Normalization is Enough: An Improved Algorithm for Online Group-wise Conformal Prediction

Lixing Lyu [1] **Wang Chi Cheung** [1]

Abstract

We study adversarial online conformal prediction with the goal of achieving group-wise coverage guarantees while producing small, informative conformal prediction sets. We propose Normalized-GCACI, a new algorithm that modifies the standard group-conditional approach through a single structural change: the prediction threshold is normalized by the number of active groups rather than summed over them. Despite its simplicity, this normalization stabilizes the threshold scale, directly tightens prediction sets, and crucially enables a new analysis that yields a strictly tighter group-wise coverage bound than prior work and extends the known single-group guarantee to the general multi-group setting. Extensive experiments on a non-stationary linear regression benchmark show that Normalized-GCACI tracks the target coverage more closely than existing baselines while producing substantially smaller prediction sets, and targeted ablations confirm that the gain is specifically attributable to the normalization rather than to parameter regularization or step-size decay. Finally, we instantiate our framework on post-retrieval filtering for retrieval-augmented generation (RAG): by constructing conformal prediction sets over candidate documents, our method substantially reduces the number of retained documents while preserving high recall of the relevant evidence, improving retrieval efficiency without sacrificing the quality of downstream generation.

1. Introduction

Uncertainty quantification is key to trustworthy machine learning, as it equips users with confidence estimates per prediction to guide better decision-making. Conformal prediction has become a principled framework for uncertainty quantification that transforms a model's point predictions into prediction sets, that is, sets of labels that are guaranteed to contain the true label with high probability. Formally, consider the supervised setting with feature space $\mathcal{X}$ and label space $\mathcal{Y}$, where each feature $x \in \mathcal{X}$ is associated with its label $y \in \mathcal{Y}$. Given a black-box machine learning model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ that provides a point prediction $\hat{f}(x)$ on the true label y, a canonical aim for the conformal predictor is to construct a set-value function $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ satisfying the following coverage guarantee:

$$\mathbb{P}(y \in \mathcal{C}(x)) \geq 1 - \delta, \tag{1}$$

where $\delta \in (0, 1)$ is the prescribed miscoverage level.

While the coverage guarantee (1) is a fundamental requirement, it is not sufficient to guarantee practical utility of a conformal predictor. In fact, consider a trivial construction that always outputs the entire label space $\mathcal{Y}$ as the prediction set $\mathcal{C}(x)$. Such a predictor clearly satisfies the coverage condition, often with coverage $1 > 1 - \delta$, yet it provides no meaningful information about the label of a specific test point. A practically useful conformal predictor must therefore produce prediction sets that are as small as possible, while still maintaining coverage at the desired level. In other words, the goal is to achieve tightness: prediction sets that are large enough to meet the coverage target, but no larger than necessary, thereby retaining informativeness and supporting effective downstream decision-making.

Beyond achieving tight marginal coverage, we seek a finer-grained guarantee: coverage that holds conditionally within each group of features. Specifically, each feature $x \in \mathcal{X}$ belongs to one or more groups, where the groups are arbitrary subsets of $\mathcal{X}$, and these groups may overlap. We require that the coverage guarantee holds simultaneously across all groups, with prediction sets as small as possible. The group-conditional coverage ((Bastani et al., 2022; Ramalingam et al., 2025; Gupta et al., 2021)) is substantially more refined than the marginal coverage, which only ensures correctness on average over the entire population. Marginal coverage (1) can mask severe disparities: some subgroups may be under-covered while others are over-covered. Group-

[1]National University of Singapore. Correspondence to: Wang Chi Cheung <isecwc@nus.edu.sg>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

conditional coverage addresses this limitation by ensuring reliable prediction sets for each subgroup, regardless of how the groups are defined.

Conformal prediction has found broad applications across diverse domains. In public health, it has been used to produce reliable predictions of influenza incidence, supporting more informed decision-making (Angelopoulos et al., 2023; Susmann et al., 2023). In finance, it has been applied to financial risk assessment and volatility forecasting (Gibbs & Candès, 2024; Bastani et al., 2022; Areces et al., 2025), and in the energy sector, to electricity price (Zaffran et al., 2022) and demand (Areces et al., 2025) forecasting. More broadly, conformal prediction has been adopted in a range of time series problems. More recently, conformal prediction has been applied to large language models (LLMs), providing a principled mechanism for quantifying uncertainty in generated outputs. Notable applications include hallucination detection and mitigation (Yadkori et al., 2024), response quality evaluation in retrieval-augmented generation (RAG) (Feng et al., 2025), and post-retrieval filtering (Ge et al., 2024). In this paper, we also demonstrate how our framework can be applied to post-retrieval filtering to improve RAG efficiency: by constructing conformal prediction sets over candidate documents, we substantially reduce the number of retrieved passages while maintaining high recall of relevant evidence, thereby improving retrieval efficiency without sacrificing downstream generation quality.

1.1. Main Contributions

Algorithm Design. We propose *Normalized-GCACI*, an online conformal prediction algorithm that improves upon GCACI (Ramalingam et al., 2025) via a single structural change: normalizing the threshold by the number of active groups rather than summing. On the theoretical side, the normalization enables a new analysis based on a uniform boundedness property of the group-level parameters, which in turn yields a strictly tighter group-wise coverage bound than GCACI. In particular, we obtain an $O(1/|\mathcal{T}(i)|)$ convergence rate for each group i, where $|\mathcal{T}(i)|$ is the number of context that belongs to group i. This removes the $\sqrt{T}$ dependence that is present in prior work (Ramalingam et al., 2025) and extends the single-group rate of (Gibbs & Candes, 2021) to the general multi-group regime, where T is the length of decision horizon.

Extensive Numerical Experiments in Synthetic Environments. On a non-stationary linear regression, *Normalized-GCACI* consistently attains desired group-wise coverage while producing small prediction, outperforming the performance of MVP (Bastani et al., 2022) and GCACI (Ramalingam et al., 2025) baselines. Ablations further show that neither parameter-level regularizers (non-negativity clipping, ℓ_2-ball projection, or their combination) nor decaying step-size schedules (Angelopoulos et al., 2024) can match these gains, confirming that the benefit arises specifically from our normalizing technique.

Application in Efficient Document Retrieval. We instantiate our framework on a practical problem of direct relevance to retrieval-augmented generation (RAG): post-retrieval filtering of candidate documents. Formulating this task as online group-wise conformal prediction, we construct, for each incoming query, a conformal subset of its candidate documents that provably retains at least one relevant document with high probability, uniformly across groups. On the HotpotQA (Yang et al., 2018) distractor benchmark with a pretrained cross-encoder and a country-level grouping, our Normalized-GCACI attains group-wise recall close to the desired target while retaining small size of conformal subset, offering a principled way to reduce RAG inference cost without sacrificing evidence coverage.

Finally, we provide concluding remarks and discuss future directions in Appendix A.

1.2. Related Works

Online conformal prediction naturally inherits the sequential "predict-then-reveal" structure of conformal inference, in which the learner first outputs a prediction set and then observes the true outcome. The foundational theory is laid out in the textbook of (Vovk et al., 2005), and we refer the reader to (Shafer & Vovk, 2008; Stephen et al., 2021; Zhou et al., 2025) for comprehensive surveys of conformal prediction. A central algorithmic paradigm in the online setting is Adaptive Conformal Inference (ACI), introduced by (Gibbs & Candes, 2021). ACI maintains a time-varying threshold q_t and updates it based on whether the most recent prediction set covered the true label, effectively performing online gradient descent on a pinball (quantile) loss in one dimension (Ramalingam et al., 2025). This simple update rule endows ACI with distribution-free coverage guarantees that remain valid even under arbitrary distribution shift, making it especially well suited for non-stationary environments. Subsequent works have refined and extended the ACI framework along several directions: (Gibbs & Candès, 2024) proposes DtACI, which aggregates multiple ACI instances with different step sizes via exponential reweighting with forgetting, enabling faster adaptation to local distribution shifts. (Bhatnagar et al., 2023) leverages strongly adaptive online learning to obtain improved regret-based coverage bounds. (Angelopoulos et al., 2024) studies the effect of decaying step sizes. (Ramalingam et al., 2025) establishes a formal connection between no-regret online learning and conformal prediction, proposing the GCACI algorithm that extends ACI to the group-conditional setting. Among these, (Ramalingam et al., 2025) is the most closely related to our work, as it considers the same group-wise

coverage objective. In contrast, our algorithm achieves comparable coverage guarantees while producing substantially smaller prediction sets.

The concept of group-wise (or multivalid) coverage has attracted increasing attention in the conformal prediction literature. In addition to the GCACI algorithm of (Ramalingam et al., 2025), group-conditional and multivalid coverage has been studied by (Bastani et al., 2022), who proposes the MVP algorithm. (Lei et al., 2018) studies the Split Conformal Prediction algorithms. (Gupta et al., 2021) investigates online multivalid learning for means, moments, and prediction intervals. In the numerical study of this work, we compare against two standard baselines, MVP and GCACI, and demonstrate that our algorithm outperforms both in terms of coverage accuracy and prediction set size.

Notation. For a positive integer n, denote $[n] = \{1, \ldots, n\}$. We adopt the $O(\cdot), o(\cdot), \Omega(\cdot)$ notation in (Cormen et al., 2009). We adopt $\tilde{O}(\cdot)$, $\tilde{o}(\cdot)$, $\tilde{\Omega}(\cdot)$ notations, which has the same meaning as $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, respectively, except that they hide the logarithmic terms. We abbreviate independently and identically distributed as "i.i.d.". The relationship $R \sim P$ means that the random variable R follows the probability distribution P.

2. Model

The model can be characterized by the tuple

$$(T, \mathcal{X}, \mathcal{Y}, \{s_t(\cdot,\cdot)\}_{t\in[T]}, \mathcal{G}, \delta),$$

where T denotes the length of decision horizon, $\mathcal{X}$ is the feature space, and $\mathcal{Y}$ is the label space. The collection $\{s_t(\cdot,\cdot)\}_{t\in[T]}$ represents a class of *non-conformity score functions*, where each s_t satisfies $s_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. Without loss of generality, we assume the score functions are bounded in $[0, 1]$, i.e. $s_t(x, y) \in [0, 1]$ for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $t \in [T]$. The set $\mathcal{G} = \{G(i)\}_{i\in[m]}$ is a collection of subsets of $\mathcal{X}$ with $|\mathcal{G}| = m$. For any $x \in \mathcal{X}$, if $x \in G(i)$ for some i, we say that x belongs to group i. The groups are allowed to overlap. Without loss of generality and for technical convenience, we assume there exists a group $G_0 \in \mathcal{G}$ such that $G_0 = \mathcal{X}$; i.e., all contexts $x \in \mathcal{X}$ belong to G_0. Thus, any $x \in \mathcal{X}$ belongs to at least one group. Finally, $\delta \in (0, 1)$ is the miscoverage level, and $1 - \delta$ denotes the target coverage. At each round t, the learning agent interacts with an adversary according to the following protocol:

1. The adversary selects a joint probability distribution $\mathcal{P}_t$ over $\mathcal{X} \times \mathcal{Y}$, where the choice of $\mathcal{P}_t$ may depend on the history up to rounds $1, \ldots, t-1$. A sample $(X_t, Y_t) \sim \mathcal{P}_t$ is then drawn.
2. The learning agent observes X_t.
3. The learning agent receives a non-conformity score function $s_t : \mathcal{X} \times \mathcal{Y} \to [0, 1]$.
4. The learning agent selects the threshold q_t, which induces the prediction set
$$\mathcal{C}_t(q_t) = \{y \in \mathcal{Y} : s_t(X_t, y) \le q_t\}.$$
5. The adversary reveals the true label Y_t to the learning agent.

Ideally, at each round t, the learning agent aims to construct a prediction set $\mathcal{C}_t(q_t)$ that covers the true label Y_t with probability $1 - \delta$. i.e.

$$\mathbb{P}_{Y|X_t}(Y \in \mathcal{C}_t(q_t)) \approx 1 - \delta.$$

By the definition of the prediction set, this is equivalent to choosing q_t such that

$$\mathbb{P}_{Y|X_t}(q_t \ge s_t(X_t, Y)) \approx 1 - \delta.$$

Therefore, we can equivalently formulate the interaction process in the following simplified form. At each round t,

1. The adversary selects a joint probability distribution $\mathcal{D}_t$ over the set $\mathcal{X} \times [0, 1]$, and samples a feature vector $X_t \in \mathcal{X}$ and a conformal score $S_t \in [0, 1]$.
2. The learning agent observes X_t.
3. The learning agent selects a threshold q_t.
4. The adversary reveals the realized conformal score S_t to the learning agent.

We aim to achieve a group-wise multi-valid coverage guarantee. For each group index $i \in [m]$, define the index set $\mathcal{T}(i) = \{t \in [T] : X_t \in G(i)\}$. Note that $\mathcal{T}(i)$ is determined by the online realization of the data sequence. Following the literature ((Bastani et al., 2022; Ramalingam et al., 2025; Gupta et al., 2021)), the goal of a conformal predictor is to ensure the following coverage guarantee: for all $i \in [m]$, we have

$$\frac{1}{|\mathcal{T}(i)|} \sum_{t\in\mathcal{T}(i)} \mathbf{1}(Y_t \in \mathcal{C}_t(q_t)) \ge (1 - \delta). \tag{2}$$

However, the coverage guarantee alone is insufficient for a useful conformal predictor. To see this, observe that by setting $\mathcal{C}_t(q_t) = \mathcal{Y}$ for all inputs, t, and all choices of q_t, one can trivially yield a coverage

$$\frac{1}{|\mathcal{T}(i)|} \sum_{t\in\mathcal{T}(i)} \mathbf{1}(Y_t \in \mathcal{C}_t(q_t)) = 1 > (1 - \delta)$$

for all groups. Such prediction sets carry no informative value. A meaningful conformal predictor should instead produce prediction sets that are as small as possible while maintaining coverage close to the target coverage $1-\delta$. That is, we aim to achieve

$$\left| \frac{1}{|\mathcal{T}(i)|} \sum_{t \in \mathcal{T}(i)} \mathbf{1}(Y_t \in \mathcal{C}_t(q_t)) - (1-\delta) \right| \leq o(1),$$

with $|\mathcal{C}_t(q_t)|$ is minimized. In other words, we seek prediction sets that are tight: large enough to attain the desired coverage, yet small enough to remain informative for downstream decision-making.

3. Algorithm: Normalized-GCACI

We now present our approach, Normalized-GCACI, described in detail in Algorithm 1. The algorithm builds upon the GCACI framework of (Ramalingam et al., 2025), with one crucial modification: rather than setting $q_t = \langle \theta_t, g_t \rangle$, we normalize it by the number of active groups, yielding $q_t = \frac{\langle \theta_t, g_t \rangle}{\|g_t\|_1}$.

The intuition behind this normalization is as follows. In the original GCACI, q_t is computed as the inner product $\langle \theta_t, g_t \rangle$, which sums the parameter values $\theta_t(i)$ over groups to which the current input X_t belongs. Therefore, the magnitude of q_t grows with the number of X_t's active group, i.e. $\|g_t\|_1$. This dependence introduces instability, since inputs belonging to many groups receive disproportionately large q_t, leading to unnecessarily wide prediction sets, while inputs belonging to few groups may receive q_t that are too small. Our normalization removes this dependence by averaging the group-level parameters rather than summing them, ensuring that q_t remains on a consistent scale regardless of how many groups the input belongs to.

This modification offers two key advantages. First, the coverage guarantee is preserved, and in practice becomes more stable. This is because the normalized threshold no longer fluctuates with the varying number of active groups across time steps. The resulting q_t values are smaller in magnitude and more tightly concentrated, which reduces the variance of the coverage process and leads to coverage rates that track the target level $1-\delta$ more closely. Second, the normalization directly improves control over the prediction set width. An inflated q_t admits more labels than necessary, resulting in overly conservative prediction sets. By keeping q_t properly scaled, our approach produces tighter prediction sets without sacrificing coverage, achieving a more favorable trade-off between reliability and informativeness. As we formalize in Theorem 3.1 below, this simple normalization also yields a strictly tighter theoretical bound on the group-wise coverage guarantee compared to the original GCACI.

Algorithm 1 Normalized GCACI

Input: Confidence parameter $\delta \in (0,1)$, learning rate η.
Initialize: $\theta_1 = (\theta_1(i))_{i \in [m]}$, with $\theta_1(i) = 0$ for all $i \in [m]$.
for $t = 1, \ldots, T$ **do**
  Receive X_t and $g_t = (g_t(i))_i$, where $g_t(i) = 1$ if $X_t \in G(i)$ else 0.
  Set
$$q_t = \frac{\langle \theta_t, g_t \rangle}{\|g_t\|_1}$$
  Receive $S_t = s_t(X_t, Y_t) \in [0,1]$.
  Update
$$\theta_{t+1} = \theta_t - \eta \cdot (\mathbf{1}\{q_t \geq s_t(X_t, Y_t)\} - (1-\delta)) \cdot g_t$$
end for

Now, we show the following theoretical coverage guarantee for our Normalized-GCACI Algorithm:

Theorem 3.1. *Algorithm 1 can induce the following: For any $i \in [m]$*

$$\left| \frac{1}{|\mathcal{T}(i)|} \sum_{t \in \mathcal{T}(i)} \mathbf{1}(Y_t \in \mathcal{C}_t(q_t)) - (1-\delta) \right| \leq \frac{1}{\eta \cdot |\mathcal{T}(i)|} \cdot \min \left\{ \sqrt{T\eta m}, 2^{m-1} \cdot (1+\eta) \right\}.$$

The theorem is proved in Appendix B. The bound in Theorem 3.1 consists of two terms in the minimum.

The first term, $\sqrt{T\eta m}$, is standard and similar as the bound in (Ramalingam et al., 2025). It reflects the $O(\sqrt{T}/\mathcal{T}(i))$ convergence rate for each group i. The main drawback here is the dependence $\sqrt{T}$ on the full length of the decision horizon T. For some group with significantly small size, i.e. $T = \Omega(\mathcal{T}(i))$, the performance for such group i can be quite poor.

The second term, $2^{m-1} \cdot (1+\eta)$, is, to the best of our knowledge, novel in the literature. This term removes the dependence on T and yields a strictly faster convergence rate of $O(1/\mathcal{T}(i))$ for each group i. The cost of this improved is an exponential dependence 2^m on the number of groups m. However, in most problem instances of practical interest, the decision horizon T is substantially larger than m, and consequently $\mathcal{T}(i)$ is large relative to 2^m for each group i, making the exponential factor a minor concern in practice. We remark that (Gibbs & Candes, 2021) establishes an $O(1/T)$ convergence rate for the single-group case $m = 1$. Our second term successfully extends it to the general multi-group setting $m > 1$, explicitly characterizing how the

convergence rate relates to and depends on the number of groups. This generalization represents a significant advance in the theory of group-wise online conformal prediction.

Deriving this second term requires new analytical techniques beyond those used in (Ramalingam et al., 2025). The key is the following lemma, which establishes a uniform bound on the absolute value of $\theta_t(i)$ for all $t \in [T]$ and $i \in [m]$. This boundedness property is a direct consequence of our normalization. Indeed, our normalized q_t averages the group-level parameters rather than summing them, the update at each round is inherently controlled, preventing any single coordinate $\theta_t(i)$ from growing unboundedly.

Lemma 3.2. *For Algorithm 1, we can obtain: for any $i \in [m]$, $t \in [T+1]$,*

$$\theta_t(i) \in [-\mathcal{B}_m, 1 + \mathcal{B}_m],$$

$$\mathcal{B}_m = 2^{m-1} \cdot (1+\eta) - 1.$$

The lemma is proved in Appendix B. The proof of Lemma 3.2 proceeds by induction on the number of groups m. The base case $m = 1$ recovers the single-group bound of (Gibbs & Candes, 2021). For the induction step, we analyze the worst-case trajectory of $\theta_t(i)$ by observing that, to drive a single coordinate $\theta_t(i)$ as negative as possible, the remaining coordinates must simultaneously attain their maximal values. This coupling across coordinates leads to a recursive characterization of the bound $\mathcal{B}_m$, satisfying the recurrence $\mathcal{B}_{m+1} = \sum_{k=1}^{m}(\mathcal{B}_k + 1) + \eta$, resulting in $\mathcal{B}_m = 2^{m-1} \cdot (1+\eta) - 1$ as shown in Lemma 3.2.

4. Experiment: Synthetic Non-stationary Linear Regression

We evaluate our algorithm on a synthetic non-stationary linear regression problem. The results demonstrate the advantage of our approach in terms of both coverage and prediction set size. Furthermore, we compare our algorithm with several variants built on the original GCACI framework using alternative regularization tricks and decaying step size, demonstrating the necessity of our normalization technique.

The feature space is $\mathcal{X} = \mathbb{R}^{250}$, and the corresponding label y is generated according to the linear model

$$y = \langle \theta, x \rangle + \epsilon,$$

where $\theta \in \mathbb{R}^{250}$ is an unknown parameter vector and ϵ is a mean-zero random noise. We define the nonconformity score function as $s_t(x, y) = |y - \hat{f}(x)|$, where $\hat{f}$ is a pretrained linear regression model. We compare our algorithm with two existing baselines: Algorithm MVP ((Bastani et al., 2022)) and Algorithm GCACI ((Ramalingam et al., 2025)), setting $\eta = 1$ for both our algorithm and GCACI. We define 100 groups based on the feature space. We evaluate algorithm performance across two metrics: *Coverage*, defined as the indicator $\mathbf{1}\{q_t \geq S_t\}$, and *Width*, defined as q_t (since the prediction interval $[\hat{f}(X_t) - q_t,\ \hat{f}(X_t) + q_t]$ has length $2q_t$). For each metric and each group, we report the average value over the entire decision horizon. Therefore, for each problem instance, the algorithm performance is summarized by two arrays, one for each metric, where each element represents the value of a particular group. Each problem instance is parameterized by the horizon length T. For each T, we repeat the experiment 40 times with independently generated data streams and report the per-group means. The data stream is designed to be non-stationary. Detailed experimental settings are provided in Appendix C.

4.1. Results: Compare with Existing Baselines

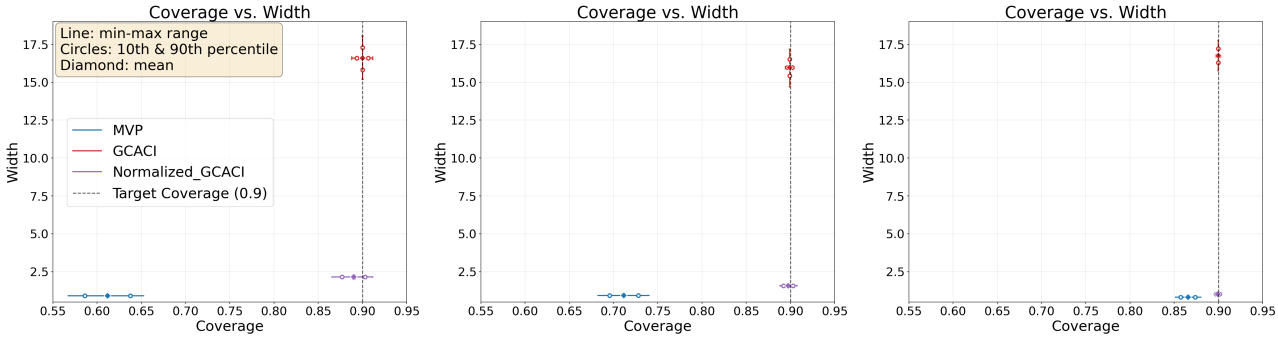
We run the experiment for $T = 600, 1200, 2400$. The results are shown in Figure 1. In each figure, we evaluate algorithm performance across two metrics: Coverage (x-axis) and Width (y-axis). Each algorithm is visualized as a cross-shaped marker: a horizontal segment representing the distribution of the Coverage array, and a vertical segment representing the distribution of the Width array. For each segment, the two endpoints denote the minimum and maximum values of the corresponding array, the two hollow circles mark the 10th and 90th percentiles, and the diamond indicates the mean.

The figures demonstrate that our Normalized-GCACI consistently achieves coverage close to the target level of $1 - \delta = 0.9$ across all groups and all values of T, while maintaining a significantly smaller prediction set width compared to the baselines. This confirms that our normalization approach effectively balances coverage accuracy and prediction efficiency.

4.2. Discussions: The necessity of our normalization trick

In this section, we demonstrate the effectiveness and necessity of our normalization approach by conducting an ablation study. A natural question arises: *Given that our Normalized-GCACI differs from the original GCACI only in normalizing q_t, could standard regularizations like clipping or projection on θ_t achieve similar gains?* To investigate, we compare our method against four closely related GCACI variants with different parameter constraints as follows, testing whether limiting the magnitude or sign of θ_t, without normalizing q_t, can replicate the benefits of our normalization:

- **GCACI**: The original GCACI algorithm in (Ramalingam et al., 2025), without any modification.
- **GCACI-NONNEG**: GCACI with non-negativity clipping: After each update on θ_t, the algorithm projects



(a) $T = 600$ *(b)* $T = 1200$ *(c)* $T = 2400$

Figure 1. Synthetic Non-stationary Linear Regression Experiment result: Coverage v.s. Width

onto the non-negative orthant via $\theta_t = \max(\theta_t, 0)$.

- **GCACI-L2**: GCACI with unit-ball projection: After each update on θ_t, the algorithm projects onto the unit ball via $\theta_t = \theta_t / \|\theta_t\|_2$ whenever $\|\theta_t\|_2 > 1$.
- **GCACI-NONNEG-L2**: GCACI with both non-negativity clipping and unit-ball projection applied sequentially after each update.

The result is shown in Figure 2. The result demonstrates that none of these alternatives achieve comparable performance, confirming that our normalization of q_t is essential and cannot be substituted by simple parameter regularization.

We further investigate whether decaying step sizes, as proposed by (Angelopoulos et al., 2024), can yield improvements comparable to our normalization technique. We compare against two additional variants:

- **GCACI-decay-uniform**: GCACI with a uniform decaying step size $\eta_t = t^{-0.5-\epsilon}$ with $\epsilon = 0.1$, following the setting suggested by (Angelopoulos et al., 2024).
- **GCACI-decay-groupwise**: GCACI with group-wise decaying step sizes $\eta_t = (\eta_t(i))_{i\in[m]}$, where $\eta_t(i) = N_t(i)^{-0.5-\epsilon}$ with $\epsilon = 0.1$ and $N_t(i) = \sum_{\tau=1}^{t} \mathbf{1}\{g_\tau(i) = 1\}$ denotes the number of rounds in which group i has appeared up to time t. The update rule becomes

$$\begin{aligned}&\theta_{t+1}(i)\\ =&\theta_t(i) - \eta_t(i) \cdot (\mathbf{1}\{q_t \geq s_t(X_t, Y_t)\} - (1-\delta)) \cdot g_t(i).\end{aligned}$$

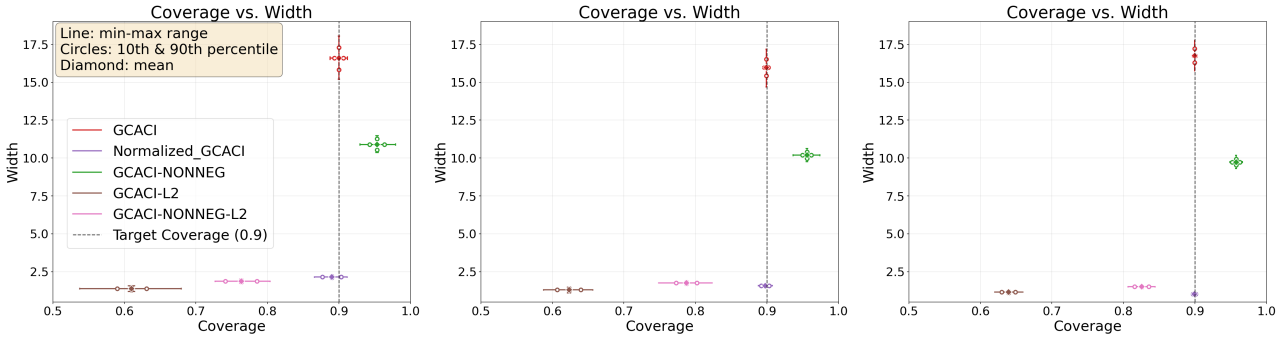
The results are shown in Figure 3. Both decaying step-size variants yield prediction sets that are substantially wider than those of Normalized-GCACI, with no meaningful gain in coverage accuracy over the original GCACI and the Normalized-GCACI. The decaying step-size variants exhibit significantly higher variance in coverage across different groups, indicating far less consistent performance than other methods. This further confirms that the benefit of our approach stems from the normalization of q_t itself, rather than from step-size scheduling. Appendix C.2 includes additional experiment results with different ϵ's.

5. Experiments: Efficient Document Retrieval

We apply our online conformal prediction framework to document retrieval, aiming to improve retrieval efficiency while preserving high recall of relevant evidence. Specifically, we are given a stream of queries. Each query is accompanied by a fixed set of candidate documents, among which one or a small number of "informative" documents contain sufficient evidence to answer the query. A pretrained encoder model assigns a relevance score to each (query, document) pair. For each query, our goal is to output a subset of the candidates that, with high probability, contains at least one informative document, while being strictly smaller than the full candidate set.

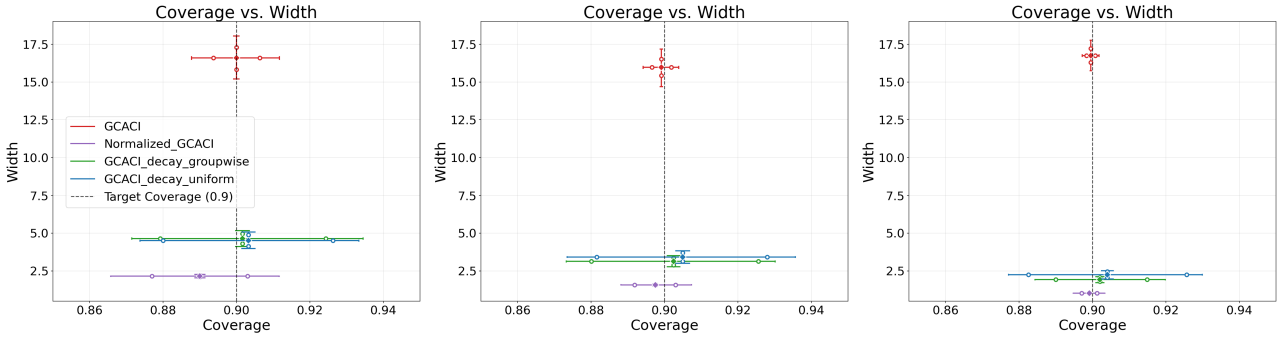
This setting is directly motivated by retrieval-augmented generation (RAG). In standard RAG, a language model (LM) conditions on a set of retrieved documents to generate an answer, and both inference latency and answer quality are sensitive to the size and relevance of that set. Passing the entire candidate pool to the LM is wasteful, and can dilute the signal with distractors. By contrast, simple strategies, like naively truncating to a fixed top-k, may silently discard critical evidence. Constructing a prediction set with a provable coverage guarantee on the informative document offers a principled middle ground. It reduces the number of passages processed by the LM, and thus the cost of downstream generation, while explicitly controlling the probability of discarding the evidence needed to answer the query.

We view this question in an online fashion: Queries arrive sequentially, one per round, and no distributional assumption is required on the query stream. We adapt our online conformal prediction framework to construct the predicted subset for each query sequentially. At each time step, the



(a) $T = 600$ *(b)* $T = 1200$ *(c)* $T = 2400$

Figure 2. Compare "Normalization" with "Clipping" or "Projecting"



(a) $T = 600$ *(b)* $T = 1200$ *(c)* $T = 2400$

Figure 3. Compare with "Decaying Step Sizes"

learner observes the query together with its candidate set and their relevance scores, and then construct a conformal prediction set, which is a subset of the candidate set. The environment subsequently reveals whether the set contains an "informative" document. Ultimately, our experiments show that our Normalized-GCACI can substantially shrink the size of the retained candidate set while maintaining a desired coverage (recall) level, and that it does so uniformly across all groups of queries.

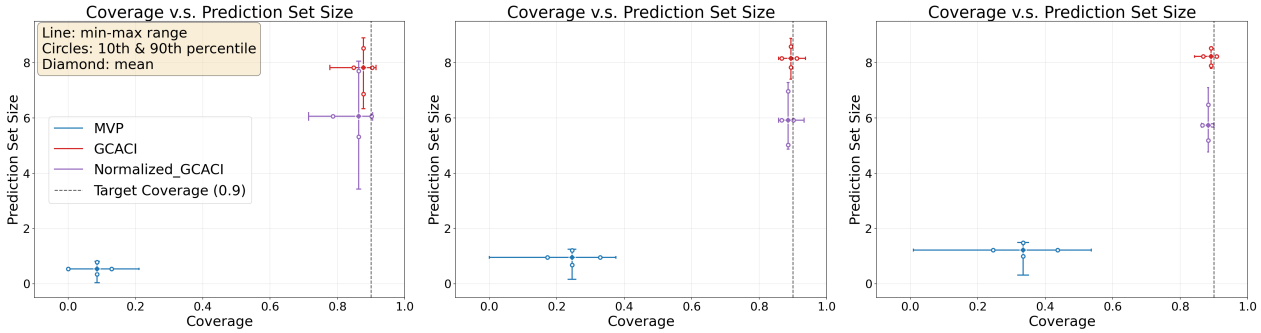
While RAG literature often calls the probability of retaining a relevant document recall, this is precisely our conformal coverage. We use "coverage" for consistency with the rest of the paper.

5.1. Experiment settings

Dataset We conduct experiments on HotpotQA dataset (Yang et al., 2018), a widely used benchmark for multi–hop question answering over Wikipedia. The dataset contains approximately 113K natural-language questions, each of which requires reasoning over and synthesizing evidence from multiple supporting documents. Every example is annotated with a question, an answer, and sentence-level supporting facts that identify the reasoning path across documents, making the dataset well suited for evaluating retrieval and evidence-aggregation methods in multi-hop settings.

We use the development set under the *distractor* setting. Here, each question is paired with 10 candidate paragraphs: One or two are guaranteed to be relevant and sufficient to answer the question, while the remaining paragraphs serve as distractors. Each paragraph is presented as a Wikipedia article title together with its text. HotpotQA further annotates each question with a reasoning type: *comparison* questions, which typically admit yes/no answers, and *bridge* questions, which require identifying an intermediate entity that connects multiple facts. See `https://hotpotqa.github.io/` for additional details.

Non-conformity Score Function and Encoder Model We define the non-conformity score as follows, following a similar way as in (Angelopoulos et al., 2020; Bastani et al., 2022; Angelopoulos et al., 2024). Given a query x and a candidate document set $\mathcal{D}$, let $\hat{f}$ be an encoder model that assigns a relevance score $\hat{f}(x, d)$ to each $d \in \mathcal{D}$, and let $\pi_1(x), \pi_2(x), \ldots$ denote the documents sorted in decreasing order of $\hat{f}(x, \cdot)$. For any document $d \in \mathcal{D}$, let $k(x, d)$



(a) $T = 240$ *(b)* $T = 360$ *(c)* $T = 480$

Figure 4. RAG experiment result: Coverage v.s. Prediction set size

be its rank in this ordering. Since the ground-truth label $y_t \subseteq \mathcal{D}$ may contain several relevant documents, we define

$$k(x, y_t) := \max_{d \in y_t} k(x, d),$$

i.e., the worst (largest) rank among all relevant documents. The non-conformity score is then defined as

$$s_t(x, y_t) \;=\; \sum_{i=1}^{k(x,y_t)} \hat{f}\big(x, \pi_i(x)\big).$$

For $\hat{f}$, we directly adopt the pretrained cross-encoder `bge-reranker-base`, without any additional training or fine-tuning. For each (query, document) pair, the model outputs a relevance logit. We then apply a softmax over the logits of all candidates to obtain normalized relevance scores, which serve as inputs to the online conformal prediction procedure.

Grouping We introduce a country-level grouping of questions based on the metadata of their associated Wikipedia documents. This grouping is not part of the original dataset and is constructed entirely as part of our experimental design. In the HotpotQA distractor setting, each question has exactly 10 Wikipedia article titles. We query the Wikipedia API to retrieve the category keywords associated with each title. Since a single article may belong to multiple categories, this yields a set of category keywords per title. Questions for which category information cannot be reliably retrieved for any of the associated titles are discarded, leaving a filtered dataset of 5,935 questions.

Using the collected category keywords, we perform fuzzy matching, over both the keywords and their lexical variants, to determine whether a document is associated with a country drawn from a predefined list. We restrict our analysis to 13 countries, chosen so that each appears sufficiently often to support meaningful group-wise evaluation. Countries with very few associated questions are excluded, since their small sample sizes would yield noisy coverage estimates and provide little information about group-conditional performance. The resulting per-country question counts are reported in Appendix 5. A question may be associated with multiple countries if its documents contain category keywords related to different countries, or with none if no such association is found. The resulting country information is attached as additional metadata to each question and is used for grouping and analysis throughout our experiments.

5.2. Results

The results are summarized in Figure 4 which compares our proposed Normalized-GCACI against the baseline Algorithm MVP and Algorithm GCACI for time horizons $T \in \{240, 360, 480\}$ on the HotpotQA retrieval task. Across all settings, Normalized-GCACI consistently achieves group-wise coverage close to the target level, while producing small prediction sets. In contrast, GCACI yields considerably larger prediction sets, and MVP exhibits slower convergence and significantly larger variance in terms of coverage. These findings highlight the advantage of our Normalized GCACI approach: It achieves a more favorable coverage v.s. prediction set size trade-off, and translates our theoretical findings into tangible gains in a practical RAG retrieval setting. We conclude our paper in Appendix A.

Acknowledgement

This work is supported by the Singapore Ministry of Education AcRF Tier 2 Grant [Proposal ID T2EP20124-0037, Award number MOE-T2EP20124-0008].

References

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Angelopoulos, A., Candes, E., and Tibshirani, R. J. Conformal pid control for time series prediction. *Advances in*

neural information processing systems, 36:23047–23074, 2023.

Angelopoulos, A. N., Barber, R. F., and Bates, S. Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*, 2024.

Areces, F., Mohri, C., Hashimoto, T., and Duchi, J. Online conformal prediction via online optimization. In *Forty-second International Conference on Machine Learning*, 2025.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramdas, A., and Roth, A. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29362–29373, 2022. URL proceedings.neurips.cc.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363. PMLR, 2023.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms, 3rd Edition*. MIT Press, 2009. ISBN 978-0-262-03384-8.

Feng, N., Sui, Y., Hou, S., Cresswell, J. C., and Wu, G. Response quality assessment for retrieval-augmented generation via conditional conformal factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2832–2836, 2025.

Ge, H., Bastani, H., and Bastani, O. Stochastic online conformal prediction with semi-bandit feedback. *arXiv preprint arXiv:2405.13268*, 2024.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Gibbs, I. and Candès, E. J. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Gupta, V., Jung, C., Noarov, G., Pai, M. M., and Roth, A. Online multivalid learning: Means, moments, and prediction intervals. *arXiv preprint arXiv:2101.01739*, 2021.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Ramalingam, R., Kiyani, S., and Roth, A. The relationship between no-regret learning and online conformal prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=JAcOYCqNo9.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Stephen, B. et al. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv: 2107.07511*, 2021.

Susmann, H., Chambaz, A., and Josse, J. Adaptiveconformal: An r package for adaptive conformal inference. *arXiv preprint arXiv:2312.00448*, 2023.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer, 2005.

Yadkori, Y. A., Kuzborskij, I., Stutz, D., György, A., Fisch, A., Doucet, A., Beloshapka, I., Weng, W.-H., Yang, Y.-Y., Szepesvári, C., et al. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*, 2018.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. Adaptive conformal predictions for time series. In *International conference on machine learning*, pp. 25834–25866. PMLR, 2022.

Zhou, X., Chen, B., Gui, Y., and Cheng, L. Conformal prediction: A data perspective. *ACM computing surveys*, 58(2):1–37, 2025.

A. Concluding Remarks

In this work, we studied group-wise online conformal prediction in the adversarial setting, and proposed Normalized-GCACI, a simple yet effective modification of the GCACI algorithm. The only change is to normalize the prediction threshold by the number of active groups, rather than summing the group-level parameters. Despite its simplicity, this single structural change yields three concrete benefits. Theoretically, it induces a uniform boundedness property on the group-level parameters, which in turn leads to a strictly tighter group-wise coverage bound than prior work and extends the known single-group convergence rate to the general multi-group setting. Empirically, it consistently produces smaller prediction sets while tracking the target coverage more closely than existing baselines, and our ablations confirm that the improvement is specifically attributable to the normalization itself, rather than to parameter regularization or step-size decay. Practically, when applied to post-retrieval filtering for retrieval-augmented generation, our method substantially reduces the number of retained documents while preserving high recall of the relevant evidence. The main takeaway is that, for group-wise online conformal prediction, a careful rescaling at the threshold level is a remarkably cost-effective lever for improving both informativeness and theoretical guarantees.

Several directions naturally extend this work. First, it would be interesting to sharpen the dependence on the number of groups, either through refining the analysis or by designing algorithms that better exploit structural relationships among overlapping groups. Second, our retrieval-augmented generation experiments point to a promising application direction: integrating online conformal prediction more deeply into modern LLM pipelines, including agentic systems, tool use, and multi-step reasoning. In such settings, calibrated uncertainty over intermediate decisions could translate directly into improved efficiency and reliability for downstream generation.

B. Proof for Theorem 3.1

This theorem is directly obtained by the following lemmas, which we will prove later:

Lemma B.1. *Algorithm 1 can induce the following: For any $i \in [m]$*

$$\left| \frac{1}{|\mathcal{T}(i)|} \sum_{t \in \mathcal{T}(i)} \mathbf{1}(Y_t \in \mathcal{C}_t(q_t)) - (1-\delta) \right| \leq \frac{|\theta_{T+1}(i)|}{\eta \cdot |\mathcal{T}(i)|}$$

Lemma B.2. *For Algorithm 1, we can obtain: for any $i \in [m]$, $t \in [T+1]$,*

$$|\theta_t(i)| \leq O\left(\sqrt{T\eta m}\right)$$

Combine Lemma B.1, 3.2, B.2, then Theorem 3.1 holds immediately.

B.1. Proof for Lemma B.1

For any $i \in [m]$, denote $\mathcal{T}(i) = \{t_1, \cdots, t_{|\mathcal{T}(i)|}\}$ (for simplicity). Then for any $j \in [|\mathcal{T}(i)|]$, we have

$$\theta_{t_{j+1}}(i) = \theta_{t_j+1}(i) = \theta_{t_j}(i) + \eta((1-\delta) - \mathbf{1}\{q_{t_j} \geq S_{t_j}\}).$$

This is obvious, since only when $g_{t,i} = 1$, $\theta_t(i)$ can update value. Thus we have

$$(1-\delta) - \mathbf{1}\{q_{t_j} \geq S_{t_j}\} = \frac{1}{\eta} \cdot \left(\theta_{t_{j+1}}(i) - \theta_{t_j}(i)\right).$$

Summing over $\mathcal{T}(i)$ and we obtain the lemma.

B.2. Proof for Lemma 3.2

We prove the lemma by induction on m.

Base case $m = 1$: This simple case is same as Lemma 4.1 in (Gibbs & Candes, 2021). We provide the proof here for completeness. In this case, $q_t = \theta_t(1)$. For any t such that $q_t \geq 0$ but $q_{t+1} < 0$, then we must have

$$q_{t+1} = q_t - \theta\delta \geq 0 - \eta\delta = -\eta\delta$$

and $q_{t+1} < 0 \leq S_t$ ($S_t \in [0, 1]$ for all t). Hence, we will have

$$q_{t+2} = q_{t+1} + \eta(1 - \delta) > q_{t+1}$$

This process will continue until $q_{t+\ell} \geq 0$ for some $\ell \in \mathbb{N}^+$. During this process, it holds that

$$q_{t+\ell} > q_{t+\ell-1} > \cdots > q_{t+1} \geq -\eta\delta \geq -\eta = -\mathcal{B}_1.$$

Thus, for all t we must have $q_t \geq -\mathcal{B}_1$. Similarly, we have $q_t \leq 1 + \eta(1 - \delta) \leq 1 + \eta = 1 + \mathcal{B}_1$.

Induction step: Assume the claim holds for all numbers of coordinates up to m. Consider the algorithm with $m + 1$ coordinates. For any $i \in [m + 1]$ and $t \in [T]$, we need to prove $-\mathcal{B}_{m+1} \leq \theta_t(i) \leq 1 + \mathcal{B}_{m+1}$. Follow by the symmetry, we only need to prove the lower bound side.

Fix i. The worst case occurs when i is updated with $q_t \geq S_t$ and other coordinates are as large as possible. This is intuitive since a larger q_t makes it easier for $q_t \geq S_t$ to hold even when $\theta_t(i) < 0$ and $\theta_t(i)$ is very negative. Hence, to push $\theta_t(i)$ to the worst possible lower bound, the other coordinates should be set to their maximal possible values. More importantly, such lower bound can neither be pushed in a single set, nor be pushed by another single coordinate. Instead, this bound is obtained by a recursive construction that repeatedly uses the maximal values of subsets of the other coordinates.

What we need to prove is that the $\mathcal{B}_{m+1}$ satisfies the recurrence

$$\mathcal{B}_{m+1} = \sum_{k=1}^{m} (\mathcal{B}_k + 1) + \eta. \tag{3}$$

If (3) holds, then by the induction hypothesis, we immediately have

$$\mathcal{B}_{m+1} = \sum_{i=1}^{m} (\mathcal{B}_i + 1) + \eta = (1 + \eta) \sum_{i=1}^{m} 2^{i-1} + \eta = 2^m(1 + \eta) - 1,$$

and the proof is finished. The remaining task is to prove that (3) is true.

To drive $\theta_t(i)$ as negative as possible, we can raise a set of m other coordinates to their maximum possible values in a hierachical way:

- Choose one coordinate j_1 and bring it to $\mathcal{B}_m + 1$. It is possible by the induction hypothesis on the m coordinates excluding i.
- Among the remaining $m - 1$ coordinates, choose j_2 and bring it to $\mathcal{B}_{m-1} + 1$. It is possible by the induction hypothesis on the $m - 1$ coordinates excluding i, j_1.
- Continue until all m other coordinates are set to $\mathcal{B}_m + 1$, $\mathcal{B}_{m-1} + 1$, ..., $\mathcal{B}_1 + 1$.

Let x be the current value of $\theta_t(i)$. To drive x to be its worst lower bound, we need to activate all coordinates simultaneously, that is $g_t = \mathbf{1}_{m+1} = (1, \ldots, 1)$. Then we have

$$q_t = \frac{x + \sum_{k=1}^{m} (1 + \mathcal{B}_k)}{m + 1}.$$

In the extreme case that $S_t = 0$, to ensure that $q_t \geq S_t$, x must satisfy

$$x \geq -\sum_{k=1}^{m} (1 + \mathcal{B}_k)$$

When it comes to this step, we still have $q_t \geq S_t = 0$ so we still have update on x by the update rule:

$$x \leftarrow x - \eta\delta$$

Then next step we will definitely have $q_t < 0 \leq S_t$ and the value of x and $\theta_t(i)$ will increase and be away from the lower bound. Consequently, we can see the worst possible lower bound for x:

$$x \geq -\sum_{k=1}^{m} (1 + \mathcal{B}_k) - \eta\delta \geq -\sum_{k=1}^{m} (1 + \mathcal{B}_k) - \eta := -\mathcal{B}_{m+1}$$

Thus we prove (3) and the lemma.

B.3. Proof for Lemma B.2

This proof follows the same line as the proof for Lemma 5.2 in (Ramalingam et al., 2025).

Since $S_t \in [0, 1]$, we have the follwing observations:

- If $q_t < 0$, then $q_t \leq S_t$.
- If $q_t > 1$, then $q_t > S_t$.

We consider $\|\theta_t\|_2$. When $q_t < S_t$,

$$\begin{aligned}\|\theta_{t+1}\|_2^2 &= \|\theta_t\|_2^2 + 2\eta(1-\delta)\langle\theta_t, g_t\rangle + \eta^2 \cdot (1-\delta)^2\|g_t\|_2^2 \\ &= \|\theta_t\|_2^2 + 2\eta(1-\delta)q_t \cdot \|g_t\|_1 + \eta^2 \cdot (1-\delta)^2\|g_t\|_2^2 \\ &\leq \|\theta_t\|_2^2 + 2\eta(1-\delta)m + \eta^2(1-\delta)^2 m\end{aligned}$$

Similarly, when $q_t > S_t$,

$$\begin{aligned}\|\theta_{t+1}\|_2^2 &= \|\theta_t\|_2^2 - 2\eta\delta\langle\theta_t, g_t\rangle + \eta^2 \cdot \delta^2\|g_t\|_2^2 \\ &= \|\theta_t\|_2^2 - 2\eta\delta q_t \cdot \|g_t\|_1 + \eta^2 \cdot \delta^2\|g_t\|_2^2 \\ &\leq \|\theta_t\|_2^2 + \eta^2\delta^2 m\end{aligned}$$

Overall we have

$$\|\theta_{T+1}\|_2^2 \leq T(\eta^2 \cdot m \cdot \max\{\delta, 1-\delta\} + 2\eta(1-\delta)m) = T\eta m(\eta \cdot \max\{\delta, 1-\delta\} + 2)$$

Thus we have

$$\theta_{t,i} \leq \|\theta_{T+1}\|_\infty \leq O(\sqrt{T\eta m})$$

C. Details for Section 4

C.1. Detailed Experiments Setup

Feature Space and Data Generation The experiment setting is similar as Section 4.1.2 in (Bastani et al., 2022). The feature space is $\mathcal{X} = \mathbb{R}^{250}$, consisting of 100 binary features and 150 continuous features. For any feature vector $x \in \mathcal{X}$, we write $x = \{x_i\}_{i=1}^{250}$, where the first 100 coordinates correspond to the binary features and the remaining 150 coordinates correspond to the continuous features. The corresponding label y is generated according to a linear model with heteroscedastic noise:

$$y = \langle\theta, x\rangle + N\left(0, \sigma_y^2 + \sum_{i=1}^{100} \sigma_{x,i}^2 x_i\right),$$

where $\theta \in \mathbb{R}^{250}$ is an unknown parameter vector, and $(\sigma_{x,i})_{i\in[100]})$ controls the feature-dependent noise level.

For better control of the data generation process, we first construct a large synthetic dataset where each row consists of a pair (x, y), whose format is described above. For each experimental run, we randomly sample data points from this dataset. For any $x \in \mathbb{R}^{250}$, the binary features are independently drawn from a Bernoulli distribution with parameter 0.5, and the continuous features are independently drawn from a Gaussian distribution $N(0, \sigma_x^2)$. To study the impact of environmental non-stationarity on algorithm performance, we partition the dataset into four equal-sized segments. Each segment follows the same data format for (x, y), but uses different parameter values $(\theta, \sigma_y, \sigma_x, (\sigma_{x,i})_{i\in[100]})$.

The parameter for the 4 segments are set as follows:

1. $\sigma_x = 0.1$, each $\theta = (\theta_j)_{j\in[250]}$ is generated by $N(0, \sigma_x^2)$ independently, each continuous X_i is generated by $N(0, \sigma_x^2)$ independently. $\sigma_y = 0.25$, $\sigma_{x,1} = 5.0$, $\sigma_{x,i} = 0.5$ for $i \neq 1$.
2. $\sigma_x = 0.5$, each $\theta = (\theta_j)_{j\in[250]}$ is generated by $N(0, \sigma_x^2)$ independently, each continuous X_i is generated by $N(0, \sigma_x^2)$ independently. $\sigma_y = 0.25$, $\sigma_{x,2} = 5.0$, $\sigma_{x,i} = 0.5$ for $i \neq 2$.

3. $\sigma_x = 0.1$, each $\theta = (\theta_j)_{j\in[250]}$ is generated by $N(0, \sigma_x^2)$ independently, each continuous X_i is generated by $N(0, \sigma_x^2)$ independently. $\sigma_y = 0.25$, $\sigma_{x,3} = 5.0$, $\sigma_{x,i} = 0.5$ for $i \neq 3$.

4. $\sigma_x = 0.5$, each $\theta = (\theta_j)_{j\in[250]}$ is generated by $N(0, \sigma_x^2)$ independently, each continuous X_i is generated by $N(0, \sigma_x^2)$ independently. $\sigma_y = 0.25$, $\sigma_{x,4} = 5.0$, $\sigma_{x,i} = 0.5$ for $i \neq 4$.

Each of the four segments contains 20,000 data points. In each experimental run, we sample data from all four segments using the same sampling ratio. The sampled data are then concatenated in the order of the first, second, third, and fourth segments to form the final data stream used in the experiment. This construction allows us to simulate a non-stationary environment where the underlying data distribution changes across different phases.

Non-conformity Score Function We define the nonconformity score function as $s_t(x, y) = |y - \hat{f}(x)|$, where $\hat{f}$ is a pre-trained linear regression model. In each experimental run, we first train $\hat{f}$ using data from the first segment of the dataset. The model is then fixed, and the data from the remaining three segments are used to run the online conformal prediction algorithms.

Grouping We use the first 50 columns of binary features to define 100 (intersecting) groups defined by the value of each of the 50 binary features. This design is motivated by the experimental setting in (Bastani et al., 2022).

C.2. Additional Results: Compare with Decaying Step Sizes

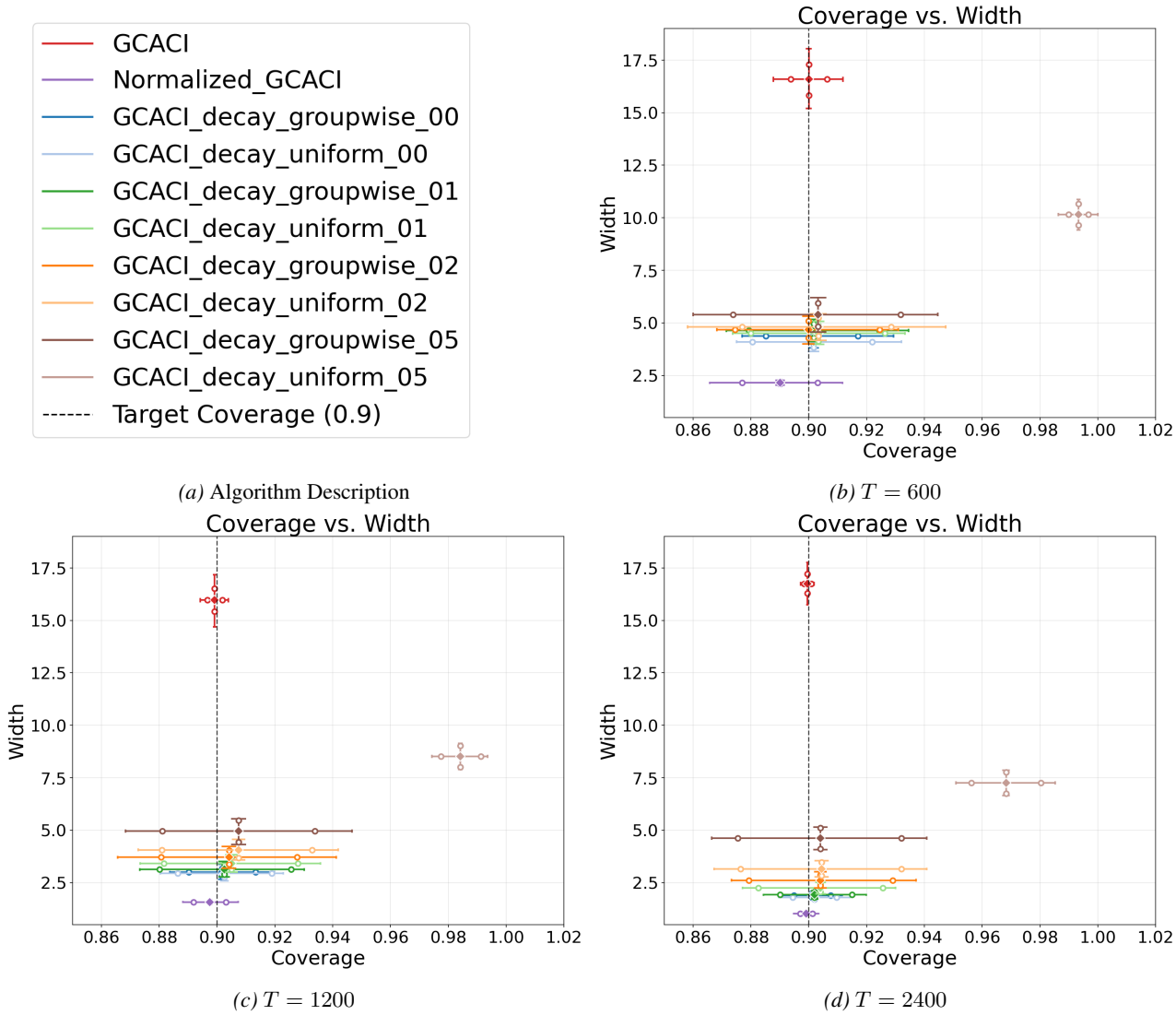
To further investigate whether decaying step sizes, as proposed by (Angelopoulos et al., 2024), can yield improvements comparable to our normalization technique. We compare against two additional variants, **GCACI-decay-uniform**, and **GCACI-decay-groupwise**, in Section 4.2. Here we provide addtional results regarding these two methods with different decaying ratio.

- **GCACI-decay-uniform-0x**: GCACI with a uniform decaying step size $\eta_t = t^{-0.5-\epsilon}$ with $\epsilon = 0.1 \cdot x$.
- **GCACI-decay-groupwise-0x**: GCACI with group-wise decaying step sizes $\eta_t = (\eta_t(i))_{i\in[m]}$, where $\eta_t(i) = N_t(i)^{-0.5-\epsilon}$ with $\epsilon = 0.1 \cdot x$ and $N_t(i) = \sum_{\tau=1}^{t} \mathbf{1}\{g_\tau(i) = 1\}$ denotes the number of rounds in which group i has appeared up to time t.

The results are shown in Figures 5. All decaying step-size variants produce prediction sets that are substantially wider than those of Normalized-GCACI, without yielding any meaningful improvement in coverage accuracy. This further confirms that the advantage of our approach arises from the normalization of q_t itself, rather than from step-size scheduling.

D. Details for Section 5

D.1. Country Count



(a) Algorithm Description

(b) $T = 600$

(c) $T = 1200$

(d) $T = 2400$

Figure 5. Compare with "Decaying Step Sizes"

Country	Count
United States of America	3,279
United Kingdom	1,449
Germany	966
France	836
Indonesia	702
Italy	508
Spain	494
Canada	421
Australia	367
Netherlands	326
Russia	296
Japan	263
China	215

Table 1. Country Count