} - V_j^{(2)} - \overline{V}^{(1)}\big) + \epsilon^2\big(V_j^{(1)}\overline{V}^{(2)} - V_j^{(2)}\overline{V}^{(1)}\big) \right| \le \sum_{j=1}^k \frac{2}{k}(4\epsilon + 2\epsilon^2) \le 10\epsilon,
\end{aligned} \tag{158}$$

where the inequalities in the last line result from $\epsilon \le 1/2$ and $\max_{j\in[k], i=1,2}\{|V_j^{(i)}|, |\overline{V}^{(i)}|\} \le 1$. As a result, if we take $\epsilon \le \min\{\mathsf{TV}_T/(20m), 1/2\}$, then for any $\{\mathcal{D}_t\}_{t=1}^T$ we have

$$\begin{aligned}
\sum_{t=1}^{T-1} \mathsf{TV}(\mathcal{D}_t, \mathcal{D}_{t+1}) &= \sum_{j=1}^m \mathsf{TV}\left(\mathcal{D}_{(j-1)\lfloor \frac{T}{m} \rfloor + 1}, \mathcal{D}_{j\lfloor \frac{T}{m} \rfloor + 1}\right) \\
&\le m \sup_{\mathcal{D}_1, \mathcal{D}_2 \in \mathcal{I}} \mathsf{TV}(\mathcal{D}_1, \mathcal{D}_2) \overset{(158)}{\le} 20m\epsilon \le \mathsf{TV}_T,
\end{aligned}$$

thus ensuring that $\mathcal{L}' \subset \mathcal{L}_4(\mathsf{TV}_T)$.

Step 3: establishing a general regret lower bound. We now look at the cumulative regret within each batch $\mathcal{T}_i$ $(i = 1, \ldots, m+1)$ defined in (157). To this end, we first establish the following lower bound on the coverage gap for a single time point; the proof is deferred to Section C.4.1.

Lemma C.6. *Suppose that $0 < \alpha \le 1/2$, $k \ge \frac{256K}{\alpha}$ and $\epsilon \le \min\left\{\frac{\alpha^{5/2}}{200}, \frac{1}{64}\sqrt{\frac{\alpha k}{n\log(2nk/\alpha^2)}}\right\}$. Consider any $n \ge 1$. Then, for any admissible algorithm $\pi \in \mathcal{P}_K$ (cf. (33)), the set-value mapping $\mathcal{C}$ induced by π satisfies*

$$\frac{1}{|\mathcal{I}|}\sum_{\mathcal{D}\in\mathcal{I}} \left\{ \mathop{\mathbb{E}}_{Y_{1:n}\sim\mathcal{D}^n,\, U\sim p_U} \left[\left| \mathbb{P}\big(Y_{n+1} \in \mathcal{C}(Y_{1:n}, U) \,|\, Y_{1:n}, U\big) - (1-\alpha) \right| \right] \right\} \ge \frac{\alpha^{\frac{5}{2}}\epsilon}{144\sqrt{k}} - \frac{\alpha^6}{4n^3k^3}, \tag{159}$$

where U is independently drawn from an arbitrary continuous distribution with density function $p_U(\cdot)$, and $\mathcal{I}$ is defined in Step 1 (with the parameter ϵ).

With this intermediate result in hand, we can now analyze the cumulative coverage gap within each batch $\mathcal{T}_i$ (cf. (157)). In fact, letting $\tau_i \coloneqq (i-1)\lfloor T/m \rfloor + 1$, one can write

$$\sum_{t\in\mathcal{T}_i} \mathbb{E}\Big[\left| \mathbb{P}\big(Y_t \in \mathcal{C}(Y_{1:t-1}, U) \,|\, Y_{1:t-1}, U\big) - (1-\alpha) \right| \Big]$$

$$= \sum_{t\in\mathcal{T}_i} \mathbb{E}\left[\left| \mathbb{P}\Big(Y_t \in \mathcal{C}(Y_{1:\tau_i-1}, Y_{\tau_i:t-1}, U) \,\Big|\, Y_{1:t-1}, U\Big) - (1-\alpha) \right|\right].$$

Note that by construction, the distribution selected for each batch is independent of the distributions assigned to all preceding batches. Therefore, we can view $(Y_{1:\tau_i-1}, U)$ jointly as a new random variable $\widetilde{U} \sim P_{\widetilde{u}}$ for some distribution $P_{\widetilde{u}}$, which is independent of all randomness within $\mathcal{T}_i$. Consequently, we shall write the prediction set $\mathcal{C}(Y_{1:\tau_i-1}, Y_{\tau_i:t-1}, U)$ as $\mathcal{C}(Y_{\tau_i:t-1}, \widetilde{U})$ in the sequel (as long as it is clear from the context), in order to underscore the role of $Y_{\tau_i:t-1}$. Armed with this simplified notation, we define, for any given distribution $\mathcal{D}$ and any index $i \in [m+1]$, the cumulative regret over batch $\mathcal{T}_i$ as

$$\begin{aligned}\mathsf{regret}_\pi(\mathcal{D}, \mathcal{T}_i) &:= \sum_{t\in\mathcal{T}_i} \mathbb{E}\left[\left|\mathbb{P}\Big(Y_t \in \mathcal{C}(Y_{1:\tau_i-1}, Y_{\tau_i:t-1}, U) \,\Big|\, Y_{1:t-1}, U\Big) - (1-\alpha)\right|\right] \\ &= \sum_{t\in\mathcal{T}_i} \mathbb{E}\left[\left|\mathbb{P}\Big(Y_t \in \mathcal{C}(Y_{\tau_i:t-1}, \widetilde{U}) \,\Big|\, Y_{\tau_i:t-1}, \widetilde{U}\Big) - (1-\alpha)\right|\right].\end{aligned}$$

Applying Lemma C.6 for each time t, we reach

$$\begin{aligned}\frac{1}{|\mathcal{I}|}\sum_{\mathcal{D}\in\mathcal{I}} \mathsf{regret}_\pi(\mathcal{D}, \mathcal{T}_i) &= \frac{1}{|\mathcal{I}|}\sum_{\mathcal{D}\in\mathcal{I}}\sum_{t\in\mathcal{T}_i} \mathbb{E}\Big[\,\Big|\mathbb{P}\Big(Y_t \in \mathcal{C}(Y_{\tau_i:t-1}, \widetilde{U}) \,\Big|\, Y_{\tau_i:t-1}, \widetilde{U}\Big) - (1-\alpha)\Big|\,\Big] \\ &= \sum_{t\in\mathcal{T}_i}\left\{\frac{1}{|\mathcal{I}|}\sum_{\mathcal{D}\in\mathcal{I}} \mathbb{E}\Big[\,\Big|\mathbb{P}\Big(Y_t \in \mathcal{C}(Y_{\tau_i:t-1}, \widetilde{U}) \,\Big|\, Y_{\tau_i,t-1}, \widetilde{U}\Big) - (1-\alpha)\Big|\,\Big]\right\} \qquad (160) \\ &\geq \sum_{t\in\mathcal{T}_i}\left(\frac{\alpha^{\frac{5}{2}}\epsilon}{144\sqrt{k}} - \frac{\alpha^6}{4(T/m)^3k^3}\right) = \left(\frac{\alpha^{\frac{5}{2}}\epsilon}{144\sqrt{k}} - \frac{\alpha^6}{4(T/m)^3k^3}\right)|\mathcal{T}_i|,\end{aligned}$$

provided that

$$\epsilon \leq \min\left\{\frac{\alpha^{\frac{5}{2}}}{200}, \frac{1}{64}\sqrt{\frac{\alpha mk}{T\log(Tk/\alpha m)}}\right\}. \qquad (161)$$

Now, putting all batches together yields the following regret lower bound: for any algorithm $\pi \in \mathcal{P}_K$,

$$\begin{aligned}\mathsf{regret}_\pi(\mathcal{L}', T, K) &= \sup_{\{\mathcal{D}_t\}_{t=1}^T \in \mathcal{L}'} \mathsf{regret}_\pi\big(\mathcal{D}_{1:T}, T\big) = \sum_{i=1}^{m+1} \sup_{\mathcal{D}\in\mathcal{I}} \mathsf{regret}_\pi(\mathcal{D}, \mathcal{T}_i) \\ &\geq \sum_{i=1}^{m+1} \frac{1}{|\mathcal{I}|}\sum_{\mathcal{D}\in\mathcal{I}} \mathsf{regret}_\pi(\mathcal{D}, \mathcal{T}_i) \overset{(160)}{\geq} \left(\frac{\alpha^{\frac{5}{2}}\epsilon}{144\sqrt{k}} - \frac{\alpha^6}{4n^3k^3}\right)T.\end{aligned} \qquad (162)$$

Step 4: instantiating the general lower bound to two drift settings. It remains to connect the above lower bound to the two distribution-drift settings, which we discuss separately below.

- *The change-point setting.* In this drift scenario, setting $m = N^{\mathsf{cp}}$ ensures that $\mathcal{L}' \subset \mathcal{L}_3(N^{\mathsf{cp}})$ (as discussed in Step 2). Then, taking $k = \frac{256K}{\alpha}$ and $\epsilon = \min\left\{\frac{\alpha^{5/2}}{200}, \frac{1}{64}\sqrt{\frac{\alpha k(N^{\mathsf{cp}}+1)}{T\log(Tk/\alpha)}}\right\}$ in (162) yields

$$\begin{aligned}\mathsf{regret}_\pi\big(\mathcal{L}_3(N^{\mathsf{cp}}), T, K\big) \geq \mathsf{regret}_\pi\big(\mathcal{L}', T, K\big) &\geq \frac{1}{300}\min\left\{\frac{\alpha^5 T}{200\sqrt{k}}, \frac{\alpha^3}{16}\sqrt{\frac{(N^{\mathsf{cp}}+1)T}{\log(Tk/\alpha)}}\right\} \\ &= \widetilde{\Omega}\left(\min\left\{\frac{T}{\sqrt{K}}, \sqrt{(N^{\mathsf{cp}}+1)T}\right\}\right).\end{aligned}$$

- *The smooth drift setting.* In order to simultaneously satisfy (161) and $\mathcal{L}' \subset \mathcal{L}_4(\mathsf{TV}_T)$ (which needs $\epsilon \leq \min\{\mathsf{TV}_T/(20m), 1/2\}$ as discussed in Step 2), we take ϵ to be

$$\epsilon = \min\left\{\frac{\alpha^{\frac{5}{2}}}{200}, \frac{\mathsf{TV}_T}{20m}, \frac{1}{64}\sqrt{\frac{\alpha mk}{T\log(Tk/\alpha m)}}\right\}.$$

Substitution into inequality (162) leads to

$$\mathsf{regret}_\pi\big(\mathcal{L}_4(\mathsf{TV}_T), T, K\big) \geq \mathsf{regret}_\pi(\mathcal{L}', T, K) \geq \frac{\alpha^{\frac{5}{2}} T}{300\sqrt{k}} \min\left\{\frac{\alpha^{\frac{5}{2}}}{200}, \frac{\mathsf{TV}_T}{20m}, \frac{1}{64}\sqrt{\frac{\alpha m k}{T\log(Tk/\alpha m)}}\right\}. \quad (163)$$

We now divide into two cases.

– If $\mathsf{TV}_T\sqrt{\frac{\alpha T}{12K}} \geq 1$, then let us take $m = \frac{\mathsf{TV}_T^{\frac{2}{3}} T^{\frac{1}{3}} \log^{\frac{2}{3}}(Tk/\alpha)}{(\alpha k)^{\frac{1}{3}}}$ and $k = \frac{256K}{\alpha}$, giving rise to

$$\begin{aligned}\frac{\alpha^{\frac{5}{2}} T}{300\sqrt{k}} \min\left\{\frac{\alpha^{\frac{5}{2}}}{200}, \frac{\mathsf{TV}_T}{20m}, \sqrt{\frac{\alpha m k}{2^8 T\log(Tk/\alpha m)}}\right\} &= \widetilde{\Omega}\left(\frac{T}{\sqrt{k}} \min\left\{1, \left(kT^{-1}\mathsf{TV}_T\right)^{\frac{1}{3}}\right\}\right) \\ &= \widetilde{\Omega}\left(\min\left\{\frac{T}{\sqrt{K}}, \frac{\mathsf{TV}_T^{\frac{1}{3}} T^{\frac{2}{3}}}{K^{\frac{1}{6}}}\right\}\right).\end{aligned}$$

Plugging this into (163) yields

$$\mathsf{regret}_\pi\big(\mathcal{L}_4(\mathsf{TV}_T), T, K\big) \geq \widetilde{\Omega}\left(\min\left\{\frac{T}{\sqrt{K}}, \frac{\mathsf{TV}_T^{\frac{1}{3}} T^{\frac{2}{3}}}{K^{\frac{1}{6}}}\right\}\right) \quad (164)$$

– Next, consider the case with $\mathsf{TV}_T\sqrt{\frac{\alpha T}{12K}} < 1$. Note that for any $\mathcal{D} \in \mathcal{I}$, the constant distribution sequence $\mathcal{D}_{1:T}$ with $\mathcal{D}_1 = \cdots = \mathcal{D}_T = \mathcal{D}$ belongs to $\mathcal{L}_4(\mathsf{TV}_T)$, since its cumulative variation equals 0. Then one can apply Lemma C.6 with $\epsilon = \min\left\{\frac{\alpha^{5/2}}{200}, \sqrt{\frac{\alpha K}{2^8 T\log(TK/\alpha)}}\right\}$ and $k = \frac{256}{\alpha}K$ to the entire horizon $[T]$ to arrive at

$$\begin{aligned}\mathsf{regret}_\pi\big(\mathcal{L}_4(\mathsf{TV}_T), T, K\big) &\geq \sup_{\mathcal{D}\in\mathcal{I}} \mathsf{regret}_\pi(\mathcal{D}^T, T, K) \geq \mathop{\mathbb{E}}_{\mathcal{D}\in\mathcal{I}}\left[\sum_{t=1}^{T}\Big|\mathbb{P}\big(Y_t \in \mathcal{C}(Y_{1:t-1})\big) - (1-\alpha)\Big|\right] \\ &= \widetilde{\Omega}\left(\min\left\{\frac{T}{\sqrt{K}}, \sqrt{T}\right\}\right). \quad (165)\end{aligned}$$

Combining the bounds (164) and (165) establishes the desired lower bound for the smooth drift setting. The proof of Theorem 4.2 is thus complete.

C.4.1 Proof of Lemma C.6

Throughout this proof, we shall often write $\mathcal{C}(Y_{1:n}, U)$ simply as $\mathcal{C}$, as long as it is clear from the context. For each $j = 1, \ldots, k$, we define

$$a_j = a_j(Y_{1:n}, U) \coloneqq \mu(\mathcal{C} \cap I_j), \quad (166a)$$

where we often abbreviate $a_j(Y_{1:n}, U)$ as a_j. Here, I_j is given in (154) and $\mu(\cdot)$ denotes the Lebesgue measure on the interval $[0, 1]$. Further, let

$$\overline{a} \coloneqq \frac{1}{k}\sum_{j=1}^{k} a_j \qquad \text{and} \qquad \widetilde{a}_j \coloneqq a_j - \overline{a} \ \ (j = 1, \ldots, k). \quad (166b)$$

The following lemma characterizes the range of $\sum_{j=1}^{k} \widetilde{a}_j^2$, which will be used repeatedly in our analysis. Its proof is deferred to Section C.4.2.

Lemma C.7. *For any given realization of $Y_{1:n}, U$, it always holds that*

$$\frac{1}{k}\left(\mu(\mathcal{C})\big(1-\mu(\mathcal{C})\big) - \frac{2K}{k}\right)_+ \leq \sum_{j=1}^{k} \widetilde{a}_j^2 \leq \frac{\mu(\mathcal{C})\big(1-\mu(\mathcal{C})\big)}{k} \leq \frac{1}{4k}.$$

We now embark on the proof, which contains a few steps below.

Step 1: an expression for the training-conditional coverage gap. Under the distribution $\mathcal{D}$ with parameters $V_{1:k}$ (see (155)), we can derive

$$\begin{aligned}\mathbb{P}\big(Y_{n+1} \in \mathcal{C} \,|\, Y_{1:n}, U\big) &= \sum_{j=1}^{k} \int \frac{1+\epsilon V_j}{1+\epsilon\overline{V}} \mathbb{1}\{y \in \mathcal{C} \cap I_j\} \mathrm{d}y \\ &= \sum_{j=1}^{k} \frac{1+\epsilon V_j}{1+\epsilon\overline{V}} a_j = \sum_{j=1}^{k} a_j + \frac{\epsilon}{1+\epsilon\overline{V}} \sum_{j=1}^{k} (V_j - \overline{V}) a_j \\ &\overset{\text{(a)}}{=} \mu(\mathcal{C}) + \frac{\epsilon}{1+\epsilon\overline{V}} \sum_{j=1}^{k} (V_j - \overline{V}) a_j = \mu(\mathcal{C}) + \frac{\epsilon}{1+\epsilon\overline{V}} \sum_{j=1}^{k} V_j \widetilde{a}_j,\end{aligned}$$

where (a) follows since the sets $\{\mathcal{C} \cap I_j\}_{j=1}^{k}$ are mutually disjoint and the last equality results from

$$\sum_{j=1}^{k} \overline{V} a_j = k\overline{V} \frac{1}{k} \sum_{j=1}^{k} a_j = \sum_{j=1}^{k} V_j \overline{a}.$$

Note that each distribution $\mathcal{D}$ in $\mathcal{I}$ (cf. (156)) is uniquely determined by the sequence $V_{1:k} \in \{\pm 1\}^k$. To make the subsequent analysis clearer and more concise, we define

$$l(V_{1:k}, \mathcal{C}) \coloneqq \big|\mathbb{P}\big(Y_{n+1} \in \mathcal{C} \,|\, Y_{1:n}, U\big) - (1-\alpha)\big|,$$

which, according to the above calculation, can be written as

$$l(V_{1:k}, \mathcal{C}) = \left|\mu(\mathcal{C}) - (1-\alpha) + \frac{\epsilon}{1+\epsilon\overline{V}} \sum_{j=1}^{k} V_j \widetilde{a}_j\right|. \tag{167}$$

As can be easily seen, averaging over $\mathcal{D} \in \mathcal{I}$ on the left-hand side of (159) is equivalent to taking expectation with respect to $V_{1:k}$ drawn from the uniform distribution $\mathcal{D}_V^k \coloneqq \mathsf{Unif}\big(\{\pm 1\}^k\big)$, namely,

$$\begin{aligned}&\frac{1}{|\mathcal{I}|} \sum_{\mathcal{D} \in \mathcal{I}} \mathop{\mathbb{E}}_{Y_{1:n} \sim \mathcal{D}^n,\, U \sim p_U} \Big[\big|\, \mathbb{P}\big(Y_{n+1} \in \mathcal{C}(Y_{1:n}, U) \,|\, Y_{1:n}, U\big) - (1-\alpha)\,\big|\Big] \\ &\qquad = \mathop{\mathbb{E}}_{V_{1:k} \sim \mathcal{D}_V^k} \Big[\mathop{\mathbb{E}}_{\mathcal{D}^n \times p_U} \big[\,\big|\mathbb{P}\big(Y_{n+1} \in \mathcal{C} \,|\, Y_{1:n}, U\big) - (1-\alpha)\big|\,\big]\Big]. \end{aligned} \tag{168}$$

By fully expanding the right-hand side of (168) w.r.t. the generative process of U, $V_{1:k}$ and $Y_{1:n}$, we obtain

$$\begin{aligned}&\mathop{\mathbb{E}}_{V_{1:k} \sim \mathcal{D}_V^k} \Big[\mathop{\mathbb{E}}_{\mathcal{D}^n \times p_U} \big[\big|\, \mathbb{P}\big(Y_{n+1} \in \mathcal{C} \,|\, Y_{1:n}, U\big) - (1-\alpha)\,\big|\big]\Big] \\ &\qquad = \frac{1}{2^k} \sum_{V_{1:k}} \left\{ \int_{[0,1]^n \times \mathcal{U}} l(V_{1:k}, \mathcal{C}) \Big(\prod_{i=1}^{n} f(y_i \mid V_{1:k}) \Big) p_U(u) \mathrm{d}y_{1:n} \mathrm{d}u \right\}, \end{aligned} \tag{169}$$

where $\mathcal{U}$ is the support of U, and the summation ranges over all 2^k choices of $V_{1:k}$ in $\{\pm 1\}^k$. Denote by

$$N_j(y_{1:n}) = |\{i \mid i \in [n], y_i \in I_j\}| \tag{170}$$

the number of responses that fall in the interval I_j; when there is no ambiguity, we abbreviate $N_j(y_{1:n})$ as N_j. This allows us to express the joint density as

$$\prod_{i=1}^{n} f(y_i \mid V_{1:k}) = \prod_{j=1}^{k} \left(\frac{1+\epsilon V_j}{1+\epsilon\overline{V}}\right)^{N_j} = \frac{\prod_{j=1}^{k} (1+\epsilon V_j)^{N_j}}{\big(1+\epsilon\overline{V}\big)^n},$$

which combined with (169) yields an expression for the expected training-conditional coverage gap:

$$\begin{aligned}&\mathop{\mathbb{E}}_{V_{1:k}\sim\mathcal{D}_V^k}\Big[\mathop{\mathbb{E}}_{\mathcal{D}^n\times p_U}\big[\big|\mathbb{P}(Y_{n+1}\in\mathcal{C}\mid Y_{1:n},U)-(1-\alpha)\big|\big]\Big]\\&\quad=\sum_{V_{1:k}}\Bigg\{\int_{[0,1]^n\times\mathcal{U}}\underbrace{\frac{p_U(u)\prod_{j=1}^k(1+\epsilon V_j)^{N_j(y_{1:n})}}{2^k(1+\epsilon\overline{V})^n}}_{=:\,p(V_{1:k},y_{1:n},u)}l(V_{1:k},\mathcal{C})\mathrm{d}y_{1:n}\mathrm{d}u\Bigg\}.\end{aligned}\tag{171}$$

Here, the term $p(V_{1:k},y_{1:n},u)$ represents the joint density of the random tuple $(V_{1:k},Y_{1:n},U)$.

Step 2: a lower bound on the coverage gap using auxiliary conditional distributions. Now, consider the conditional distribution of $V_{1:k}$ given $y_{1:n}$ and u, denoted by $p(V_{1:k}\mid y_{1:n},u)$. As it turns out, one can construct a set of auxiliary conditional densities $q_j(\cdot\mid y_{1:n},u)$ for $j=1,\ldots,k$, whose product provides a good approximation to $p(V_{1:k}\mid y_{1:n},u)$. This is formalized in the following lemma, whose proof is given in Section C.4.3.

Lemma C.8. *For any $y_{1:n}$ and u, let $p(y_{1:n},u)\coloneqq\sum_{v_{1:k}}p(v_{1:k},y_{1:n},u)$. There exists a collection of conditional distributions $\{q_j(V_j\mid y_{1:n},u)\}_{j=1}^k$ such that*

$$\begin{aligned}&\mathop{\mathbb{E}}_{V_{1:k}\sim\mathcal{D}_V^k}\Big[\mathop{\mathbb{E}}_{\mathcal{D}^n\times p_U}\big[\big|\,\mathbb{P}\big(Y_{n+1}\in\mathcal{C}\,\big|\,Y_{1:n},U\big)-(1-\alpha)\,\big|\big]\Big]\\&\qquad\qquad\geq\frac{1}{3}\int_{[0,1]^n\times\mathcal{U}}\mathcal{L}_Q(y_{1:n},u)p(y_{1:n},u)\mathrm{d}y_{1:n}\mathrm{d}u-\frac{\alpha^6}{4n^3k^3}\end{aligned}\tag{172}$$

as long as $\epsilon\leq\sqrt{\frac{k}{12n\log(2nk/\alpha^2)}}$, where

$$\mathcal{L}_Q(y_{1:n},u)\coloneqq\sum_{V_{1:k}}\Big(\prod_{j=1}^k q_j(V_j\mid y_{1:n},u)\Big)\,l(V_{1:k},\mathcal{C}).\tag{173}$$

As can be seen, each summand in $\mathcal{L}_Q(y_{1:n},u)$ (cf. (173)) involves the product distribution based on $\{q_j(\cdot\mid y_{1:n},u)\}$.

Step 3: a decomposition of $\mathcal{L}_Q(y_{1:n},u)$. With Lemma C.8, we now turn to the analysis of the quantity $\mathcal{L}_Q(y_{1:n},u)$ (cf. (173)). Given $(y_{1:n},u)$, define $\mathcal{D}_Q$ as the product distribution

$$\mathcal{D}_Q(V_{1:k})\coloneqq\prod_{j=1}^k q_j(V_j\mid y_{1:n},u),\tag{174}$$

and draw an independent copy $V'_{1:k}$ of $V_{1:k}$ from the same distribution $\mathcal{D}_Q$. Then, invoking (167) and the triangle inequality, we can derive

$$\begin{aligned}\mathcal{L}_Q(y_{1:n},u)&=\sum_{V_{1:k}}\Big(\prod_{j=1}^k q_j(V_j\mid y_{1:n},u)\Big)\,l(V_{1:k},\mathcal{C})=\mathbb{E}_{\mathcal{D}_Q}\big[l(V_{1:k},\mathcal{C})\big]\\&=\frac{1}{2}\mathbb{E}_{\mathcal{D}_Q}\big[l(V_{1:k},\mathcal{C})\big]+\frac{1}{2}\mathbb{E}_{\mathcal{D}_Q}\big[l(V'_{1:k},\mathcal{C})\big]\\&\overset{(167)}{=}\frac{1}{2}\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\big(\mu(\mathcal{C})-(1-\alpha)\big)+\frac{\epsilon}{1+\epsilon\overline{V}}\langle V_{1:k},\widetilde{a}_{1:k}\rangle\right|+\left|\big(\mu(\mathcal{C})-(1-\alpha)\big)+\frac{\epsilon}{1+\epsilon\overline{V}'}\langle V'_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]\\&\geq\frac{1}{2}\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\left(\frac{\epsilon}{1+\epsilon\overline{V}}\langle V_{1:k},\widetilde{a}_{1:k}\rangle-\frac{\epsilon}{1+\epsilon\overline{V}}\langle V'_{1:k},\widetilde{a}_{1:k}\rangle\right)+\left(\frac{\epsilon}{1+\epsilon\overline{V}}-\frac{\epsilon}{1+\epsilon\overline{V}'}\right)\langle V'_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]\\&\geq\frac{1}{2}\mathbb{E}_{\mathcal{D}_Q^2}\left[\frac{\epsilon}{1+\epsilon\overline{V}}\left|\langle V_{1:k}-V'_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]-\frac{1}{2}\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\left(\frac{\epsilon}{1+\epsilon\overline{V}}-\frac{\epsilon}{1+\epsilon\overline{V}'}\right)\langle V'_{1:k},\widetilde{a}_{1:k}\rangle\right|\right],\end{aligned}$$

from which it follows that

$$\mathbb{E}_{Y_{1:n},U}\big[\mathcal{L}_Q(Y_{1:n},U)\big] \geq \underbrace{\frac{1}{2}\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\frac{\epsilon}{1+\epsilon\overline{V}}\left|\langle V_{1:k}-V'_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]\right]}_{=:\mathcal{L}_1} - \underbrace{\frac{1}{2}\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\left(\frac{\epsilon}{1+\epsilon\overline{V}}-\frac{\epsilon}{1+\epsilon\overline{V}'}\right)\langle V'_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]\right]}_{=:\mathcal{L}_2}. \tag{175}$$

This leaves us two terms to control.

Step 4: lower bounds on $\mathcal{L}_1$ and $\mathcal{L}_2$. Next, we would like to control the two terms on the right-hand side of (175) separately. Before continuing, we introduce some additional notation: for any $j = 1, 2, \ldots, k$, define

$$\widetilde{V}_j \coloneqq \frac{1}{2}(V_j - V'_j), \qquad \check{V} \coloneqq \frac{1}{2}(\overline{V} - \overline{V}'), \qquad \widetilde{V}_{1:k} \coloneqq \frac{1}{2}(V_{1:k} - V'_{1:k}). \tag{176}$$

It is straightforward to see that since $V_j, V'_j \in \{\pm 1\}$, we have $\widetilde{V}_j \in \{-1, 0, 1\}$ for all $j = 1, 2, \ldots, k$.

- With regards to $\mathcal{L}_1$, recognizing that $\left|\epsilon\overline{V}\right| \leq |\epsilon| \leq 1/4$, one has

$$\mathcal{L}_1 \geq \frac{4\epsilon}{3}\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\langle \widetilde{V}_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]\right], \tag{177}$$

the right-hand side of which is further controlled by the following lemma. The proof can be found in Section C.4.4.

Lemma C.9. *Let $k \geq 64/\alpha$ and $\epsilon \leq \min\left\{\frac{\alpha^{5/2}}{200}, \frac{1}{64}\sqrt{\frac{\alpha k}{n\log(2nk/\alpha^2)}}\right\}$. For any admissible algorithm whose resulting prediction set $\mathcal{C}$ is the union of at most k intervals, we have*

$$\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\langle \widetilde{V}_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]\right] \geq \left(\frac{\sigma_\pi^2}{2} - \frac{\alpha}{16}\right)^{\frac{5}{2}}\frac{2}{\sqrt{k}},$$

where we define

$$\sigma_\pi^2 \coloneqq \mathbb{E}_{Y_{1:n},U}\left[\left(\mu(\mathcal{C})\big(1-\mu(\mathcal{C})\big) - \frac{2K}{k}\right)_+\right]. \tag{178}$$

Consequently, taking (177) and Lemma C.9 collectively reveals that

$$\mathcal{L}_1 \geq \frac{8\epsilon}{3\sqrt{k}}\left(\frac{\sigma_\pi^2}{2} - \frac{\alpha}{16}\right)^{\frac{5}{2}}. \tag{179}$$

- When it comes to $\mathcal{L}_2$, we make the following observation:

$$\begin{aligned}\mathcal{L}_2 &= \mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\left(\frac{\epsilon}{1+\epsilon\overline{V}}-\frac{\epsilon}{1+\epsilon\overline{V}'}\right)\langle V'_{1:k},\widetilde{a}_{1:k}\rangle\right|\right]\right] \\ &\overset{\text{(a)}}{\leq} \mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\frac{\epsilon}{1+\epsilon\overline{V}}-\frac{\epsilon}{1+\epsilon\overline{V}'}\right|\right]\right] \overset{\text{(b)}}{\leq} 4\epsilon^2\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\overline{V}'-\overline{V}\right|\right]\right] \\ &= 8\epsilon^2\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{\mathcal{D}_Q^2}\left[\left|\check{V}\right|\right]\right] \overset{\text{(c)}}{\leq} 8\epsilon^2\mathbb{E}_{Y_{1:n},U}\left[\left(\mathbb{E}_{\mathcal{D}_Q^2}\left[\frac{1}{k^2}\sum_{j=1}^k \widetilde{V}_j^2\right]\right)^{1/2}\right] \overset{\text{(d)}}{\leq} \frac{8\epsilon^2}{\sqrt{k}}.\end{aligned} \tag{180}$$

Here, (a) is valid since $|\langle V'_{1:k},\widetilde{a}_{1:k}\rangle| \leq \sum_{j=1}^k |\widetilde{a}_j| \leq \mu([0,1]) = 1$ (see (166)); (b) follows from the fact that $\min\{1+\epsilon\overline{V}, 1+\epsilon\overline{V}'\} \geq 1-\epsilon \geq 1/2$; (c) applies Jensen's inequality and uses the properties that $\{\widetilde{V}_j\}$ are independent zero-mean random variables; and (d) arises from the inequality $\widetilde{V}_j^2 \leq 1$.

Combine Eqns. (179) and (180) with (175) to yield

$$\begin{aligned}\mathbb{E}_{Y_{1:n},U}\left[\mathcal{L}_Q(Y_{1:n},U)\right] &\geq \frac{1}{2}\mathcal{L}_1 - \frac{1}{2}\mathcal{L}_2 \\ &\geq \frac{4\epsilon}{3\sqrt{k}}\left(\frac{\sigma_\pi^2}{2}-\frac{\alpha}{16}\right)^{\frac{5}{2}} - \frac{4\epsilon^2}{\sqrt{k}} = \frac{4\epsilon}{3\sqrt{k}}\left(\left(\frac{\sigma_\pi^2}{2}-\frac{\alpha}{16}\right)^{\frac{5}{2}} - 3\epsilon\right).\end{aligned} \tag{181}$$

Step 5: putting all pieces together. To finish up, we divide into two cases and analyze them separately.

- If $\mathbb{P}\left(|\mu(\mathcal{C})-(1-\alpha)|>\alpha/8\right)\geq 1/4$, then one has

$$\begin{aligned}&\mathbb{E}_{V_{1:n}}\left[\mathop{\mathbb{E}}_{\mathcal{D}^n\times p_U}\left[\left|\mathbb{P}\big(Y_{n+1}\in\mathcal{C}\mid Y_{1:n},U\big)-(1-\alpha)\right|\right]\right] \geq \mathbb{E}\left[\left|\mu(C)-(1-\alpha)\right| - \left|\frac{\epsilon}{1+\epsilon\overline{\widetilde{V}}}\left\langle V_{1:k},\widetilde{a}_{1:k}\right\rangle\right|\right] \\ &\geq \frac{\alpha}{32} - 2\epsilon\mathbb{E}\left[\left|\sum_{j=1}^k V_j\widetilde{a}_j\right|\right] \geq \frac{\alpha}{32} - 2\epsilon\mathbb{E}\left[\sqrt{\Big(\sum_{j=1}^k V_j^2\Big)\Big(\sum_{j=1}^k \widetilde{a}_j^2\Big)}\right] \\ &\overset{\text{(e)}}{\geq} \frac{\alpha}{32}-\epsilon \overset{\text{(f)}}{\geq} \frac{\alpha}{32}-\frac{\alpha}{64} = \frac{\alpha}{64},\end{aligned}$$

  where (e) invokes Lemma C.7, and (f) is due to our choice of ϵ.

- If instead $\mathbb{P}\left(|\mu(\mathcal{C})-(1-\alpha)|>\alpha/8\right)<1/4$, then it holds that

$$\mathbb{P}\left(1-\frac{9\alpha}{8}\leq\mu(\mathcal{C})\leq 1-\frac{7\alpha}{8}\right)\geq\frac{3}{4}. \tag{182}$$

  We can then lower bound σ_π^2 (cf. (178)) by

$$\begin{aligned}\sigma_\pi^2 &= \mathbb{E}\left[\left(\mu(\mathcal{C})\big(1-\mu(\mathcal{C})\big)-\frac{2K}{k}\right)_+\right] \\ &\geq \mathbb{E}\left[\left(\mu(\mathcal{C})\big(1-\mu(\mathcal{C})\big)-\frac{2K}{k}\right)_+ \mathbb{1}\left\{1-\frac{9\alpha}{8}\leq\mu(\mathcal{C})\leq 1-\frac{7\alpha}{8}\right\}\right] \\ &\geq \mathbb{E}\left[\left(\left(1-\frac{9\alpha}{8}\right)\frac{7\alpha}{8}-\frac{2K}{k}\right)_+ \mathbb{1}\left\{1-\frac{9\alpha}{8}\leq\mu(\mathcal{C})\leq 1-\frac{7\alpha}{8}\right\}\right] \\ &\overset{\text{(g)}}{\geq} \left(\frac{49\alpha}{128}-\frac{2K}{k}\right)_+ \mathbb{P}\left(1-\frac{9\alpha}{8}\leq\mu(\mathcal{C})\leq 1-\frac{7\alpha}{8}\right) \geq \frac{3\alpha}{8}\times\frac{3}{4}\geq\frac{9\alpha}{32}.\end{aligned}$$

  Here, (g) holds when $\alpha\leq 1/2$, and the last inequality follows by combining $k\geq\frac{256K}{\alpha}$ and (182). Taking this together with (181) and $\epsilon\leq\frac{\alpha^{5/2}}{200}$, we arrive at

$$\mathbb{E}_{Y_{1:n},U}\left[\mathcal{L}_Q(Y_{1:n},U)\right]\geq\frac{4\epsilon}{3\sqrt{k}}\left(\left(\frac{9\alpha}{32}-\frac{\alpha}{32}\right)^{\frac{5}{2}}-\frac{\alpha^{\frac{5}{2}}}{64}\right)\geq\frac{\alpha^{\frac{5}{2}}\epsilon}{48\sqrt{k}}. \tag{183}$$

  Substituting Eqn. (183) into (172), we obtain

$$\mathop{\mathbb{E}}_{V_{1:n}}\left[\mathop{\mathbb{E}}_{\mathcal{D}^n\times p_U}\left[\left|\mathbb{P}(Y_{n+1}\in\mathcal{C}\mid Y_{1:n},U)-(1-\alpha)\right|\right]\right]\geq\frac{1}{3}\mathop{\mathbb{E}}_{Y_{1:n},U}[\mathcal{L}_Q(Y_{1:n},U)]-\frac{\alpha^6}{4n^3k^3}\geq\frac{\alpha^{\frac{5}{2}}\epsilon}{144\sqrt{k}}-\frac{\alpha^6}{4n^3k^3}.$$

The above two cases taken collectively conclude the proof of Lemma C.6.

C.4.2 Proof of Lemma C.7

Upper bound. According to the definition of a_j and $\overline{a}$ (see (166)), it is readily seen that

$$\sum_{j=1}^{k} \widetilde{a}_j^2 = \sum_{j=1}^{k} \big(a_j - \overline{a}\big)^2 = \sum_{j=1}^{k} a_j^2 - k\overline{a}^2 \le \frac{1}{k}\sum_{j=1}^{k} a_j - \frac{\mu(\mathcal{C})^2}{k} = \frac{\mu(\mathcal{C})\big(1-\mu(\mathcal{C})\big)}{k} \le \frac{1}{4k},$$

where the first inequality follows since $a_j \le \mu(I_j) = 1/k$ and $\overline{a} = \mu(\mathcal{C})/k$, and the last inequality comes from the AM-GM inequality.

Lower bound. To establish the lower bound, we exploit the structural property of the prediction set $\mathcal{C}$ (i.e., $\mathcal{C}$ is the union of at most K intervals). Consider the following conditions:

- suppose l intervals from $\{I_j\}_{j=1}^k$ (without loss of generality, $\{I_j\}_{j=1}^l$) are completely contained in $\mathcal{C}$;
- at most $2K$ of the $\{I_j\}_{j=1}^k$ (without loss of generality, $\{I_{j_i}\}_{i=1}^{2K}$) are only partially covered by $\mathcal{C}$, which always holds due to the structural property of $\mathcal{C}$.

Since the intervals $\{I_j\}_{j=1}^k$ are mutually disjoint, the total Lebesgue measure of these at most $l+2K$ intervals must be at least $\mu(\mathcal{C})$. This leads to

$$\frac{l+2K}{k} = \sum_{j=1}^{l} \big|\, I_j \,\big| + \sum_{i=1}^{2K} \big|\, I_{j_i} \,\big| \ge \mu(\mathcal{C}),$$

thus indicating that

$$l \ge \big(\mu(\mathcal{C})k - 2K\big)_+. \tag{184}$$

Now, let us look at $\{\widetilde{a}_j\}_{j=1}^k$. For each $1 \le j \le l$, we have $a_j = \mu(\mathcal{C} \cap I_j) = 1/k$ and $k\overline{a} = \sum_{j=1}^k a_j = \mu(\mathcal{C})$, allowing one to derive

$$\begin{aligned}
\sum_{j=1}^{k} \widetilde{a}_j^2 = \sum_{j=1}^{k} \big(a_j - \overline{a}\big)^2 = \sum_{j=1}^{k} a_j^2 - k\overline{a}^2 = \sum_{j=1}^{k} a_j^2 - \frac{\mu(\mathcal{C})^2}{k} \\
\ge \sum_{j=1}^{l} a_j^2 - \frac{\mu(\mathcal{C})^2}{k} = \frac{l}{k^2} - \frac{\mu(\mathcal{C})^2}{k} \\
\overset{(184)}{\ge} \frac{1}{k}\left\{\Big(\mu(\mathcal{C}) - \frac{2K}{k}\Big)_+ - \mu(\mathcal{C})^2\right\} \ge \frac{1}{k}\left\{\mu(\mathcal{C})\big(1-\mu(\mathcal{C})\big) - \frac{2K}{k}\right\}.
\end{aligned}$$

This combined with the trivial fact $\sum_{j=1}^k \widetilde{a}_j^2 \ge 0$ establishes the advertised lower bound on $\sum_{j=1}^k \widetilde{a}_j^2$.

C.4.3 Proof of Lemma C.8

By virtue of Bayes's rule, the conditional density $p(V_{1:k} \mid y_{1:n}, u)$ admits the following expression:

$$\begin{aligned}
p(V_{1:k} \mid y_{1:n}, u) = \frac{p(V_{1:k}, y_{1:n}, u)}{\sum_{v_{1:k}} p(v_{1:k}, y_{1:n}, u)} = \frac{\Big(p_U(u)\prod_{j=1}^k (1+\epsilon V_j)^{N_j}\Big)/2^k(1+\epsilon\overline{V})^n}{\sum_{v_{1:k}}\Big(p_U(u)\prod_{j=1}^k (1+\epsilon v_j)^{N_j}\Big)/2^k(1+\epsilon\overline{v})^n} \\
= \frac{\Big(\prod_{j=1}^k (1+\epsilon V_j)^{N_j}\Big)/(1+\epsilon\overline{V})^n}{\sum_{v_{1:k}}\Big[\Big(\prod_{j=1}^k (1+\epsilon v_j)^{N_j}\Big)/(1+\epsilon\overline{v})^n\Big]}.
\end{aligned} \tag{185}$$

For any given $V_{1:k}$, we observe that

$$\begin{aligned}
\Big(\prod_{j=1}^k (1+\epsilon V_j)^{N_j}\Big) \Big/ \big(1+\epsilon\overline{V}\big)^n = \Big(\prod_{j=1}^k (1+\epsilon V_j)^{N_j}\Big) e^{-\epsilon\overline{V}n} \cdot \frac{e^{\epsilon\overline{V}n}}{\big(1+\epsilon\overline{V}\big)^n} \\
\ge \prod_{j=1}^k (1+\epsilon V_j)^{N_j} e^{-\epsilon V_j \frac{n}{k}} =: \prod_{j=1}^k q_j(V_j, y_{1:n}),
\end{aligned} \tag{186}$$

where the last line follows from the elementary inequality $(e^x)^n \geq (1+x)^n$ for all $x \in [-1,1]$, and we define

$$q_j(V_j, y_{1:n}) \coloneqq (1+\epsilon V_j)^{N_j} e^{-\epsilon V_j \frac{n}{k}}.$$

Further, define

$$q_j(y_{1:n}) \coloneqq \sum_{V_j} (1+\epsilon V_j)^{N_j} e^{-\epsilon V_j \frac{n}{k}}, \qquad j = 1, \ldots, k; \tag{187a}$$

$$q(y_{1:n}) \coloneqq \sum_{V_{1:k}} \prod_{j=1}^{k} \left[(1+\epsilon V_j)^{N_j} e^{-\epsilon V_j \frac{n}{k}} \right] = \prod_{j=1}^{k} q_j(y_{1:n}); \tag{187b}$$

$$q_j(V_j \mid y_{1:n}) \coloneqq \frac{q_j(V_j, y_{1:n})}{q_j(y_{1:n})}, \qquad j = 1, \ldots, k \tag{187c}$$

$$q(V_{1:k} \mid y_{1:n}) \coloneqq \prod_{j=1}^{k} q_j(V_j \mid y_{1:n}). \tag{187d}$$

Combining the above analysis and notation with a little algebra, we reach

$$\begin{aligned}
&\mathbb{E}_{V_{1:k}}\left[\mathbb{E}_{\mathcal{D}^n \times p_U}\left[\left|\, \mathbb{P}(Y_{n+1} \in \mathcal{C} \,|\, Y_{1:n}, U) - (1-\alpha) \,\right|\right]\right] \\
&\overset{(171)}{=} \int_{[0,1]^n \times \mathcal{U}} \left\{ \sum_{V_{1:k}} p(V_{1:k} \mid y_{1:n}, u) l(V_{1:k}, \mathcal{C}) \right\} p(y_{1:n}, u) \mathrm{d}y_{1:n} \mathrm{d}u \\
&\overset{(185)}{=} \int_{[0,1]^n \times \mathcal{U}} \left\{ \sum_{V_{1:k}} \frac{\left(\prod_{j=1}^{k} (1+\epsilon V_j)^{N_j}\right)\left(1+\epsilon \overline{V}\right)^{-n}}{\sum_{v_{1:k}} \left(\prod_{j=1}^{k} (1+\epsilon v_j)^{N_j}\right)(1+\overline{v})^{-n}} l(V_{1:k}, \mathcal{C}) \right\} p(y_{1:n}, u) \mathrm{d}y_{1:n} \mathrm{d}u \\
&\overset{(186)}{\geq} \int_{[0,1]^n \times \mathcal{U}} \left\{ \sum_{V_{1:k}} \frac{\prod_{j=1}^{k} q_j(V_j, y_{1:n})}{\sum_{v_{1:k}} \left(\prod_{j=1}^{k} (1+\epsilon v_j)^{N_j}\right)(1+\epsilon \overline{v})^{-n}} l(V_{1:k}, \mathcal{C}) \right\} p(y_{1:n}, u) \mathrm{d}y_{1:n} \mathrm{d}u \\
&\overset{(187)}{=} \int_{[0,1]^n \times \mathcal{U}} \underbrace{\left(\frac{q(y_{1:n})}{\sum_{v_{1:k}} \left(\prod_{j=1}^{k} (1+\epsilon v_j)^{N_j}\right)(1+\epsilon \overline{v})^{-n}} \right)}_{=: \mathcal{G}(y_{1:n})} \left\{ \sum_{V_{1:k}} \prod_{j=1}^{k} q_j(V_j \mid y_{1:n}) l(V_{1:k}, \mathcal{C}) \right\} p(y_{1:n}, u) \mathrm{d}y_{1:n} \mathrm{d}u.
\end{aligned} \tag{188}$$

The next step is to control $\mathcal{G}(y_{1:n})$. Specifically, by expanding $q(y_{1:n})$ and applying Eqn. (185), we can establish the following inequality:

$$\begin{aligned}
\mathcal{G}(y_{1:n}) &\overset{(187b)}{=} \frac{\sum_{V_{1:k}} \left(\prod_{j=1}^{k} (1+\epsilon V_j)^{N_j}\right)\left(1+\epsilon \overline{V}\right)^{-n} \frac{(1+\epsilon \overline{V})^n}{e^{\epsilon \overline{V} n}}}{\sum_{v_{1:k}} \left(\prod_{j=1}^{k} (1+\epsilon v_j)^{N_j}\right)\left(1+\epsilon \overline{v}\right)^{-n}} \overset{(185)}{=} \sum_{V_{1:k}} p(V_{1:k} \mid y_{1:n}) \frac{\left(1+\epsilon \overline{V}\right)^n}{e^{\epsilon \overline{V} n}} \\
&\geq \sum_{V_{1:k}} p(V_{1:k} \mid y_{1:n}) \frac{\left(1+\epsilon \overline{V}\right)^n}{e^{\epsilon \overline{V} n}} \mathbb{1}\left\{ |\overline{V}| \leq \sqrt{6k^{-1} \log(2nk/\alpha^2)} \right\} \\
&\overset{(a)}{\geq} \frac{1}{3} \sum_{V_{1:k}} p(V_{1:k} \mid y_{1:n}) \mathbb{1}\left\{ |\overline{V}| \leq \sqrt{6k^{-1} \log(2nk/\alpha^2)} \right\} \\
&= \frac{1}{3}\left(1 - \mathbb{P}\left(\overline{V} > \sqrt{\frac{6 \log(2nk/\alpha^2)}{k}} \;\middle|\; Y_{1:n} = y_{1:n} \right)\right).
\end{aligned} \tag{189}$$

To justify why (a) holds, note that

$$\epsilon |\overline{V}| \leq \sqrt{\frac{k}{12n \log(2nk/\alpha^2)}} \sqrt{\frac{6 \log(2nk/\alpha^2)}{k}} = \sqrt{\frac{1}{2n}}$$

holds under the condition that $\left|\overline{V}\right| \le \sqrt{2k^{-1}\log(nk/\alpha)}$, which combined with the elementary fact $1+x+x^2 \ge e^x$ $(x \in [-1,1])$ gives

$$\frac{e^{\epsilon\overline{V}n}}{\left(1+\epsilon\overline{V}\right)^n} \le \left(\frac{1+\epsilon\overline{V}+(\epsilon\overline{V})^2}{1+\epsilon\overline{V}}\right)^n \le \left(1+2(\epsilon\overline{V})^2\right)^n \le \left(1+\frac{1}{n}\right)^n \le e < 3.$$

As a consequence, substituting Eqn. (189) into Eqn. (188) yields

$$\begin{aligned}
&\mathbb{E}_{V_{1:k}}\left[\mathbb{E}_{\mathcal{D}^n\times p_U}\left[\left|\mathbb{P}\big(Y_{n+1}\in\mathcal{C}\,|\,Y_{1:n},U\big)-(1-\alpha)\right|\right]\right] \\
&\quad\ge \frac{1}{3}\int_{[0,1]^n\times\mathcal{U}}\left(1-\mathbb{P}\left(\overline{V}>\sqrt{\frac{6\log(2nk/\alpha^2)}{k}}\;\middle|\; y_{1:n}\right)\right)\left\{\sum_{V_{1:k}}\prod_{j=1}^k q_j(V_j\mid y_{1:n})l(V_{1:k},\mathcal{C})\right\}p(y_{1:n},u)\mathrm{d}y_{1:n}\mathrm{d}u \\
&\quad\overset{\text{(b)}}{\ge} \frac{1}{3}\int_{[0,1]^n\times\mathcal{U}}\left\{\sum_{V_{1:k}}\prod_{j=1}^k q_j(V_j\mid y_{1:n})l(V_{1:k},\mathcal{C})\right\}p(y_{1:n},u)\mathrm{d}y_{1:n}\mathrm{d}u \\
&\qquad\qquad -\frac{1}{3}\int_{[0,1]^n\times\mathcal{U}}\mathbb{P}\left(\left|\overline{V}\right|>\sqrt{\frac{6\log(2nk/\alpha^2)}{k}}\;\middle|\; y_{1:n}\right)p(y_{1:n},u)\mathrm{d}y_{1:n}\mathrm{d}u \\
&\quad= \frac{1}{3}\int_{[0,1]^n\times\mathcal{U}}\mathcal{L}_Q(y_{1:n})p(y_{1:n},u)\mathrm{d}y_{1:n}\mathrm{d}u-\frac{1}{3}\mathbb{P}\left(\left|\overline{V}\right|>\sqrt{\frac{6\log(2nk/\alpha^2)}{k}}\right),
\end{aligned} \tag{190}$$

where (b) relies on the fact that

$$\sum_{V_{1:k}}\prod_{j=1}^k q_j(V_j\mid y_{1:n})l(V_{1:k},\mathcal{C}) \overset{(167)}{=} \mathop{\mathbb{E}}_{V_{1:k}\sim q(V_{1:k}\mid y_{1:n})}\left[\left|\,\mathbb{P}\big(Y_{n+1}\in\mathcal{C}\,|\,Y_{1:n},U\big)-(1-\alpha)\,\right|\right]\le 1.$$

Finally, recalling that $V_{1:k}$ are k independent Rademacher random variables, we can apply Hoeffding's inequality to yield

$$\mathbb{P}\left(\left|\overline{V}\right|>\sqrt{\frac{6\log(2nk/\alpha^2)}{k}}\right)\le 2\exp\left\{-\frac{6k\log\frac{2nk}{\alpha^2}}{2k}\right\}\le\frac{\alpha^6}{4n^3k^3}.$$

Substitution into the above bound (190) concludes the proof of Lemma C.8.

C.4.4 Proof of Lemma C.9

As discussed earlier, we have $\widetilde{V}_j\in\{-1,0,1\}$ (see (176)). Further, define

$$\delta_j \coloneqq \mathbb{1}\{\widetilde{V}_j\neq 0\},\qquad\text{and}\qquad \xi_j\coloneqq\begin{cases}\operatorname{sgn}\big(\widetilde{V}_j\big), & \text{if } \delta_j=1,\\ \zeta_j, & \text{if } \delta_j=0,\end{cases}$$

where $\{\zeta_j\}_{j=1}^k$ are k independent Rademacher random variables, generated independent of $(\delta_j)_{j=1}^k$, $Y_{1:n}$ and U. It is then straightforward to verify that $\widetilde{V}_j=\xi_j\delta_j$. Additionally, it is easy to see that for any given $y_{1:n}$ and u, one has

$$\mathbb{P}(\xi_j=1,\delta_j=1\mid y_{1:n},u)=\mathbb{P}(V_j>V_j'\mid y_{1:n},u)=\mathbb{P}(V_j<V_j'\mid y_{1:n},u)=\mathbb{P}(\xi_j=-1,\delta_j=1\mid y_{1:n},u),$$

where the second identity follows from the fact that V_j and V_j' are independently and identically distributed. This implies that, conditional on $\delta_j=1$ and $(y_{1:n},u)$, the variable ξ_j is a Rademacher random variable. Similarly, when $\delta_j=0$ and when $(y_{1:n},u)$ are given, we have $\xi_j=\zeta_j$, which is also Rademacher.

Moreover, since $\{V_j\}_{j=1}^k$ and $\{V_j'\}_{j=1}^k$ are independent within each sequence (with $V_{1:k} \sim \prod_{j=1}^k q_j(V_j \mid y_{1:n}, u)$), it follows that $\{\xi_j\}_{j=1}^k$ are independent Rademacher random variables, which are also independent of $\big(\{\delta_j\}_{j=1}^k, Y_{1:n}, U\big)$. Therefore, for any fixed $\big(\{\delta_j\}_{j=1}^k, y_{1:n}, u\big)$, applying Lemma E.2 to $\{\xi_j\}_{j=1}^k$ yields

$$\begin{aligned}\mathbb{E}\Big[\big|\langle \widetilde{V}_{1:k}, \widetilde{a}_{1:k}\rangle\big|\Big] &= \mathbb{E}\left[\left|\sum_{j=1}^k \xi_j\delta_j\widetilde{a}_j\right|\right] = \mathbb{E}\left[\mathbb{E}\left[\left|\sum_{j=1}^k \xi_j\delta_j\widetilde{a}_j\right| \;\middle|\; \{\delta_j\}_{j=1}^k,\ \{\widetilde{a}_j\}_{j=1}^k\right]\right] \\ &\geq \frac{1}{\sqrt{2}}\mathbb{E}\left[\left(\sum_{j=1}^k \widetilde{a}_j^2\delta_j^2\right)^{\frac{1}{2}}\right] = \frac{1}{\sqrt{2}}\mathbb{E}\left[\left(\sum_{j=1}^k \widetilde{a}_j^2\delta_j\right)^{\frac{1}{2}}\right],\end{aligned} \tag{191}$$

where the last equality holds since $\delta_j^2 = \delta_j$.

To continue, let us first examine $\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2\delta_j\right]$. Note that, given $(y_{1:n}, u)$, under the distribution $\prod_{j=1}^k q_j(V_j \mid y_{1:n}, u)$ one has

$$\begin{aligned}\mathbb{E}[\delta_j \mid y_{1:n}, u] &= \mathbb{P}(V_j \neq V_j' \mid y_{1:n}, u) = 2q_j(1 \mid y_{1:n}, u)q_j(0 \mid y_{1:n}, u) \\ &= \frac{1}{2}\left(1 - (2q_j(1 \mid y_{1:n}, u) - 1)^2\right) =: \frac{\big(1 - (2q_j - 1)^2\big)}{2},\end{aligned}$$

where we define

$$q_j \coloneqq q_j(1 \mid y_{1:n}, u) = \frac{(1+\epsilon)^{N_j} e^{-\epsilon\frac{n}{k}}}{\sum_{V=\pm 1}(1+\epsilon V)^{N_j} e^{-\epsilon\frac{n}{k}V}}. \tag{192}$$

From this, we can derive that

$$\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2\delta_j\right] = \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2\right] - \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2(2q_j - 1)^2\right]. \tag{193}$$

As for the first term on the right-hand side of (193), applying Lemma C.7 and then taking expectation over $Y_{1:n}$ and U yield

$$\frac{1}{2}\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2\right] \geq \frac{1}{2k}\mathbb{E}\left[\Big(\mu(\mathcal{C})\big(1 - \mu(\mathcal{C})\big) - \frac{2K}{k}\Big)_+\right] \tag{194}$$

Also, the second term on the right-hand side of (193) satisfies

$$\frac{1}{2}\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2(2q_j - 1)^2\right] \leq \frac{1}{2k^2}\mathbb{E}\left[\sum_{j=1}^k (2q_j - 1)^2\right]. \tag{195}$$

To control $\mathbb{E}\big[\sum_{j=1}^k (2q_j - 1)^2\big]$, we resort to the following lemma, with the proof given in Section C.4.5.

Lemma C.10. *Recall that q_j is defined in (192). For any admissible algorithm whose resulting prediction set $\mathcal{C}$ is the union of at most k intervals, we have*

$$\mathbb{E}\left[\sum_{j=1}^k (2q_j - 1)^2\right] \leq 321n\epsilon^2 \log\Big(\frac{2nk}{\alpha^2}\Big) + \frac{\alpha^6}{32n^3k^3}.$$

Taking Eqns. (193), (194), (195) and Lemma C.10 together yields

$$\begin{aligned}\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2\delta_j\right] &\geq \frac{1}{2k}\mathbb{E}\left[\Big(\mu(\mathcal{C})\big(1 - \mu(\mathcal{C})\big) - \frac{2K}{k}\Big)_+\right] - \frac{1}{k^2}\left(161n\epsilon^2 \log\Big(\frac{2nk}{\alpha^2}\Big) + \frac{\alpha^6}{64n^3k^3}\right) \\ &\geq \frac{\sigma_\pi^2}{2k} - \frac{\alpha}{16k},\end{aligned} \tag{196}$$

where the last line holds as long as $k \geq \frac{256K}{\alpha}$ and $\epsilon \leq \min\left\{\frac{\alpha^{5/2}}{200}, \frac{1}{64}\sqrt{\frac{\alpha k}{n\log(2nk/\alpha^2)}}\right\}$.

To finish up, denoting $\mathbb{E}\left[\sum_{j=1}^k \widetilde{a}_j^2\delta_j\right]$ as λ, and applying Lemma E.3 with $\theta = 1/2$, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\sum_{j=1}^k \widetilde{a}_j^2\delta_j}\right] &\geq \frac{1}{\sqrt{2}}\sqrt{\lambda}\mathbb{P}\left(\sum_{j=1}^k \widetilde{a}_j^2\delta \geq \frac{\lambda}{2}\right) \geq \frac{1}{4\sqrt{2}}\sqrt{\lambda}\frac{\lambda^2}{\mathbb{E}\left[\left(\sum_{j=1}^k \widetilde{a}_j^2\delta_j\right)^2\right]} \\
&\geq \frac{1}{4\sqrt{2}}\frac{\lambda^{5/2}}{\mathbb{E}\left[\left(\sum_{j=1}^k \widetilde{a}_j^2\right)^2\right]} \overset{\text{Lemma C.7}}{\geq} \frac{k^2}{4\sqrt{2}}\frac{\lambda^{5/2}}{\mathbb{E}\left[\mu(\mathcal{C})^2\big(1-\mu(\mathcal{C})\big)^2\right]} \\
&\geq 2\sqrt{2}k^2\lambda^{5/2} \overset{(196)}{\geq} \frac{2\sqrt{2}}{\sqrt{k}}\left(\frac{\sigma_\pi^2}{2} - \frac{\alpha}{16}\right)^{\frac{5}{2}}.
\end{aligned}$$

Combining this with Eqn. (191) establishes the advertised result of the lemma.

C.4.5 Proof of Lemma C.10

To begin with, the TV distance between two Bernoulli distributions $\mathrm{Ber}(q_j)$ and $\mathrm{Ber}(1/2)$ obeys

$$\mathsf{TV}\big(\mathrm{Ber}(q_j),\ \mathrm{Ber}(1/2)\big) = \left|q_j - \frac{1}{2}\right| + \left|(1-q_j) - \frac{1}{2}\right| = |2q_j - 1|,$$

which combined with Pinsker's inequality (Tsybakov, 2009, Lemm 2.5) yields

$$(2q_j-1)^2 = \big(\mathsf{TV}\big(\mathrm{Ber}(q_j),\ \mathrm{Ber}(1/2)\big)\big)^2 \leq \frac{1}{2}\mathsf{KL}\left(\mathrm{Ber}(q_j)\,\middle\|\,\mathrm{Ber}(0.5)\right). \tag{197}$$

Note that the distribution $q(V_{1:k} \mid y_{1:n}) = \prod_{j=1}^k q_j(V_j \mid y_{1:n})$ factorizes. Therefore, recalling the definition of q_j in (192), one can apply the chain rule of the KL divergence to derive

$$\begin{aligned}
2\mathbb{E}\left[\sum_{j=1}^k (2q_j-1)^2\right] &\leq \mathbb{E}_{Y_{1:n}}\left[\sum_{j=1}^k \mathsf{KL}\left(\mathrm{Ber}(q_j)\,\middle\|\,\mathrm{Ber}(0.5)\right)\right] \\
&= \mathbb{E}_{Y_{1:n}}\left[\sum_{j=1}^k \left(\mathop{\mathbb{E}}_{V_j\sim q_j(\cdot|Y_{1:n})}\left[\log\big(2q_j(V_j\,|\,Y_{1:n})\big)\right]\right)\right] \\
&= \mathbb{E}_{Y_{1:n}}\left[\mathop{\mathbb{E}}_{V_{1:k}\sim q(\cdot|Y_{1:n})}\left[\log\left(2^k\prod_{j=1}^k q_j(V_j\,|\,Y_{1:n})\right)\right]\right] \\
&= \mathbb{E}_{Y_{1:n}}\left[\sum_{V_{1:k}} q(V_{1:k}\,|\,Y_{1:n})\log\big(2^k q(V_{1:k}\,|\,Y_{1:n})\big)\right].
\end{aligned}$$

Recognizing that $q(V_{1:k}\,|\,Y_{1:n})$ is an approximation of $p(V_{1:k}\,|\,Y_{1:n}, U)$ (recall the arguments in (185)-(187)), we consider the following decomposition by incorporating the term $p(V_{1:k}\,|\,Y_{1:n}, U)$ into the preceding identity:

$$\begin{aligned}
2\mathbb{E}\left[\sum_{j=1}^k (2q_j-1)^2\right] &= \mathbb{E}_{Y_{1:n},U}\left[\sum_{V_{1:k}} p(V_{1:k}\mid Y_{1:n},U)\frac{q(V_{1:k}\,|\,Y_{1:n})}{p(V_{1:k}\mid Y_{1:n},U)}\log\left(2^k q(V_{1:k}\,|\,Y_{1:n})\right)\right] \\
&= \mathbb{E}_{Y_{1:n},U}\left[\sum_{V_{1:k}} \frac{q(V_{1:k}\mid Y_{1:n})}{p(V_{1:k}\mid Y_{1:n},U)}\log\big(2^k p(V_{1:k}\mid Y_{1:n},U)\big)p(V_{1:k}\mid Y_{1:n},U)\right] \\
&\quad + \mathbb{E}_{Y_{1:n},U}\left[\sum_{V_{1:k}} \frac{q(V_{1:k}\mid Y_{1:n})}{p(V_{1:k}\mid Y_{1:n},U)}\log\left(\frac{q(V_{1:k}\mid Y_{1:n})}{p(V_{1:k}\mid Y_{1:n},U)}\right)p(V_{1:k}\mid Y_{1:n},U)\right].
\end{aligned} \tag{198}$$

Next, we examine the ratio term $\frac{q(V_{1:k}|Y_{1:n})}{p(V_{1:k}|Y_{1:n},U)}$, which appears multiple times in (198). From the elementary fact $1+x \le e^x$ as well as the definitions (185) and (187), we can demonstrate that

$$\begin{aligned}
\frac{q(V_{1:k} \mid y_{1:n})}{p(V_{1:k} \mid y_{1:n}, u)} &= \frac{\prod_{j=1}^k \left((1+\epsilon V_j)^{N_j} e^{-\epsilon V_j \frac{n}{k}}\right)}{\sum_{v_{1:k}} \prod_{j=1}^k \left((1+\epsilon v_j)^{N_j} e^{-\epsilon v_j \frac{n}{k}}\right)} \cdot \frac{\sum_{v_{1:k}} \left(\prod_{j=1}^k (1+\epsilon v_j)^{N_j}\right)(1+\epsilon \overline{v})^{-n}}{\left(\prod_{j=1}^k (1+\epsilon V_j)^{N_j}\right)(1+\epsilon \overline{V})^{-n}} \\
&= \frac{(1+\epsilon\overline{V})^n}{e^{\epsilon \overline{V} n}} \cdot \frac{\sum_{v_{1:k}} \left(\prod_{j=1}^k (1+\epsilon v_j)^{N_j}\right)(1+\epsilon \overline{v})^{-n}}{\sum_{v_{1:k}} \prod_{j=1}^k \left((1+\epsilon v_j)^{N_j} e^{-\epsilon v_j \frac{n}{k}}\right)} = \frac{(1+\epsilon\overline{V})^n}{e^{\epsilon \overline{V} n}} \Bigg/ \left(\sum_{v_{1:k}} p(v_{1:k} \mid y_{1:n}, u) \frac{(1+\epsilon\overline{v})^n}{e^{\epsilon \overline{v} n}}\right) \\
&= \frac{F_n(\epsilon \overline{V})}{\mathbb{E}_{V_{1:k} \sim p(\cdot|y_{1:n},u)}\left[F_n(\epsilon \overline{V})\right]} = \frac{F_n(\epsilon \overline{V})}{F(y_{1:n},u)} \le \frac{1}{F(y_{1:n},u)},
\end{aligned} \tag{199}$$

where we define

$$F_n(\epsilon \overline{V}) \coloneqq \frac{(1+\epsilon\overline{V})^n}{e^{\epsilon \overline{V} n}} \qquad \text{and} \qquad F(y_{1:n},u) \coloneqq \mathbb{E}_{V_{1:k} \sim p(\cdot|y_{1:n},u)}\left[F_n(\epsilon \overline{V})\right]. \tag{200}$$

Further, given that the choice of ϵ guarantees that $\left|\epsilon \overline{V}\right| < 1/2$, we can invoke the elementary inequality $1+x \le e^x \le 1+x+x^2$ $(|x|<1/2)$ to derive

$$1 \le \frac{1}{F_n(\epsilon \overline{V})} = \frac{e^{\epsilon \overline{V} n}}{\left(1+\epsilon \overline{V}\right)^n} \le \left(\frac{1+\epsilon\overline{V}+\epsilon^2\overline{V}^2}{1+\epsilon\overline{V}}\right)^n \le \left(1+2\epsilon^2\overline{V}^2\right)^n \le \exp\left\{2\left(\epsilon\overline{V}\right)^2 n\right\}. \tag{201}$$

Substituting (199) and (201) into (198) leads to

$$\begin{aligned}
2\mathbb{E}\left[\sum_{j=1}^k (2q_j-1)^2\right] &= \mathbb{E}_{Y_{1:n},U}\left[\sum_{V_{1:k}} \frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)} \log\left(2^k p(V_{1:k} \mid Y_{1:n},U)\right) p(V_{1:k} \mid Y_{1:n},U)\right] \\
&\quad + \mathbb{E}_{Y_{1:n},U}\left[\sum_{V_{1:k}} \frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)} \log\left(\frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)}\right) p(V_{1:k} \mid Y_{1:n},U)\right] \\
&= \underbrace{\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{V_{1:k} \sim p(\cdot|Y_{1:n},U)}\left[\log\left(2^k p(V_{1:k} \mid Y_{1:n},U)\right)\right]\right]}_{=:\mathcal{S}_0} \\
&\quad + \underbrace{\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{V_{1:k} \sim p(\cdot|Y_{1:n},U)}\left[\left(\frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)}-1\right)\log\left(2^k p(V_{1:k} \mid Y_{1:n},U)\right)\right]\right]}_{=:\mathcal{S}_1} \\
&\quad + \underbrace{\mathbb{E}_{Y_{1:n},U}\left[\mathbb{E}_{V_{1:k} \sim p(\cdot|Y_{1:n},U)}\left[\frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)}\log\left(\frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)}\right)\right]\right]}_{=:\mathcal{S}_2},
\end{aligned} \tag{202}$$

leaving us with three terms to cope with.

Bounding the term $\mathcal{S}_0$. Towards this end, we first define $I(V_{1:k};Y_{1:n})$ to be the mutual information between $V_{1:k}$ and $Y_{1:n}$, namely,

$$I(V_{1:k};Y_{1:n}) \coloneqq \sum_{V_{1:k}} \int_{[0,1]^n} \log\left(\frac{p(V_{1:k},y_{1:n})}{2^{-k}p(y_{1:n})}\right) p(V_{1:k},y_{1:n})\,\mathrm{d}y_{1:n}, \tag{203}$$

where

$$p(V_{1:k},y_{1:n}) \coloneqq \int_{\mathcal{U}} p(V_{1:k},y_{1:n},u)\,\mathrm{d}u \qquad \text{and} \qquad p(y_{1:n}) \coloneqq \sum_{V_{1:k}} p(V_{1:k},y_{1:n}). \tag{204}$$

As it turns out, the term $\mathcal{S}_0$ (cf. (202)) is equivalent to this mutual information quantity since

$$\begin{aligned}\mathcal{S}_0 &= \mathbb{E}_{Y_{1:n},U}\left[\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\log\left(2^k p(V_{1:k}\,|\,Y_{1:n},U)\right)\right]\right] \\ &= \sum_{V_{1:k}}\int_{[0,1]^n\times\mathcal{U}} p(V_{1:k},y_{1:n})p_U(u)\log\Big(\frac{p(V_{1:k},y_{1:n})p_U(u)}{2^{-k}p(y_{1:n})p_U(u)}\Big)\mathrm{d}y_{1:n}\mathrm{d}u \\ &= \sum_{V_{1:k}}\int_{[0,1]^n} p(V_{1:k},y_{1:n})\log\Big(\frac{p(V_{1:k},y_{1:n})}{2^{-k}p(y_{1:n})}\Big)\mathrm{d}y_{1:n} = I(V_{1:k};Y_{1:n}),\end{aligned}$$

where the second line holds since U is independent of $(V_{1:k},Y_{1:n})$ (as it only affects the construction of the prediction set $\mathcal{C}$).

According to the data-generating mechanism and the chain rule of the KL divergence, one has

$$\begin{aligned}I(V_{1:k},Y_{1:n}) &= \sum_{V_{1:k}}\int_{[0,1]^n} p(V_{1:k})p(y_{1:n}\mid V_{1:k})\log\frac{p(V_{1:k})p(y_{1:n}\mid V_{1:k})}{p(V_{1:k})p(y_{1:n})}\mathrm{d}y_{1:n} \\ &= \sum_{i=1}^n\sum_{V_{1:k}}\int_{[0,1]}\frac{p(y_i\mid V_{1:k})}{2^k}\log\frac{p(y_i\mid V_{1:k})}{p(y_i)}\mathrm{d}y_i \\ &\overset{\text{(a)}}{\le} \sum_{i=1}^n\sum_{V_{1:k}}\frac{1}{2^k}\int_{[0,1]}\left(p(y_i\mid V_{1:k})-p(y_i)+p(y_i\mid V_{1:k})\left(\frac{p(y_i)}{p(y_i\mid V_{1:k})}-1\right)^2\right)\mathrm{d}y_i \\ &\overset{\text{(b)}}{=} \sum_{i=1}^n\sum_{V_{1:k}}\int_{[0,1]}\frac{p(y_i\mid V_{1:k})}{2^k}\left(\frac{p(y_i)}{p(y_i\mid V_{1:k})}-1\right)^2\mathrm{d}y_i \overset{\text{(c)}}{\le} 36n\epsilon^2,\end{aligned} \tag{205}$$

where (b) follows because both $p(y_i)$ and $p(y_i\mid V_{1:k})$ are density functions; (a) and (c) are proven below.

- To justify (c), suppose that $y_i\in I_j$ (defined in (154)). Denoting by $V'_{1:k}$ an independent copy of $V_{1:k}$, one can derive

$$\begin{aligned}\left(\frac{p(y_i)}{p(y_i\mid V_{1:k})}-1\right)^2 &= \left(\frac{\mathbb{E}_{V'_{1:k}}[p(y_i\mid V'_{1:k})]}{p(y_i\mid V_{1:k})}-1\right)^2 \\ &\le \mathbb{E}_{V'_{1:k}}\left[\left(\frac{p(y_i\mid V'_{1:k})}{p(y_i\mid V_{1:k})}-1\right)^2\right] = \mathbb{E}_{V'_{1:k}}\left[\left(\frac{(1+\epsilon V'_j)(1+\epsilon\overline{V})}{(1+\epsilon V_j)(1+\epsilon\overline{V}')}-1\right)^2\right] \\ &= \mathbb{E}_{V'_{1:k}}\left[\left(\frac{\epsilon(V'_j+\overline{V}-V_j-\overline{V}')+\epsilon^2(V'_j\overline{V}-V_j\overline{V}')}{(1+\epsilon V_j)(1+\epsilon\overline{V}')}\right)^2\right] \le 36\epsilon^2,\end{aligned}$$

  with the proviso that $\epsilon\le 1/6$. This also implies that

$$\left|\frac{p(y_i)}{p(y_i\mid V_{1:k})}-1\right|\le 1 \qquad \text{as long as } \epsilon\le\frac{1}{6}. \tag{206}$$

- Regarding inequality (a), we invoke the elementary fact $\log(1+x)\ge x-x^2$ for $x\in[-1,1]$ along with (206) to reach

$$\begin{aligned}p(y_i\mid V_{1:k})\log\frac{p(y_i\mid V_{1:k})}{p(y_i)} &= -p(y_i\mid V_{1:k})\log\frac{p(y_i)}{p(y_i\mid V_{1:k})} \\ &\le -p(y_i\mid V_{1:k})\left(\frac{p(y_i)}{p(y_i\mid V_{1:k})}-1-\left(\frac{p(y_i)}{p(y_i\mid V_{1:k})}-1\right)^2\right) \\ &= p(y_i\mid V_{1:k})-p(y_i)+p(y_i\mid V_{1:k})\left(\frac{p(y_i)}{p(y_i\mid V_{1:k})}-1\right)^2.\end{aligned}$$

Bounding the term $\mathcal{S}_1$. Based on the definition (cf. (202)), we can write

$$\mathcal{S}_1 = \mathbb{E}_{Y_{1:n},U}\Bigg[\underbrace{\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\Big(F_n(\epsilon\overline{V}) - F(Y_{1:n},U)\Big)\log\big(2^k p(V_{1:k}\,|\,Y_{1:n},U)\big)\Big]}_{=:\,\mathcal{S}_1^{\mathsf{num}}(Y_{1:n},U)}\Big/F(Y_{1:n},U)\Bigg]. \tag{207}$$

We begin by examining $\mathcal{S}_1^{\mathsf{num}}(Y_{1:n},U)$. For notational simplicity, set

$$\sigma \coloneqq \sqrt{16k^{-1}\log(2nk/\alpha^2)}. \tag{208}$$

Consider any given $y_{1:n}, u$, and any two realizations $v_{1:k}$, $v'_{1:k}$. If $|\overline{v}| \vee |\overline{v}'| \le \sigma$, then by the mean value theorem, there exists a real number ζ between $\epsilon\overline{v}$ and $\epsilon\overline{v}'$ (which means $|\zeta| \le \epsilon\big(|\overline{v}| \vee |\overline{v}'|\big) \le \epsilon\sigma$) satisfying

$$\Big|\,F_n(\epsilon\overline{v}) - F_n(\epsilon\overline{v}')\,\Big| \le \big|\,F_n'(\zeta)\,\big|\cdot\big|\,\epsilon\overline{v} - \epsilon\overline{v}'\,\big| \le 2\epsilon n\sigma|\zeta|e^{-\zeta}\Big(\frac{1+\zeta}{e^{\zeta}}\Big)^{n-1} \le 6\epsilon^2\sigma^2 n, \tag{209}$$

where the choice of ϵ ensures $\epsilon\sigma < 1$ and we have used the definition (200) of $F_n(\cdot)$. Splitting the expectation in the definition of $\mathcal{S}_1^{\mathsf{num}}(Y_{1:n},U)$ into two parts based on whether or not $|\overline{V}| \vee |\overline{V}'| \le \sigma$, we can derive

$$\begin{aligned}
\mathcal{S}_1^{\mathsf{num}}(Y_{1:n},U) &= \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\Big(F_n(\epsilon\overline{V}) - \mathop{\mathbb{E}}_{V'_{1:k}\sim p(\cdot|Y_{1:n},U)}\big[F_n\big(\epsilon\overline{V}'\big)\big]\Big)\log\big(2^k p(V_{1:k}\mid Y_{1:n},U)\big)\Big] \\
&\le \mathop{\mathbb{E}}_{V_{1:k},V'_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\Big|F_n(\epsilon\overline{V}) - F_n\big(\epsilon\overline{V}'\big)\Big|\log\Big(\frac{1}{p(V_{1:k}\mid Y_{1:n},U)}\Big)\Big] \\
&\overset{(209)}{\le} 6\epsilon^2\sigma^2 n \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\log\frac{1}{p(V_{1:k}\mid Y_{1:n},U)}\Big] \\
&\quad + \mathop{\mathbb{E}}_{V_{1:k},V'_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\Big|F_n(\epsilon\overline{V}) - F_n\big(\epsilon\overline{V}'\big)\Big|\mathbb{1}\{|\overline{V}| \vee |\overline{V}'| > \sigma\}\log\Big(\frac{1}{p(V_{1:k}\mid Y_{1:n},U)}\Big)\Big],
\end{aligned} \tag{210}$$

where the second line follows from Jensen's inequality and the following:

$$\mathop{\mathbb{E}}_{V_{1:k},V'_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\Big(F_n(\epsilon\overline{V}) - F_n\big(\epsilon\overline{V}'\big)\Big)\log\big(2^k\big)\Big] = 0.$$

- Regarding the first term on the right-hand side of (210), it is observed that for any given $y_{1:n}$ and u,

$$\begin{aligned}
\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|y_{1:n},u)}\Big[\log\frac{1}{p(V_{1:k}\mid y_{1:n},u)}\Big] &= \log 2^k - \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|y_{1:n},u)}\big[\log\big(2^k p(V_{1:k}\mid y_{1:n},u)\big)\big] \\
&= k\log 2 - \mathsf{KL}\big(p(V_{1:k}\mid y_{1:n},u)\,\|\,p(V_{1:k})\big) \le k\log 2.
\end{aligned}$$

  Plug this into the first term on the right-hand side of (210) and use (208) to yield

$$6\epsilon^2\sigma^2 n \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\log\frac{1}{p(V_{1:k}\mid Y_{1:n},U)}\Big] \le 96\epsilon^2 n\log\frac{2nk}{\alpha^2}. \tag{211}$$

- As for the second term on the right-hand side of (210), it can be derived that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{V_{1:k},V'_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\Big|F_n(\epsilon\overline{V}) - F_n\big(\epsilon\overline{V}'\big)\Big|\mathbb{1}\{|\overline{V}| \vee |\overline{V}'| > \sigma\}\log\Big(\frac{1}{p(V_{1:k}\mid Y_{1:n},U)}\Big)\Big] \\
&\overset{(201)}{\le} \mathop{\mathbb{E}}_{V_{1:k},V'_{1:k}\sim p(\cdot|Y_{1:n},U)}\Big[\mathbb{1}\Big\{|\overline{V}| \vee |\overline{V}'| > \sigma\Big\}\log\Big(\frac{1}{p(V_{1:k}\mid Y_{1:n},U)}\Big)\Big] \\
&\overset{(a)}{\le} 4\epsilon n\mathbb{P}\Big(|\overline{V}| \vee |\overline{V}'| > \sigma\,\Big|\,Y_{1:n},U\Big).
\end{aligned} \tag{212}$$

To justify step (a) of (212), consider any two different realization $v_{1:k}$ and $v'_{1:k}$, which obey

$$\frac{p(v'_{1:k} \mid y_{1:n}, u)}{p(v_{1:k} \mid y_{1:n}, u)} = \frac{p(v'_{1:k})p(y_{1:n} \mid v'_{1:k})p_U(u)}{p(v_{1:k})p(y_{1:n} \mid v_{1:k})p_U(u)} = \frac{\prod_{j=1}^k (1+\epsilon v'_j)^{N_j}}{\prod_{j=1}^k (1+\epsilon v_j)^{N_j}} \le e^{4\epsilon n},$$

$$\Longrightarrow \qquad \log\Big(\frac{1}{p(V_{1:k} \mid y_{1:n}, u)}\Big) = \log\left(\mathop{\mathbb{E}}_{V'_{1:k}|y_{1:n},u}\left[\frac{p(V'_{1:k} \mid y_{1:n}, u)}{p(V_{1:k} \mid y_{1:n}, u)}\right]\right) \le 4\epsilon n.$$

Consequently, by substituting (212) and (211) into (210), we arrive at

$$\mathcal{S}_1^{\mathsf{num}}(Y_{1:n}, U) \le 96\epsilon^2 n \log\frac{2nk}{\alpha^2} + 8\epsilon n \mathbb{P}\Big(|\overline{V}| > \sigma \,\Big|\, Y_{1:n}, U\Big). \tag{213}$$

Now, we switch attention to the term $\frac{1}{F(Y_{1:n},U)}$. Applying Jensen's inequality yields

$$\begin{aligned}
\frac{1}{\mathbb{E}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[F_n(\epsilon\overline{V})\right]} &\le \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{e^{\epsilon\overline{V}n}}{\left(1+\epsilon\overline{V}\right)^n}\right] \\
&= \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{e^{\epsilon\overline{V}n}}{\left(1+\epsilon\overline{V}\right)^n}\mathbb{1}\left\{|\overline{V}| \le \sigma\right\}\right] + \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{e^{\epsilon\overline{V}n}}{\left(1+\epsilon\overline{V}\right)^n}\mathbb{1}\left\{|\overline{V}| > \sigma\right\}\right] \\
&\overset{(201)}{\le} \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[e^{2\left(\epsilon\overline{V}\right)^2 n}\mathbb{1}\left\{|\overline{V}| \le \sigma\right\}\right] + \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[e^{2\left(\epsilon\overline{V}\right)^2 n}\mathbb{1}\left\{|\overline{V}| > \sigma\right\}\right] \\
&\le 3 + \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\exp\left\{2\left(\epsilon\overline{V}\right)^2 n\right\}\mathbb{1}\left\{|\overline{V}| > \sigma\right\}\right],
\end{aligned} \tag{214}$$

where the last line follows since $2\left(\epsilon\overline{V}\right)^2 n \le \frac{2k}{32n}\cdot\frac{16n}{k} = 1$ provided that $\epsilon \le \min\left\{\frac{\alpha^{5/2}}{128}, \frac{1}{64}\sqrt{\frac{\alpha k}{n\log(2nk/\alpha^2)}}\right\}$.

We are now ready to bound $\mathcal{S}_1$. It is readily seen from (207) that

$$\begin{aligned}
\mathcal{S}_1 &\le \mathbb{E}_{Y_{1:n},U}\left[\frac{\mathcal{S}_1^{\mathsf{num}}(Y_{1:n},U)}{F(Y_{1:n},U)}\right] \overset{(213)}{\le} \mathbb{E}_{Y_{1:n},U}\left[\frac{96\epsilon^2 n\log\left(\frac{2nk}{\alpha^2}\right) + 8\epsilon n\mathbb{P}\Big(|\overline{V}| > \sigma \mid Y_{1:n},U\Big)}{F(Y_{1:n},U)}\right] \\
&\overset{(214)}{\le} \mathop{\mathbb{E}}_{Y_{1:n},U}\left[\left(96\epsilon^2 n\log\left(\frac{2nk}{\alpha^2}\right) + 8\epsilon n\mathbb{P}\Big(|\overline{V}| > \sigma \mid Y_{1:n},U\Big)\right)\left(3 + \mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[e^{2(\epsilon\overline{V})^2 n}\mathbb{1}\{|\overline{V}| > \sigma\}\right]\right)\right] \\
&\le 100\epsilon^2 n\log\left(\frac{2nk}{\alpha^2}\right)\left(3 + \mathbb{E}_{V_{1:k}}\left[e^{2(\epsilon\overline{V})^2 n}\mathbb{1}\{|\overline{V}| > \sigma\}\right]\right) + 24\epsilon n\mathbb{P}\big(|\overline{V}| > \sigma\big) \\
&\quad + 8\epsilon n \mathop{\mathbb{E}}_{Y_{1:n},U}\left[\mathbb{P}\Big(|\overline{V}| > \sigma \mid Y_{1:n},U\Big)\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[e^{2(\epsilon\overline{V})^2 n}\mathbb{1}\{|\overline{V}| > \sigma\}\right]\right].
\end{aligned} \tag{215}$$

Note that Hoeffding's inequality tells us that

$$\mathbb{P}_{V_{1:k}}\big(|\overline{V}| > \sigma\big) = \mathbb{P}_{V_{1:k}}\left(|\overline{V}| > \sqrt{\frac{16\log(2nk/\alpha^2)}{k}}\right) \le \frac{\alpha^{16}}{128n^8k^8}, \tag{216}$$

we also have for any $\Delta \ge 0$ that

$$\begin{aligned}
\mathbb{E}_{V_{1:k}}\left[e^{\frac{k}{4}\overline{V}^2}\mathbb{1}\left\{e^{\frac{k}{4}\overline{V}^2} > e^{\frac{k}{4}\Delta^2}\right\}\right] &\overset{(a)}{=} e^{\frac{k}{4}\Delta^2}\mathbb{P}_{V_{1:k}}\left(e^{\frac{k}{4}\overline{V}^2} > e^{\frac{k}{4}\Delta^2}\right) + \int_{e^{\frac{k}{4}\Delta^2}}^{\infty}\mathbb{P}_{V_{1:k}}\left(e^{\frac{k}{4}\overline{V}^2} > y\right)\mathrm{d}y \\
&= e^{\frac{k\Delta^2}{4}}\mathbb{P}_{V_{1:k}}\Big(|\overline{V}| > \Delta\Big) + \int_{e^{\frac{k\Delta^2}{4}}}^{\infty}\mathbb{P}\left(|\overline{V}| > \sqrt{\frac{4\log y}{k}}\right)\mathrm{d}y \\
&\le e^{-\frac{k\Delta^2}{4}} + \int_{e^{\frac{k\Delta^2}{4}}}^{\infty} 2e^{-2\log y}\mathrm{d}y \le e^{-\frac{k\Delta^2}{4}} - 2y^{-1}\Big|_{e^{\frac{k\Delta^2}{4}}}^{\infty} = 3e^{-\frac{k\Delta^2}{4}},
\end{aligned} \tag{217}$$

where (a) follows from Fubini's formula, namely, for any non-negative random variable X with CDF F_X and any $x \geq 0$,

$$\begin{aligned}\mathbb{E}[X\mathbb{1}\{X > x_0\}] &= \int_{x_0}^{\infty} xF_X(\mathrm{d}x) = \int_{x_0}^{\infty}\left(\int_0^x \mathrm{d}t\right)F_X(\mathrm{d}x) = \int_0^{\infty}\left(\int_{x_0\vee t}^{\infty} F_X(\mathrm{d}x)\right)\mathrm{d}t \\ &= \int_0^{x_0}\mathbb{P}(X > x_0)\mathrm{d}t + \int_{x_0}^{\infty}\mathbb{P}(X > t)\mathrm{d}t = x_0\mathbb{P}(X > x_0) + \int_{x_0}^{\infty}\mathbb{P}(X > t)\mathrm{d}t.\end{aligned}$$

Further, when $\epsilon \leq \frac{1}{64}\sqrt{\frac{\alpha k}{n\log(2nk/\alpha^2)}}$, one has

$$2\epsilon^2\overline{V}^2 n \leq \frac{k}{128n}\overline{V}^2 n \leq \frac{k}{4}\overline{V}^2.$$

Combining this with (216) and (217) (with Δ set to σ), and substituting the resulting bounds into (215), yields

$$\begin{aligned}\mathcal{S}_1 &\leq 100\epsilon^2 n\log\left(\frac{2nk}{\alpha^2}\right)\left(3 + 3e^{-\frac{k\sigma^2}{4}}\right) + \frac{\alpha^{16}}{4n^7k^8} + 8\epsilon n\mathbb{E}_{V_{1:k}}\left[e^{\frac{k}{4}\overline{V}^2}\mathbb{1}\{|\overline{V}| > \sigma\}\right] \\ &\overset{(217)}{\leq} 600\epsilon^2 n\log\left(\frac{2nk}{\alpha^2}\right) + \frac{\alpha^{16}}{4n^7k^8} + 24\epsilon n e^{-\frac{k\sigma^2}{4}} \\ &\overset{(208)}{\leq} 600\epsilon^2 n\log\left(\frac{2nk}{\alpha^2}\right) + \frac{\alpha^{16}}{4n^7k^8} + \frac{\alpha^8}{4n^3k^4}.\end{aligned} \tag{218}$$

Bounding the term $\mathcal{S}_2$. We first rewrite $\mathcal{S}_2$ (cf. (202)) slightly by using (201) as follows:

$$\begin{aligned}\mathcal{S}_2 &= \mathbb{E}_{Y_{1:n},U}\left[\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)}\log\left(\frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)}\right)\right]\right] \\ &\overset{(201)}{\leq} \mathbb{E}_{Y_{1:n},U}\left[\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{F_n(\epsilon\overline{V})}{F(Y_{1:n},U)}\log\left(\frac{1}{F(Y_{1:n},U)}\right)\right]\right] \\ &\leq \mathbb{E}_{Y_{1:n},U}\left[\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{1}{F(Y_{1:n},U)}\log\left(\frac{1}{F(Y_{1:n},U)}\right)\right]\right],\end{aligned}$$

where the last line holds since $\log\left(\frac{1}{F(Y_{1:n},U)}\right) \geq 0$. It can be easily verified that the function $\frac{1}{x}\log\frac{1}{x}$ is convex when $0 < x \leq 1$. Further, since $F_n(\epsilon\overline{v}) = \frac{(1+\epsilon\overline{v})^n}{e^{\epsilon\overline{v}n}} \leq 1$, one can invoke Jensen's inequality to obtain

$$\begin{aligned}\mathcal{S}_2 &\leq \mathbb{E}_{Y_{1:n},U}\left[\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{1}{F(Y_{1:n},U)}\log\left(\frac{1}{F(Y_{1:n},U)}\right)\right]\right] \\ &= \mathbb{E}_{Y_{1:n},U}\left[\frac{1}{F(Y_{1:n},U)}\log\left(\frac{1}{F(Y_{1:n},U)}\right)\right] \leq \mathbb{E}_{Y_{1:n},U}\left[\mathop{\mathbb{E}}_{V_{1:k}\sim p(\cdot|Y_{1:n},U)}\left[\frac{1}{F_n(\epsilon\overline{V})}\log\frac{1}{F_n(\epsilon\overline{V})}\right]\right] \\ &\leq \mathbb{E}_{V_{1:k}}\left[\left(2\left(\epsilon\overline{V}\right)^2 n\right)\exp\left\{2\left(\epsilon\overline{V}\right)^2 n\right\}\right],\end{aligned}$$

where the last inequality follows from (201), together with the fact that the function $x \mapsto x\log x$ is monotonically increasing on $[1,\infty)$. By letting Δ in (217) be 0, we arrive at

$$\mathcal{S}_2 \leq 2\epsilon^2 n\mathbb{E}_{V_{1:k}}\left[e^{\frac{k}{4}\overline{V}^2}\right] \leq 6\epsilon^2 n. \tag{219}$$

Putting all this together. Finally, combining (205), (218) and (219) with (202) yields

$$2\mathbb{E}\left[\sum_{j=1}^{k}(2q_j - 1)^2\right] \leq \mathcal{S}_0 + \mathcal{S}_1 + \mathcal{S}_2 \leq 642n\epsilon^2\log\left(\frac{2nk}{\alpha^2}\right) + \frac{\alpha^6}{16n^3k^3},$$

thereby concluding the proof of Lemma C.10.

C.5 Proof of Proposition 4.2 and Proposition 4.3

C.5.1 Proof of Proposition 4.2

Denote $p_{\mathcal{D}}(\cdot)$ as the density function of the distribution $\mathcal{D}$. According to the definition of Riemann-integrability, for every $\varepsilon > 0$, the following holds for large enough n:

$$\begin{aligned}\left|\mathbb{P}(Y \in \mathcal{C}_n) - (1-\alpha)\right| &\leq \left|\mathbb{P}(Y \in \mathcal{C}_n) - (1-\alpha)\sum_{i=0}^{n-1} p_{\mathcal{D}}(i/n)\right| + (1-\alpha)\left|\sum_{i=0}^{n-1}\int_{i/n}^{(i+1)/n}\big(p_{\mathcal{D}}(i/n) - p_{\mathcal{D}}(y)\big)\mathrm{d}y\right| \\ &\leq \left|\sum_{i=0}^{n-1}\mathbb{P}\Big(Y \in \big[i/n, \big(i+(1-\alpha)\big)/n\big]\Big) - (1-\alpha)\sum_{i=0}^{n-1} p_{\mathcal{D}}(i/n)\right| + \varepsilon \\ &= \left|\sum_{i=0}^{n-1}\int_{i/n}^{(i+(1-\alpha))/n}\big(p_{\mathcal{D}}(y) - p_{\mathcal{D}}(i/n)\big)\mathrm{d}y\right| + \varepsilon \leq (2-\alpha)\varepsilon.\end{aligned}$$

This immediately establishes the result of Proposition 4.2.

C.5.2 Proof of Proposition 4.3

The claim follows immediately by invoking Lemma C.6 with

$$k \coloneqq \frac{256K}{\alpha} \quad \text{and} \quad \epsilon \coloneqq \min\left\{\frac{\alpha^{5/2}}{200}, \frac{1}{64}\sqrt{\frac{\alpha k}{n\log\big(2nk/\alpha^2\big)}}\right\}.$$

D Examples of stable learning algorithms

To illustrate the applicability of Assumption 4.3, this section verifies it for several learning algorithms commonly used in statistical applications. In particular, Section D.3 describes how to incorporate stochastic optimization methods into our online conformal framework: the fitted model can be updated incrementally using the newly arrived data at each time step, without retraining from scratch.

D.1 Constrained M-estimation

We begin with the classical constrained M-estimator (Van der Vaart, 2000) that minimizes the empirical loss:

$$\widehat{\vartheta}_n = \operatorname*{argmin}_{\vartheta\in\mathcal{C}} \widehat{L}_n(\vartheta) \coloneqq \operatorname*{argmin}_{\vartheta\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}\ell(\vartheta; Z_i),$$

where $\{Z_i\} \subset \mathbb{R}^{d_Z}$ denote n independent data samples with Z_i drawn from the distribution $\mathcal{D}_i$, $\mathcal{C} \subset \mathbb{R}^d$ represents a closed convex constraint set, and $\ell(\cdot; z)$ is a loss metric assumed to be differentiable in ϑ for every z. We also introduce the population risk

$$L(\vartheta) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{Z_i\sim\mathcal{D}_i}\big[\ell(\vartheta; Z_i)\big].$$

Note that we allow for a non-identically distributed sequence $\{Z_i\}_{i=1}^n$, which is compatible with the drifting environments considered in the present paper.

We impose the following standard assumptions ensuring well-posedness and curvature of the loss functions.

Assumption D.1. *Suppose that the loss functions satisfy the following properties:*

1. *L is μ-strongly convex on $\mathcal{C}$ for some $\mu > 0$, in the sense that*

$$\nabla^2 L(\vartheta) \succeq \mu I_d \qquad \text{for all } \vartheta \in \mathcal{C}.$$

2. *For any* $\vartheta \in \mathcal{C}$, $\|\vartheta\|_2 \le D_{\mathcal{C}}$.

3. *For any* $\vartheta, \vartheta' \in \mathcal{C}$ *and* Z, $\|\nabla_\vartheta \ell(\vartheta; Z)\| \le \beta_L$ *and* $\left\|\nabla^2_\vartheta \ell(\vartheta; Z) - \nabla^2_\vartheta \ell(\vartheta'; Z)\right\|_{\infty,\infty} \le \beta_s \|\vartheta - \vartheta'\|_2$. *Here* $\|A\|_{\infty,\infty} := \max_{i,j\in[d]} |A_{ij}|$.

With Assumption D.1 in place, we obtain the following stability guarantees for the constrained M-estimator.

Proposition D.1. *Suppose that Assumption D.1 holds. For any* $n \ge \frac{32 d \beta_s^2 D_{\mathcal{C}}^2}{\mu^2} \log\left(\frac{12 \beta_s D_{\mathcal{C}}}{\mu}\right)$, *We can find a typical set* $\mathcal{E}$ *in the* n*-sample space* $\mathbb{R}^{d_Z \times n}$ *obeying* $\mathbb{P}(\mathcal{E}^c) \le d^2 \exp\left\{-\frac{\mu^2 n}{32 \beta_s^2 D_{\mathcal{C}}^2}\right\}$ *such that the following holds: consider two adjacent datasets in* $\mathcal{E}$, $\{Z_i\}_{i=1}^n$ *and* $\{Z_i'\}_{i=1}^n$, *that differ only in the last coordinate, i.e.,* $Z_i' = Z_i$ *for* $i = 1, \dots, n-1$ *and* $Z_n' \ne Z_n$, *and denote by* $\widehat{\vartheta}_n$ *and* $\widehat{\vartheta}_n'$ *the corresponding constrained M-estimates*

$$\widehat{\vartheta}_n := \operatorname*{argmin}_{\vartheta \in \mathcal{C}} \left\{ \sum_{i=1}^n \ell(\vartheta; Z_i) \right\}, \quad \widehat{\vartheta}_n' = \operatorname*{argmin}_{\vartheta \in \mathcal{C}} \left\{ \sum_{i=1}^{n-1} \ell(\vartheta; Z_i) + \ell(\vartheta; Z_n') \right\},$$

then one has

$$\left\|\widehat{\vartheta}_n - \widehat{\vartheta}_n'\right\|_2 \le \frac{4\beta_L}{\mu n}.$$

The proof of Proposition D.1 is deferred to Section D.4.1. With the above proposition in hand, consider a prediction model $\mu(\cdot \mid \vartheta)$. Assume that, for every input x, the mapping $\vartheta \to \mu(x \mid \vartheta)$ is L_0-Lipschitz. Then, by Proposition D.1, for any two neighboring data sequences $\{Z_n\}$ and $\{Z_n'\}$ that both lie in $\mathcal{E}$ (which happens with high probability), the corresponding fitted models $\mu(\cdot \mid \widehat{\vartheta})$ and $\mu(\cdot \mid \widehat{\vartheta}')$ satisfy:

$$\left|\mu(x \mid \widehat{\vartheta}) - \mu(x \mid \widehat{\vartheta}')\right| \le L_0 \left\|\widehat{\vartheta} - \widehat{\vartheta}'\right\|_2 \le \frac{4\beta_L L_0}{\mu n}, \quad \forall x \in \mathcal{X},$$

thereby validating Assumption 4.3 for this setting with $L_2 = \frac{4\beta_L L_0}{\mu}$.

D.2 Linear stochastic approximation

Next, we verify stability for an important class of *online* learning methods based on stochastic approximation. Unlike the previous example in Section D.1, where the estimator is obtained via empirical risk minimization, stochastic approximation updates the parameter incrementally as new data arrive (Bottou et al., 2018). This makes it particularly well-suited to our online conformal framework, as it avoids retraining from scratch while still enabling control over the sensitivity of the fitted model to individual observations.

To be concrete, consider a linear prediction model

$$\widehat{\mu}(x \mid Z_{1:n}) = x^\top \vartheta_n,$$

where the parameter $\vartheta_n \in \mathbb{R}^d$ is updated online using the data $Z_{1:n} = \{(X_i, Y_i)\}_{i=1}^n$. At iteration n, the squared loss function is defined as

$$\ell_n(\vartheta) = (Y_n - X_n^\top \vartheta)^2 = \vartheta^\top X_n X_n^\top \vartheta - 2 Y_n X_n^\top \vartheta + Y_n^2.$$

Its gradient at ϑ_n is given by

$$\nabla \ell_n(\vartheta_n) = \underbrace{2 X_n X_n^\top}_{=: \widehat{A}_n} \vartheta_n - \underbrace{2 Y_n X_n}_{=: \widehat{b}_n} = \widehat{A}_n \vartheta_n - \widehat{b}_n.$$

The linear stochastic approximation (LSA) recursion with a decaying stepsize η_n for iteration n is

$$\vartheta_{n+1} = \vartheta_n - \eta_n\left(\widehat{A}_n \vartheta_n - \widehat{b}_n\right) = \left(I - \eta_n \widehat{A}_n\right)\vartheta_n + \eta_n \widehat{b}_n = \left(I - \eta_n A_n - \eta_n \widetilde{A}_n\right)\vartheta_n + \eta_n \widehat{b}_n, \tag{220}$$

where we denote $A_n := \mathbb{E}[\widehat{A}_n]$ and $\widetilde{A}_n := \widehat{A}_n - A_n$.

To validate stability, we compare the LSA iterates generated from two adjacent data streams. Consider two sequences $\{(\widehat{A}_i, \widehat{b}_i)\}_{i=1}^n$ and $\{(\widehat{A}_i', \widehat{b}_i')\}_{i=1}^n$ that differ in exactly one index l, i.e., $(\widehat{A}_i, \widehat{b}_i) = (\widehat{A}_i', \widehat{b}_i')$ for all $i \ne l$, but $(\widehat{A}_l, \widehat{b}_l) \ne (\widehat{A}_l', \widehat{b}_l')$. We impose the following assumptions.

Assumption D.2. *Suppose that there exist constants $L, \mu > 0$ and $\widehat{\sigma}, \sigma \geq 1$ satisfying the following properties:*

1. *$A_i \succeq \mu I$ for all $i \geq 1$;*
2. *$\|\widehat{A}_i\| \leq \widehat{\sigma}$, $\|\widetilde{A}_i\| \leq \sigma$, $\|\widehat{b}_i\|_2 \leq L$ for all $i \geq 1$.*

With Assumption D.2 in place, we establish stability of the terminal LSA iterate. In particular, changing a single observation in the data stream alters ϑ_{n+1} by at most $O((\log^3 n)/n)$ with high probability.

Proposition D.2 (Bounded differences for LSA)**.** *Consider any fixed $\zeta \geq 1$, and suppose that Assumption D.2 holds. Assume that the LSA recursion (220) adopts the stepsize $\eta_n = \min\{1/\widehat{\sigma}, \gamma_n/n\}$ at iteration n, where $\gamma_n = C \log n$ for some constant $C \geq \frac{2(\zeta+1)}{\mu} > 0$. Then there exists a constant $K > 0$ (independent of n) such that, for any $n \geq d^{\frac{1}{\zeta}}$ and any two adjacent datasets differing in a single time index l, the corresponding LSA iterates $\{\vartheta_n\}_{n=1}^{\infty}$ and $\{\vartheta'_n\}_{n=1}^{\infty}$ satisfy*

$$\left\|\vartheta_{n+1} - \vartheta'_{n+1}\right\|_2 \leq K \frac{\log^3 n}{n}$$

with probability at least $1 - n^{-\zeta}$.

The proof of Proposition D.2 is postponed to Section D.4.2. We now return to verify Assumption 4.3. In this example, our prediction model takes the form $\widehat{\mu}(x \mid Z_{1:n}) = x^\top \vartheta_n$. Assume that the covariate X is essentially bounded, i.e., $\|X\| \leq B_x$ almost surely. Then, for any two neighboring samples $Z_{1:n}$ and $Z'_{1:n}$ in a typical set $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1 - n^{-\zeta}$, the corresponding parameter estimates—denoted by ϑ_n and ϑ'_n, respectively—obtained by LSA satisfy

$$\left| x^\top \vartheta_n - x^\top \vartheta'_n \right| \leq \|x\|_2 \, \|\vartheta_n - \vartheta'_n\|_2 \leq B_x K \frac{\log^3 n}{n}.$$

Hence, this justifies Assumption 4.3 with $L_2 = B_x K \log^3(m+1)$ (with m the size of the training set used in this assumption).

D.3 Stochastic strongly convex optimization

We now turn to another case where the predictive model $\widehat{\mu}(\cdot \mid \{Z_i\}_{i=1}^n)$ is trained in an adaptive manner. For parametric statistical models, a natural approach is to maintain a parameter vector and update it using an incremental optimization rule. Suppose that the model used at iteration τ is $\widehat{\mu}_\tau(\cdot) = \widehat{\mu}(\cdot \mid \vartheta_\tau)$, where ϑ_τ is learned from the data $\{(X_i, Y_i)\}_{i=1}^{\tau-1}$ via stochastic optimization (or, more generally, an iterative online training procedure). Starting from $\vartheta_{\tau-1}$, after observing the new data point $(X_{\tau-1}, Y_{\tau-1})$ we update

$$\vartheta_\tau = \vartheta_{\tau-1} - \eta_{\tau-1} f\big(\vartheta_{\tau-1}; (X_{\tau-1}, Y_{\tau-1})\big), \tag{221}$$

where $f(\vartheta; (X, Y))$ denotes the update direction and $\eta_{\tau-1}$ the stepsize. Let Θ denote the parameter domain and $\mathcal{Z}$ the domain of data points.

To study stability for such adaptive methods, we impose several standard conditions on the update map f.

Assumption D.3. *There exist constants $0 < \mu \leq L$ and $B > 0$ such that, for any $\vartheta, \vartheta' \in \Theta$ and $(x, y) \in \mathcal{Z}$,*

- *Strong convexity: $\langle f(\vartheta; (x,y)) - f(\vartheta'; (x,y)), \vartheta - \vartheta' \rangle \geq \mu \|\vartheta - \vartheta'\|_2^2$;*
- *Smoothness: $\|f(\vartheta; (x,y)) - f(\vartheta'; (x,y))\|_2 \leq L \|\vartheta - \vartheta'\|_2$;*
- *Boundedness: $\|f(\vartheta; (x,y))\|_2 \leq B$.*

With Assumption D.3 in place and suitably decaying stepsizes, a single-sample perturbation has an $O(1/n)$ effect on the parameter iterate, as asserted in the following lemma. The proof is provided in Section D.4.3.

Proposition D.3. *Suppose that Assumption D.3 holds. Consider the parameter sequence $\{\vartheta_n\}_{n=1}^{\infty}$ updated according to (221) with stepsize $\eta_n = \min\{\gamma/n,\, 1/L\}$, where $\gamma > 3/\mu$. Then there exists a constant $K > 0$ (independent of n) such that, for any two adjacent datasets differing in a single time index l, the corresponding iterates $\{\vartheta_n\}_{n=1}^{\infty}$ and $\{\vartheta'_n\}_{n=1}^{\infty}$ satisfy*

$$\left\|\vartheta_n - \vartheta'_n\right\|_2 \le \frac{K}{n}.$$

We now discuss how to incorporate the above adaptive updates into the online full conformal construction. In particular, when forming the augmented dataset $(X_{n,r,l}, y)$ in the proposed full conformal procedure, we update the model parameter using the same one-step rule. For any pair $(x, y) \in \mathcal{X} \times \mathbb{R}$, define

$$\vartheta_\tau^{(x,y)} := \vartheta_{\tau-1} - \eta_{\tau-1} f\big(\vartheta_{\tau-1}; (x, y)\big).$$

Accordingly, in round r of stage n, we redefine the residual score $s_i(X_{n,r,l}, y)$ from Eqn. (26) as follows:

$$s_i^{(X,y)} := \left|Y_{n,r-1,i} - \widehat{\mu}\big(X_{n,r-1,i}^{\mathsf{cal}} \,|\, \vartheta_{\tau_{n,r}}^{(X,y)}\big)\right|, \qquad i = 1, \ldots, T_{r-1}, \tag{222a}$$

$$s_{\mathsf{test}}^{(X,y)} := \left|y - \widehat{\mu}\big(X \,|\, \vartheta_{\tau_{n,r}}^{(X,y)}\big)\right|. \tag{222b}$$

Equipped with (222), we can then construct the prediction set according to (27). This leads to a modification of Algorithm 4 that incorporates adaptive updates of the fitted model.

D.4 Detailed proofs

D.4.1 Proof of Propostion D.1

Since $\mathcal{C}$ is convex and $\ell(\cdot; z)$ is differentiable in the first argument, the standard optimality condition (Boyd and Vandenberghe, 2004) yields

$$\big\langle \nabla \widehat{L}_n(\widehat{\vartheta}_n), \vartheta - \widehat{\vartheta}_n \big\rangle \ge 0 \quad \text{for all } \vartheta \in \mathcal{C}, \tag{223}$$

$$\big\langle \nabla \widehat{L}'_n(\widehat{\vartheta}'_n), \vartheta - \widehat{\vartheta}'_n \big\rangle \ge 0 \quad \text{for all } \vartheta \in \mathcal{C}. \tag{224}$$

In particular, taking $\vartheta = \widehat{\vartheta}'_n$ in (223) and $\vartheta = \widehat{\vartheta}_n$ in (224), adding these two inequalities, and setting $\Delta_n := \widehat{\vartheta}'_n - \widehat{\vartheta}_n$, we obtain

$$\big\langle \nabla \widehat{L}_n(\widehat{\vartheta}_n) - \nabla \widehat{L}'_n(\widehat{\vartheta}'_n), \Delta_n \big\rangle \ge 0. \tag{225}$$

By construction, the gradients of the two empirical loss functions satisfy

$$\nabla \widehat{L}'_n(\vartheta) = \nabla \widehat{L}_n(\vartheta) + \frac{1}{n}\big(\nabla \ell(\vartheta; Z'_n) - \nabla \ell(\vartheta; Z_n)\big),$$

which, when evaluated at $\vartheta = \widehat{\vartheta}'_n$, yields

$$\nabla \widehat{L}'_n(\widehat{\vartheta}'_n) = \nabla \widehat{L}_n(\widehat{\vartheta}'_n) + \frac{1}{n}\big(\nabla \ell(\widehat{\vartheta}'_n, Z'_n) - \nabla \ell(\widehat{\vartheta}'_n, Z_n)\big).$$

Substituting this identity into (225) yields

$$\big\langle \nabla \widehat{L}_n(\widehat{\vartheta}_n) - \nabla \widehat{L}_n(\widehat{\vartheta}'_n), \Delta_n \big\rangle - \frac{1}{n}\big\langle \nabla \ell(\widehat{\vartheta}'_n; Z'_n) - \nabla \ell(\widehat{\vartheta}'_n; Z_n), \Delta_n \big\rangle \ge 0.$$

Rearranging terms, we are left with

$$\big\langle \nabla \widehat{L}_n(\widehat{\vartheta}'_n) - \nabla \widehat{L}_n(\widehat{\vartheta}_n), \Delta_n \big\rangle \le \frac{1}{n}\big\langle \nabla \ell(\widehat{\vartheta}'_n; Z_n) - \nabla \ell(\widehat{\vartheta}'_n; Z'_n), \Delta_n \big\rangle. \tag{226}$$

According to Assumption D.1 there exists a quantity $\beta_L > 0$ such that

$$\left\|\nabla \ell(\vartheta; z) - \nabla \ell(\vartheta; z')\right\|_2 \le 2\beta_L \qquad \text{for all } \vartheta \in \mathcal{C},\ z, z'. \tag{227}$$

With (227) in mind, we can bound the right-hand side of (226) using the Cauchy–Schwarz inequality:

$$\begin{aligned}\frac{1}{n}\left|\left\langle \nabla \ell(\widehat{\vartheta}_n'; Z_n') - \nabla \ell(\widehat{\vartheta}_n'; Z_n), \Delta_n \right\rangle\right| &\le \frac{1}{n}\left\|\nabla \ell(\widehat{\vartheta}_n'; Z_n') - \nabla \ell(\widehat{\vartheta}_n'; Z_n)\right\|_2 \|\Delta_n\|_2 \\ &\le \frac{2\beta_L}{n}\|\Delta_n\|_2.\end{aligned}$$

Combine this with (226) to arrive at the upper bound

$$\left\langle \nabla \widehat{L}_n(\widehat{\vartheta}_n') - \nabla \widehat{L}_n(\widehat{\vartheta}_n), \Delta_n \right\rangle \le \frac{2\beta_L}{n}\|\Delta_n\|_2. \tag{228}$$

On the other hand, the left-hand side of (228) can be expressed in terms of the empirical Hessian. Specifically, by the fundamental theorem of calculus for vector-valued functions,

$$\nabla \widehat{L}_n(\widehat{\vartheta}_n') - \nabla \widehat{L}_n(\widehat{\vartheta}_n) = \int_0^1 \nabla^2 \widehat{L}_n\big(\widehat{\vartheta}_n + t\Delta_n\big)\,\Delta_n \,\mathrm{d}t,$$

and as a consequence,

$$\left\langle \nabla \widehat{L}_n(\widehat{\vartheta}_n') - \nabla \widehat{L}_n(\widehat{\vartheta}_n), \Delta_n \right\rangle = \int_0^1 \Delta_n^\top \nabla^2 \widehat{L}_n\big(\widehat{\vartheta}_n + t\Delta_n\big)\Delta_n \,\mathrm{d}t.$$

In particular, there exists some $\widetilde{\vartheta}$ lying within the line segment between $\widehat{\vartheta}_n$ and $\widehat{\vartheta}_n'$ such that

$$\left\langle \nabla \widehat{L}_n(\widehat{\vartheta}_n') - \nabla \widehat{L}_n(\widehat{\vartheta}_n), \Delta_n \right\rangle = \Delta_n^\top \nabla^2 \widehat{L}_n(\widetilde{\vartheta})\Delta_n.$$

We now control the empirical Hessian $\nabla^2 \widehat{L}_n(\vartheta)$ uniformly over all $\vartheta \in \mathcal{C}$. Denote $\mathcal{N}\big(\frac{\mu}{4\beta_s}, \mathcal{C}\big)$ as the $\frac{\mu}{4}$-cover of $\mathcal{C}$. Then we have $|\mathcal{N}\big(\frac{\mu}{4\beta_s}, \mathcal{C}\big)| \le \left(\frac{12\beta_s D_{\mathcal{C}}}{\mu}\right)^d$. According to Bernstein's inequality, we have

$$\sup_{\vartheta \in \mathcal{N}\left(\frac{\mu}{4\beta_s}, \mathcal{C}\right)} \left\{\left\|\nabla^2 \widehat{L}_n(\vartheta) - \nabla^2 L(\vartheta)\right\|_{\infty,\infty}\right\} \le \frac{\mu}{4} \tag{229}$$

with probability at least $1 - d^2 \exp\big\{-\frac{\mu^2 n}{32\beta_s^2 D_{\mathcal{C}}^2}\big\}$, provided that $n \ge \frac{32 d\beta_s^2 D_{\mathcal{C}}^2}{\mu^2}\log\big(\frac{12\beta_s D_{\mathcal{C}}}{\mu}\big)$. Combining this with Assumption D.1 leads to

$$\begin{aligned}\sup_{\vartheta \in \mathcal{C}}\left\{\left\|\nabla^2 \widehat{L}_n(\vartheta) - \nabla^2 L(\vartheta)\right\|_{\infty,\infty}\right\} &\le \beta_s \sup_{\vartheta \in \mathcal{C}} \inf_{\vartheta^* \in \mathcal{N}\left(\frac{\mu}{4\beta_s}, \mathcal{C}\right)} \left\{\|\vartheta - \vartheta^*\|_2\right\} \\ &\quad + \sup_{\vartheta^* \in \mathcal{N}\left(\frac{\mu}{4\beta_s}, \mathcal{C}\right)} \left\{\left\|\nabla^2 \widehat{L}_n(\vartheta) - \nabla^2 L(\vartheta)\right\|_{\infty,\infty}\right\} \overset{(229)}{\le} \beta_s \frac{\mu}{4\beta_s} + \frac{\mu}{4} \le \frac{\mu}{2}.\end{aligned}$$

This combined with the strong convexity assumption and (229) tells us that: for all $\vartheta \in \mathcal{C}$ and every unit vector $u \in \mathbb{R}^d$,

$$u^\top \nabla^2 \widehat{L}_n(\vartheta) u \ge u^\top \nabla^2 L(\vartheta) u - \left\|\nabla^2 \widehat{L}_n(\vartheta) - \nabla^2 L(\vartheta)\right\| \ge \mu - \frac{\mu}{2} = \frac{\mu}{2}.$$

Putting all these pieces together, we arrive at

$$\frac{\mu}{2}\|\Delta_n\|_2^2 \le \Delta_n^\top \nabla^2 \widehat{L}_n(\widetilde{\vartheta})\Delta_n = \left\langle \nabla \widehat{L}_n(\widehat{\vartheta}_n') - \nabla \widehat{L}_n(\widehat{\vartheta}_n), \Delta_n \right\rangle \le \frac{2\beta_L}{n}\|\Delta_n\|_2.$$

If $\Delta_n = 0$, the bound is trivial. Otherwise, cancelling $\|\Delta_n\|_2$ from both sides yields $\|\Delta_n\| \le \frac{4\beta_L}{\mu n}$.

D.4.2 Proof of Proposition D.2

Without loss of generality, we assume that $\vartheta_1 = 0$. Note that under the stepsize choice $\eta_n = \min\{1/\widehat{\sigma}, \gamma_n/n\}$, it can be easily seen that: for any unit vector u,

$$1 = \|u\|_2^2 \geq u^\top (I - \eta_n \widehat{A}_n) u \geq \|u\|_2^2 - \eta_n \widehat{\sigma} = 1 - \eta_n \widehat{\sigma} \geq 0.$$

This implies that for any n, $\|I - \eta_n \widehat{A}_n\| \leq 1$. According to the LSA update rule (220), we have

$$\begin{aligned} \|\vartheta_n\|_2 &\leq \left\|\left(I - \eta_{n-1}\widehat{A}_{n-1}\right)\vartheta_{n-1}\right\|_2 + \eta_{n-1}\left\|\widehat{b}_n\right\|_2 \\ &\overset{\text{(a)}}{\leq} \|\vartheta_{n-1}\|_2 + \frac{CL\log(n-1)}{n-1} \leq \ldots \leq \sum_{i=1}^{n} \frac{CL\log i}{i} \leq CL\log^2 n, \end{aligned} \tag{230}$$

where we have made use of Assumption D.2 and the choice that $\eta_n \leq \gamma_n/n$.

Stability at iteration l. By construction, $\vartheta_i = \vartheta_i'$ holds for all $i = 1, \ldots, l$. At the perturbed iteration $l+1$, it holds that

$$\vartheta_{l+1} - \vartheta_{l+1}' = \left[\vartheta_l - \eta_l\left(\widehat{A}_l\vartheta_l - \widehat{b}_l\right)\right] - \left[\vartheta_l' - \eta_l\left(\widehat{A}_l'\vartheta_l' - \widehat{b}_l'\right)\right] = -\eta_l\left[(\widehat{A}_l - \widehat{A}_l')\vartheta_l - (\widehat{b}_l - \widehat{b}_l')\right].$$

Taking this together with $\eta_l \leq \gamma_l/l$, $\sigma \geq 1$, and Assumption D.2 leads to

$$\begin{aligned} \left\|\vartheta_{l+1} - \vartheta_{l+1}'\right\|_2 &\leq \frac{\gamma_l}{l}\left(\left\|\widetilde{A}_l\right\| + \left\|\widetilde{A}_l'\right\|\right)\|\vartheta_l\|_2 + \frac{2\gamma_l L}{l} \\ &\overset{(230)}{\leq} \frac{2\gamma_l}{l}\left(\sigma CL\log^2 l + L\right) \leq \frac{4C\sigma L\log^3 l}{l}. \end{aligned} \tag{231}$$

Error propagation after iteration l. Let us define, for all $i \geq 1$, $\Delta_i \coloneqq \vartheta_i - \vartheta_i'$. For every $i > l$, the two recursions share the same $(\widehat{A}_i, \widehat{b}_i)$, so subtracting their updates yields

$$\Delta_{i+1} = \left(I - \eta_i \widehat{A}_i\right)\Delta_i.$$

Iterating over $i = l, \ldots, n$ gives

$$\Delta_{n+1} = \Gamma_{l,n}^{\gamma}\Delta_{l+1}, \qquad \Gamma_{l,n}^{\gamma} \coloneqq \prod_{i=l+1}^{n}\left(I - \eta_i\widehat{A}_i\right), \tag{232}$$

where the matrix product is ordered from $i = l+1$ (the rightmost) up to n (the leftmost). Combining (231) and (232) yields

$$\|\Delta_{n+1}\|_2 = \|\Gamma_{l,n}^{\gamma}\Delta_{l+1}\|_2 \leq \|\Gamma_{l,n}^{\gamma}\|\,\|\Delta_{l+1}\|_2 \leq \frac{4C\sigma L\log^3 l}{l}\|\Gamma_{l,n}^{\gamma}\|. \tag{233}$$

Hence, it boils down to controlling the operator norm of the random matrix product $\Gamma_{l,n}^{\gamma}$.

Tools for analyzing products of random matrices. To bound $\|\Gamma_{l,n}^{\gamma}\|$, we follow the framework of Durmus et al. (2021); Huang et al. (2022). For any matrix $B \in \mathbb{R}^{d\times d}$, let $(\sigma_\ell(B))_{\ell=1}^d$ denote its singular values, and for $p \geq 1$ define the Schatten p-norm $\|B\|_p \coloneqq \left(\sum_{\ell=1}^d \sigma_\ell^p(B)\right)^{1/p}$. For $p, q \geq 1$ and a random matrix X, define $\|X\|_{p,q} \coloneqq \left(\mathbb{E}[\|X\|_p^q]\right)^{1/q}$. We record the following two useful lemmas from Durmus et al. (2021).

Lemma D.1 (Proposition 2 in Durmus et al. (2021))**.** *Let $\{Y_\ell : \ell \in \mathbb{N}\}$ be an independent sequence. Assume that for each $\ell \in \mathbb{N}$, there exist $m_\ell \in (0,1)$ and $\sigma_\ell > 0$ such that*

$$\|\mathbb{E}[Y_\ell]\|^2 \leq 1 - m_\ell \quad \text{and} \quad \|Y_\ell - \mathbb{E}[Y_\ell]\| \leq \sigma_\ell \quad \text{almost surely.}$$

Define $Z_n = \prod_{\ell=0}^n Y_\ell = Y_n Z_{n-1}$ for $n \geq 1$, with an arbitrary starting point Z_0. Then, for any $2 \leq q \leq p$ and $n \geq 1$, one has

$$\|Z_n\|_{p,q}^2 \leq \prod_{\ell=1}^{n}\left(1 - m_\ell + (p-1)\sigma_\ell^2\right)\left\|Z_0\right\|_{p,q}^2.$$

Lemma D.2 (Lemma 1 in Durmus et al. (2021)). *Let $A \in \mathbb{R}$, $B > 0$, $C \geq 1$, and $p_0, p_1 \in \mathbb{R}$ with $1 \leq p_0 \leq p_1 < \infty$. Let X be a real random variable satisfying, for any $p \in [p_0, p_1]$,*

$$\mathbb{E}[|X|^p] \leq C \exp(-Ap + Bp^2). \tag{234}$$

Then, for all $\delta \in (0, 1]$, with probability at least $1 - \delta$ one has

$$|X| \leq \exp\Big(-A + Bp_0 + 2\sqrt{B \log(C/\delta)} + \frac{\log(C/\delta)}{p_1}\Big).$$

High-probability bound on $\|\Gamma^{\gamma}_{l,n}\|$. We intend to apply Lemma D.1 to analyze the matrix product

$$\Gamma^{\gamma}_{j,n} \coloneqq \prod_{i=j}^{n} (I - \eta_i A_i + \eta_i \widetilde{A}_i), \qquad j \leq n.$$

Take $Y_i = I - \eta_i A_i + \eta_i \widetilde{A}_i$, and $Z_0 = I_d$. Then $\mathbb{E}[Y_i] = I - \eta_i A_i$ and $Y_i - \mathbb{E}[Y_i] = \eta_i \widetilde{A}_i$. With Assumption D.2 in place, by setting

$$m_i = \mu \eta_i, \qquad \sigma_i = \sigma \eta_i,$$

we see that $\|\mathbb{E}[Y_i]\|^2 \leq 1 - m_i$ and $\|Y_i - \mathbb{E}[Y_i]\| \leq \sigma_i$. Lemma D.1 then tells us that, for any $2 \leq q \leq p$,

$$\begin{aligned} \mathbb{E}\big[\|\Gamma^{\gamma}_{j,n}\|_p^q\big]^{1/q} = \|\Gamma^{\gamma}_{j,n}\|_{p,q} &\leq \prod_{i=j}^{n} \Big(1 - \mu\eta_i + (p-1)\sigma^2\eta_i^2\Big) \|I_d\|_p \\ &\leq d^{1/p} \exp\left\{-\mu \sum_{i=j}^{n} \eta_i + (p-1)\sigma^2 \sum_{i=j}^{n} \eta_i^2\right\}, \end{aligned} \tag{235}$$

where the last step makes use of the elementary inequality $\log(1 + x) \leq x$.

Since $\|I - \eta_i \widehat{A}_i\| \leq 1$ for each i, we have $\|\Gamma^{\gamma}_{l,n}\| \leq 1$ for all $l \leq n$. By introducing an index

$$j \coloneqq \max\Big\{l, \lceil 12C\sigma^2/\mu \rceil, \lceil 2C\widehat{\sigma} \log(C\widehat{\sigma}) \rceil\Big\},$$

we can easily check that $\eta_i = \gamma_i / i$ for every $i \geq j$. Further, we can derive the following inequality:

$$\begin{aligned} \sigma^2 \sum_{i=j}^{n} \frac{\gamma_i^2}{i^2} &\leq C^2\sigma^2 \log^2 n \left(\frac{1}{j-1} - \frac{1}{n}\right) \leq \frac{C\mu}{12} \log^2 n \\ &\overset{(a)}{\leq} \frac{\mu}{6} \int_j^n \frac{C \log x}{x} \mathrm{d}x \leq \frac{\mu}{6} \sum_{i=j}^{n} \frac{C \log i}{i} = \frac{\mu}{6} \sum_{i=j}^{n} \frac{\gamma_i}{i}. \end{aligned} \tag{236}$$

Here, (a) holds whenever $n \geq 3j$. In the regime $n \leq j$, the desired claim follows directly by combining (233) with the bound $\left\|\Gamma^{\gamma}_{l,n}\right\| \leq 1$. Taking $p = q$ in (235), we obtain

$$\mathbb{E}\big[\|\Gamma^{\gamma}_{j,n}\|_p^p\big] \leq d \exp\left\{-p\mu \sum_{i=j}^{n} \frac{\gamma_i}{i} + p^2\sigma^2 \sum_{i=j}^{n} \frac{\gamma_i^2}{i^2}\right\}. \tag{237}$$

Applying Lemma D.2 with $p_0 = 2$, $p_1 = \infty$, $C = d$, $A = \mu \sum_{i=j}^{n} \frac{\gamma_i}{i}$, $B = \sigma^2 \sum_{i=j}^{n} \frac{\gamma_i^2}{i^2}$, and using the fact that (237) implies (234) for all $p \geq 2$, we can deduce that, with probability at least $1 - \delta/n$,

$$\|\Gamma^{\gamma}_{j,n}\| \leq \exp\left\{-\mu \sum_{i=j}^{n} \frac{\gamma_i}{i} + 3\sigma^2 \sum_{i=j}^{n} \frac{\gamma_i^2}{i^2} + \log \frac{dn}{\delta}\right\} \overset{(236)}{\leq} \exp\left\{-\frac{\mu}{2} \sum_{i=j}^{n} \frac{\gamma_i}{i} + \log \frac{dn}{\delta}\right\}.$$

Using the elementary fact $\sum_{i=j}^{n} \frac{\log i}{i} \geq \log^2 n - \log^2 j = \log(n/j)\log(nj) \geq \log(n/j)\log n$, we further obtain

$$\|\Gamma_{j,n}^{\gamma}\| \leq \frac{dn}{\delta}\Big(\frac{j}{n}\Big)^{\frac{C\mu\log n}{2}}. \tag{238}$$

Since $\|\Gamma_{l,n}^{\gamma}\| \leq \|\Gamma_{j,n}^{\gamma}\|$ holds by construction and $\|\Gamma_{l,n}^{\gamma}\| \leq 1$, combining (238) with the trivial upper bound 1 yields, for all $l \leq n$ and with probability at least $1-\delta$,

$$\|\Gamma_{l,n}^{\gamma}\| \leq \min\left\{1,\ \frac{dn}{\delta}\left(\frac{\max\{l, \lceil 12C\sigma^2/\mu\rceil, \lceil 2C\widehat{\sigma}\log(C\widehat{\sigma})\rceil\}}{n}\right)^{\frac{C\mu\log n}{2}}\right\}, \quad l = 1,\dots,n. \tag{239}$$

Bounded-difference property of ϑ_{n+1}. For ease of exposition, we introduce the following notation:

$$C_0 \coloneqq \max\left\{\lceil 12C\sigma^2/\mu\rceil, \lceil 2C\widehat{\sigma}\log(C\widehat{\sigma})\rceil\right\}.$$

Substituting (239) into (233), we arrive at

$$\|\Delta_{n+1}\|_2 \leq \frac{4C\sigma L\log^3 l}{l}\min\left\{1,\ \frac{dn}{\delta}\Big(\frac{\max\{l, C_0\}}{n}\Big)^{\frac{C\mu\log n}{2}}\right\} \tag{240}$$

holds for any $l = 1,\dots,n$. Let $\kappa \coloneqq \sigma/\mu$ and choose $\delta = n^{-\zeta}$, $C = \frac{2(\zeta+1)}{\mu}$. For $l \leq C_0$, use (240) to obtain

$$\begin{aligned}\|\Delta_{n+1}\|_2 &\leq 4C\sigma L\log^3 C_0 \min\left\{1, dn^{\zeta+1}\left(\frac{C_0}{n}\right)^{(\zeta+1)\log n}\right\}\\ &\overset{\text{(a)}}{\leq} 4C\sigma L\log^3 C_0\left(\frac{dC_0^{(\zeta+1)\log n}}{n^{(\zeta+1)(\log n-1)}}\right)^{\frac{1}{(\zeta+1)(\log n-1)}} \overset{\text{(b)}}{\leq} 4C\sigma L\log^3 C_0\frac{eC_0^2}{n}.\end{aligned}$$

Here, (a) results from the fact that $a \leq a^{\lambda}$ for any $a, \lambda \in [0,1]$, and (b) holds since $d^{\frac{1}{(\zeta+1)(\log n-1)}} \leq e$ provided that $n^{\zeta} \geq d$. Thus, in this setting, Proposition D.2 holds by taking $K = 8e(\zeta+1)\kappa LC_0^2\log^3 C_0$.

For $l > C_0$, similarly we can derive

$$\begin{aligned}\|\Delta_{n+1}\|_2 &\leq \frac{4C\sigma L\log^3 n}{l}\min\left\{1, dn^{\zeta+1}\left(\frac{l}{n}\right)^{(\zeta+1)\log n}\right\}\\ &\leq \frac{4C\sigma L\log^3 n}{l}\left(\frac{dl^{(\zeta+1)\log n}}{n^{(\zeta+1)(\log n-1)}}\right)^{\frac{1}{(\zeta+1)(\log n-1)}}\\ &\leq \frac{4eC\sigma L\log^3 n}{n}\cdot l^{\frac{(\zeta+1)\log n}{(\zeta+1)(\log n-1)}-1} \leq \frac{4eC\sigma L\log^3 n}{n}\cdot n^{\frac{1}{\log n-1}} \leq \frac{16eC\sigma L\log^3 n}{n}.\end{aligned}$$

In this setting, Proposition D.2 follows by taking $K = 16eC\sigma L$.

D.4.3 Proof of Proposition D.3

We let $K \coloneqq \frac{2BL\gamma}{\mu}$ and prove this result by induction. Fix two adjacent data streams, and let $\{\vartheta_k\}_{k\geq 1}$ and $\{\vartheta_k'\}_{k\geq 1}$ denote the iterates generated by (221) from the same initialization. Assume for the moment that

$$\|\vartheta_n - \vartheta_n'\|_2 \leq \min\left\{\frac{K}{n}, \frac{2B}{L}\right\} \tag{241}$$

holds for some $n \geq 1$, and we would like to bound $\|\vartheta_{n+1} - \vartheta_{n+1}'\|_2$.

We divide into several cases according to the index at which the two data streams differ. First, consider the case where the two streams differ at the most recent observation, i.e., $(X_n, Y_n) \neq (X_n', Y_n')$ while

$(X_i, Y_i) = (X_i', Y_i')$ for all $i \le n-1$. In this case, the iterates coincide up to time n, hence $\vartheta_n = \vartheta_n'$. In view of the update rule (221),

$$\begin{aligned}\|\vartheta_{n+1} - \vartheta_{n+1}'\|_2 &= \eta_n \Big\| f\big(\vartheta_n; (X_n, Y_n)\big) - f\big(\vartheta_n; (X_n', Y_n')\big) \Big\|_2 \\ &\le \min\left\{\frac{2\gamma B}{n}, \frac{2B}{L}\right\} \le \min\left\{\frac{K}{n+1}, \frac{2B}{L}\right\},\end{aligned}$$

where penultimate inequality arises from Assumption D.3 and our stepsize choice, and the last inequality holds since

$$\frac{2\gamma B}{n} \le \frac{4\gamma B}{n+1} \le \frac{2BL\gamma}{\mu(n+1)} = \frac{K}{n+1}.$$

Next, consider the case where the two data streams differ at some index $l \le n-1$. Then the current update at time n is computed from the same observation $Z_n = (X_n, Y_n)$ in both streams. Since $\eta_n \le 1/L$, Lee and Zhang (2025, Lemma 2) implies that the update map is nonexpansive, which, taken together with the induction hypothesis (241), yields

$$\|\vartheta_{n+1} - \vartheta_{n+1}'\|_2 \le \|\vartheta_n - \vartheta_n'\|_2 \le \min\Big\{\frac{K}{n}, \frac{2B}{L}\Big\}.$$

In addition, a direct expansion of the recursion yields

$$\begin{aligned}\|\vartheta_{n+1} - \vartheta_{n+1}'\|_2^2 &= \big\|\vartheta_n - \vartheta_n' - \eta_n\big(f(\vartheta_n; Z_n) - f(\vartheta_n'; Z_n)\big)\big\|_2^2 \\ &\le \|\vartheta_n - \vartheta_n'\|_2^2 - 2\eta_n \langle f(\vartheta_n; Z_n) - f(\vartheta_n'; Z_n), \vartheta_n - \vartheta_n'\rangle + L^2\eta_n^2 \|\vartheta_n - \vartheta_n'\|_2^2 \\ &\le \big(1 - \mu\eta_n + L^2\eta_n^2\big)\|\vartheta_n - \vartheta_n'\|_2^2 \le \frac{K^2\big(1 - 2\mu\eta_n + L^2\eta_n^2\big)}{n^2},\end{aligned} \tag{242}$$

which follows from Assumption D.3. Further, it can be derived that

$$\big(1 - 2\mu\eta_n + L^2\eta_n^2\big)\frac{(n+1)^2}{n^2} \overset{\text{(a)}}{\le} (1 - \mu\eta_n)(1 + 3/n) \overset{\text{(b)}}{\le} 1,$$

where (a) is valid provided that $n + 1 \ge \frac{L^2\gamma}{\mu}$ and (b) holds as long as $\gamma \ge 3/\mu$. Substitution into (242) yields

$$\big\|\vartheta_{n+1} - \vartheta_{n+1}'\big\|_2^2 \le \frac{K^2}{(n+1)^2}.$$

Moreover, if $n + 1 < \frac{L^2\gamma}{\mu}$, then it still holds that

$$\big\|\vartheta_{n+1} - \vartheta_{n+1}'\big\|_2 \le \frac{2B}{L} \le \frac{2BL^2\gamma}{\mu L(n+1)} = \frac{K}{n+1}.$$

We have thus concluded the proof of Proposition D.3.

E Auxiliary concentration inequalities

This appendix collects several classical concentration inequalities that will be used repeatedly in our analysis. Their proofs can be found in, e.g., Boucheron et al. (2013); Vershynin (2018).

Lemma E.1 (McDiarmid inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables taking values in measurable spaces $\mathcal{X}_1, \dots, \mathcal{X}_n$, and let $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be a measurable function satisfying the bounded differences condition: there exist constants $c_1, \dots, c_n \ge 0$ such that for all $x_1, \dots, x_n, x_i' \in \mathcal{X}_i$,*

$$\big|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)\big| \le c_i.$$

Then, for all $t > 0$,

$$\mathbb{P}\big(\big|f(X_1, \dots, X_n) - \mathbb{E}\big[f(X_1, \dots, X_n)\big]\big| \ge t\big) \;\le\; 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma E.2 (Khintchine inequality for $p = 1$)**.** *Let $\{\varepsilon_i\}_{i=1}^n$ be independent Rademacher random variables, that is, $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$. Then for any real coefficients $a_1, \dots, a_n$, one has*

$$\mathbb{E}\Big[\Big|\sum_{i=1}^n a_i\varepsilon_i\Big|\Big] \geq \frac{1}{\sqrt{2}}\Big(\sum_{i=1}^n a_i^2\Big)^{1/2}.$$

Lemma E.3 (Paley–Zygmund inequality)**.** *Let Z be a nonnegative random variable with $\mathbb{E}[Z^2] < \infty$. Then, for any $\theta \in [0,1]$, the following inequality holds:*

$$\mathbb{P}\big(Z \geq \theta\,\mathbb{E}[Z]\big) \;\geq\; (1-\theta)^2\frac{(\mathbb{E}[Z])^2}{\mathbb{E}[Z^2]}.$$

Lemma E.4 (Generalized DKW inequality)**.** *Let $X_1, \dots, X_n$ be independent random elements taking values in $\mathcal{X}$. Let $F : \mathcal{X} \times \mathbb{R} \to [0,1]$ satisfy that, for every $u \in \mathcal{X}$, the map $x \mapsto F(u,x)$ is nondecreasing and right-continuous. Define*

$$\widehat{F}_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^n F(X_i, x), \qquad \overline{F}(x) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[F(X_i,x)\big].$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{x\in\mathbb{R}}\big|\widehat{F}_n(x) - \overline{F}(x)\big| \leq \frac{4}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Specifically, when $\mathcal{X}_i = \mathbb{R}$, $i = 1, \dots, n$ and $F(X_i, x) = \mathbb{1}\{X_i \leq x\}$ we have:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^n\big(\mathbb{1}\{X_i \leq x\} - \mathbb{P}(X_i \leq x)\big)\right| \leq \frac{4}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}\right) \geq 1-\delta.$$

Proof of Lemma E.4. The proof comprises the following steps.

Step 1: applying McDiarmid around the mean. Let

$$g \coloneqq \sup_{x\in\mathbb{R}}\big|\widehat{F}_n(x) - \overline{F}(x)\big|.$$

Fix any $i \in [n]$. Replace X_i by an arbitrary $\widetilde{X}_i$, and define

$$\widetilde{F}_n(x) \coloneqq \frac{1}{n}\Big(F(\widetilde{X}_i, x) + \sum_{j\neq i} F(X_j, x)\Big), \qquad \widetilde{g} \coloneqq \sup_{x\in\mathbb{R}}\big|\widetilde{F}_n(x) - \overline{F}(x)\big|.$$

Since $0 \leq F \leq 1$, for every $x \in \mathbb{R}$,

$$\big|\widehat{F}_n(x) - \widetilde{F}_n(x)\big| = \frac{1}{n}\big|F(X_i,x) - F(\widetilde{X}_i, x)\big| \leq \frac{1}{n}.$$

Using $||u| - |v|| \leq |u - v|$ and $|\sup_x a(x) - \sup_x b(x)| \leq \sup_x |a(x) - b(x)|$, we obtain

$$|g - \widetilde{g}| \leq \sup_{x\in\mathbb{R}}\big|\widehat{F}_n(x) - \widetilde{F}_n(x)\big| \leq \frac{1}{n},$$

so g satisfies bounded differences with constants $c_i = 1/n$. By McDiarmid's inequality, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$g \leq \mathbb{E}[g] + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{243}$$

Step 2: controlling $\mathbb{E}[g]$ via symmetrization and a reduction to indicators. Introduce a ghost sample $X_1', \dots, X_n'$, independent of $(X_1, \dots, X_n)$ satisfying $X_i' \overset{\mathrm{d}}{=} X_i$, and let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher signs, independent of everything else. Recognizing that $\mathbb{E}[F(X_i, x)] = \mathbb{E}[F(X_i', x)]$, we can express

$$\widehat{F}_n(x) - \overline{F}(x) = \mathbb{E}_{X'}\left[\frac{1}{n}\sum_{i=1}^{n}\big(F(X_i, x) - F(X_i', x)\big) \,\middle|\, X\right].$$

By the Jensen inequality and the convexity of sup, we have

$$\mathbb{E}\Big[\sup_{x\in\mathbb{R}}\big|\widehat{F}_n(x) - \overline{F}(x)\big|\Big] \le \mathbb{E}_{X,X'}\left[\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n}\big(F(X_i, x) - F(X_i', x)\big)\right|\right].$$

Using exchangeability of (X_i, X_i') to insert Rademacher signs and then invoking the triangle inequality gives

$$\begin{aligned}
\mathbb{E}_{X,X'}\left[\sup_x\left|\sum_{i=1}^{n}\big(F(X_i, x) - F(X_i', x)\big)\right|\right] &= \mathbb{E}_{X,X',\epsilon}\left[\sup_x\left|\sum_{i=1}^{n}\epsilon_i\big(F(X_i, x) - F(X_i', x)\big)\right|\right] \\
&\le \mathbb{E}_{X,X',\epsilon}\left[\sup_x\left|\sum_{i=1}^{n}\epsilon_i F(X_i, x)\right|\right] + \mathbb{E}_{X,X',\epsilon}\left[\sup_x\left|\sum_{i=1}^{n}\epsilon_i F(X_i', x)\right|\right] = 2\,\mathbb{E}_{X,\epsilon}\left[\sup_x\left|\sum_{i=1}^{n}\epsilon_i F(X_i, x)\right|\right].
\end{aligned}$$

Taking the above inequalities together yields

$$\mathbb{E}[g] = \mathbb{E}\Big[\sup_{x\in\mathbb{R}}\big|\widehat{F}_n(x) - \overline{F}(x)\big|\Big] \le \frac{2}{n}\mathbb{E}_{X,\epsilon}\Big[\sup_{x\in\mathbb{R}}\Big|\sum_{i=1}^{n}\epsilon_i F(X_i, x)\Big|\Big]. \tag{244}$$

We next bound the Rademacher term. Since $0 \le F(X_i, x) \le 1$, for each i and x,

$$F(X_i, x) = \int_0^1 \mathbb{1}\{F(X_i, x) \ge t\}\,\mathrm{d}t.$$

Therefore, for any fixed $x \in \mathbb{R}$,

$$\Big|\sum_{i=1}^{n}\epsilon_i F(X_i, x)\Big| = \Big|\int_0^1\sum_{i=1}^{n}\epsilon_i\mathbb{1}\{F(X_i, x) \ge t\}\,\mathrm{d}t\Big| \le \int_0^1\Big|\sum_{i=1}^{n}\epsilon_i\mathbb{1}\{F(X_i, x) \ge t\}\Big|\,\mathrm{d}t,$$

and hence

$$\sup_{x\in\mathbb{R}}\Big|\sum_{i=1}^{n}\epsilon_i F(X_i, x)\Big| \le \int_0^1\sup_{x\in\mathbb{R}}\Big|\sum_{i=1}^{n}\epsilon_i\mathbb{1}\{F(X_i, x) \ge t\}\Big|\,\mathrm{d}t. \tag{245}$$

Fix $t \in [0, 1]$. For any $u \in \mathcal{X}$, define the threshold

$$a(u, t) \coloneqq \inf\{x \in \mathbb{R} :\ F(u, x) \ge t\} \in [-\infty, +\infty].$$

Since $x \mapsto F(u, x)$ is nondecreasing and right-continuous, we have

$$\{x \in \mathbb{R} :\ F(u, x) \ge t\} = [a(u, t), \infty),$$

and thus, for every $x \in \mathbb{R}$,

$$\mathbb{1}\{F(X_i, x) \ge t\} = \mathbb{1}\{x \ge a(X_i, t)\}. \tag{246}$$

Consequently,

$$\sup_{x\in\mathbb{R}}\Big|\sum_{i=1}^{n}\epsilon_i\mathbb{1}\{F(X_i, x) \ge t\}\Big| = \sup_{x\in\mathbb{R}}\Big|\sum_{i=1}^{n}\epsilon_i\mathbb{1}\{x \ge a(X_i, t)\}\Big|.$$

Let $a_{(1)}(t) \le \cdots \le a_{(n)}(t)$ be the order statistics of $\{a(X_i,t)\}_{i=1}^n$, and let $\epsilon_{(1)}(t),\dots,\epsilon_{(n)}(t)$ be the corresponding reordered signs. Then the mapping $x \mapsto \sum_{i=1}^n \epsilon_i \mathbb{1}\{x \ge a(X_i,t)\}$ is piecewise constant and, as x increases, it takes values $\sum_{j=1}^k \epsilon_{(j)}(t)$ for some $k \in \{0,1,\dots,n\}$. Hence,

$$\sup_{x\in\mathbb{R}} \Big| \sum_{i=1}^n \epsilon_i \mathbb{1}\{x \ge a(X_i,t)\} \Big| = \max_{0\le k\le n} \Big| \sum_{j=1}^k \epsilon_{(j)}(t) \Big|.$$

Since $(\epsilon_1,\dots,\epsilon_n)$ are i.i.d. and independent of the X_i's, conditionally on $X_{1:n}$ the reordered sequence $(\epsilon_{(1)}(t),\dots,\epsilon_{(n)}(t))$ has the same joint distribution as $(\epsilon_1,\dots,\epsilon_n)$. Therefore,

$$\mathbb{E}_\epsilon\Big[\sup_{x\in\mathbb{R}} \Big| \sum_{i=1}^n \epsilon_i \mathbb{1}\{F(X_i,x) \ge t\} \Big| \;\Big|\; X_{1:n} \Big] = \mathbb{E}_\epsilon\Big[\max_{0\le k\le n} \Big| \sum_{j=1}^k \epsilon_j \Big| \Big].$$

Let $S_k := \sum_{j=1}^k \epsilon_j$, $k = 0,1,\dots,n$. Then $(S_k)_{k=0}^n$ is a martingale, and Doob's L^2 maximal inequality yields

$$\mathbb{E}\Big[\max_{0\le k\le n} |S_k|^2 \Big] \le 4\mathbb{E}\big[|S_n|^2\big] = 4n.$$

By Cauchy–Schwarz,

$$\mathbb{E}\Big[\max_{0\le k\le n} |S_k| \Big] \le 2\sqrt{n}.$$

Combining this with (245) and applying Tonelli's theorem gives

$$\mathbb{E}_{X,\epsilon}\Big[\sup_{x\in\mathbb{R}} \Big| \sum_{i=1}^n \epsilon_i F(X_i,x) \Big| \Big] \le \int_0^1 2\sqrt{n}\,\mathrm{d}t = 2\sqrt{n}.$$

Substituting into (244) yields

$$\mathbb{E}[g] \le \frac{2}{n} \cdot 2\sqrt{n} = \frac{4}{\sqrt{n}}. \tag{247}$$

Step 3: completing the proof. Substituting (247) into (243) completes the proof. □

References

Ai, J. and Ren, Z. (2024). Not all distributional shifts are equal: Fine-grained robust conformal inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 641–665. PMLR.

Amann, N., Leeb, H., and Steinberger, L. (2023). Assumption-lean conditional predictive inference via the jackknife and the jackknife+. *arXiv preprint arXiv:2312.14596*.

Angelopoulos, A., Candes, E., and Tibshirani, R. J. (2023). Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074.

Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024a). Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, pages 1616–1630.

Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024b). Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Angelopoulos, A. N., Jordan, M. I., and Tibshirani, R. J. (2025). Gradient equilibrium in online learning: Theory and applications. *arXiv preprint arXiv:2501.08330*.

Aolaritei, L., Zhu, Q. J., Wang, Z. O., Jordan, M. I., and Marzouk, Y. (2025). Conformal prediction under lévy-prokhorov distribution shifts: Robustness to local and global perturbations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Areces, F., Cheng, C., Duchi, J., and Rohith, K. (2024). Two fundamental limits for uncertainty quantification in predictive inference. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 186–218. PMLR.

Areces, F., Mohri, C., Hashimoto, T., and Duchi, J. (2025). Online conformal prediction via online optimization. In *International Conference on Machine Learning*.

Auer, A., Gauch, M., Klotz, D., and Hochreiter, S. (2023). Conformal prediction for time series with modern hopfield networks. *Advances in Neural Information Processing Systems*, 36:56027–56074.

Bao, Y., Huo, Y., Ren, H., and Zou, C. (2024). Cap: A general algorithm for online selective conformal prediction with fcr control. *arXiv preprint arXiv:2403.07728*.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.

Barber, R. F. and Pananjady, A. (2025). Predictive inference for time series: why is split conformal effective despite temporal dependence? *arXiv preprint arXiv:2510.02471*.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2022). Practical adversarial multivalid conformal prediction. *Advances in neural information processing systems*, 35:29362–29373.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27.

Besbes, O., Gur, Y., and Zeevi, A. (2019). Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. (2023). Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pages 2337–2363. PMLR.

Bian, M. and Barber, R. F. (2023). Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2024). Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119(548):3033–3044.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press. Cambridge University Press, Cambridge.

Chen, B., Ren, Z., and Cheng, L. (2024). Conformalized time series with semantic features. *Advances in Neural Information Processing Systems*, 37:121449–121474.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory*, pages 732–749. PMLR.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2019). Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR.

Cleaveland, M., Lee, I., Pappas, G. J., and Lindemann, L. (2024). Conformal prediction regions for time series using linear complementarity programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 20984–20992.

Collina, N., Lu, J., Noarov, G., and Roth, A. (2026). Optimal lower bounds for online multicalibration. *arXiv preprint arXiv:2601.05245*.

Duchi, J. C. (2025). A few observations on sample-conditional coverage in conformal prediction. *arXiv preprint arXiv:2503.00220*.

Durmus, A., Moulines, E., Naumov, A., Samsonov, S., Scaman, K., and Wai, H.-T. (2021). Tight high probability bounds for linear stochastic approximation with fixed stepsize. *Advances in Neural Information Processing Systems*, 34:30063–30074.

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., and Jordan, M. I. (2022). Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119.

Farinhas, A., Zerva, C., Ulmer, D. T., and Martins, A. (2024). Non-exchangeable conformal risk control. In *Twelfth International Conference on Learning Representations*. PMLR.

Feldman, S., Ringel, L., Bates, S., and Romano, Y. (2022). Achieving risk control in online learning settings. *arXiv preprint arXiv:2205.09095*.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.

Gibbs, I. and Candès, E. J. (2024). Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36.

Gibbs, I. and Candès, E. J. (2025). Characterizing the training-conditional coverage of full conformal inference in high dimensions. *arXiv preprint arXiv:2502.20579*.

Gui, Y., Barber, R. F., and Ma, C. (2024). Distributionally robust risk evaluation with an isotonic constraint. *arXiv preprint arXiv:2407.06867*.

Hajihashemi, E. and Shen, Y. (2024). Multi-model ensemble conformal prediction in dynamic environments. *Advances in Neural Information Processing Systems*, 37:118678–118700.

Han, E., Huang, C., and Wang, K. (2024a). Distribution-free predictive inference under unknown temporal drift. *arXiv preprint arXiv:2406.06516*.

Han, E., Huang, C., and Wang, K. (2024b). Model assessment and selection under temporal distribution shift. *arXiv preprint arXiv:2402.08672*.

Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, D., Niles-Weed, J., Tropp, J. A., and Ward, R. (2022). Matrix concentration for products. *Foundations of Computational Mathematics*, 22(6):1767–1799.

Humbert, P., Gazin, U., Heller, R., and Roquain, E. (2025). Online selective conformal inference: adaptive scores, convergence rate and optimality. *arXiv preprint arXiv:2508.10336*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lee, B. and Matni, N. (2024). Single trajectory conformal prediction. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 3019–3024. IEEE.

Lee, K. and Zhang, Y. (2025). Leave-one-out stable conformal prediction. In *The Thirteenth International Conference on Learning Representations*. PMLR.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., Robins, J., and Wasserman, L. (2011). Efficient nonparametric conformal prediction regions. *arXiv preprint arXiv:1111.1418*.

Liang, R. and Barber, R. F. (2025). Algorithmic stability implies training-conditional coverage for distribution-free prediction methods. *The Annals of Statistics*, 53(4):1457–1482.

Lin, Z., Trivedi, S., and Sun, J. (2022). Conformal prediction intervals with temporal dependence. *arXiv preprint arXiv:2205.12940*.

Liu, T., Dobriban, E., and Orabona, F. (2026). Online conformal prediction via universal portfolio algorithms. *arXiv preprint arXiv:2602.03168*.

Ndiaye, E. (2022). Stable conformal prediction sets. In *International Conference on Machine Learning*, pages 16462–16479. PMLR.

Oliveira, R. I., Orenstein, P., Ramos, T., and Romano, J. V. (2024). Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25(225):1–38.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pages 844–853. PMLR.

Podkopaev, A., Xu, D., and Lee, K.-C. (2024). Adaptive conformal inference by betting. In *International Conference on Machine Learning*, pages 40886–40907. PMLR.

Pournaderi, M. and Xiang, Y. (2024). Training-conditional coverage bounds under covariate shift. *arXiv preprint arXiv:2405.16594*.

Ramalingam, R., Kiyani, S., and Roth, A. (2025). The relationship between no-regret learning and online conformal prediction. In *International Conference on Machine Learning*. PMLR.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Sale, Y. and Ramdas, A. (2025). Online selective conformal prediction: Errors and solutions. *CoRR*, abs/2503.16809.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Si, W., Park, S., Lee, I., Dobriban, E., and Bastani, O. (2024). PAC prediction sets under label shift. In *The Twelfth International Conference on Learning Representations*.

Steinberger, L. and Leeb, H. (2023). Conditional predictive inference for stable algorithms. *The Annals of Statistics*, 51(1):290–311.

Stocker, M., Fontana, M., Taieb, S. B., et al. (2025). A gentle introduction to conformal time series forecasting. *arXiv preprint arXiv:2511.13608*.

Su, X., Zhou, Z., and Luo, R. (2024). Adaptive conformal inference by particle filtering under hidden markov models. *arXiv preprint arXiv:2411.01558*.

Sun, S. H. and Yu, R. (2023). Copula conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28.

Vovk, V., Gammerman, A., and Saunders, C. (1999). Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2009). On-line predictive linear regression. *The Annals of Statistics*, pages 1566–1590.

Weinstein, A. and Ramdas, A. (2020). Online control of the false coverage rate and false sign rate. In *International Conference on Machine Learning*, pages 10193–10202. PMLR.

Xu, C., Jiang, H., and Xie, Y. (2024). Conformal prediction for multi-dimensional time series by ellipsoidal sets. *arXiv preprint arXiv:2403.03850*.

Xu, C. and Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR.

Xu, C. and Xie, Y. (2023a). Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587.

Xu, C. and Xie, Y. (2023b). Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pages 38707–38727. PMLR.

Yang, Z., Candès, E., and Lei, L. (2024). Bellman conformal inference: Calibrating prediction intervals for time series. *arXiv preprint arXiv:2402.05203*.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR.

Zhang, Z., Bombara, D., and Yang, H. (2024a). Discounted adaptive online learning: towards better regularization. In *International Conference on Machine Learning*, pages 58631–58661. PMLR.

Zhang, Z., Lu, Z., and Yang, H. (2024b). The benefit of being bayesian in online conformal prediction. *arXiv preprint arXiv:2410.02561*.

Zhao, P., Zhang, L., Jiang, Y., and Zhou, Z.-H. (2020). A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 746–755. PMLR.

Zhou, X., Chen, B., Gui, Y., and Cheng, L. (2025). Conformal prediction: A data perspective. *ACM Computing Surveys*.