DUAL-CALIBRATED ONLINE CONFORMAL INFERENCE WITH HYBRID QUANTILE LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal inference provides rigorous coverage guarantees for uncertainty quantification, yet its dynamic coverage accuracy is highly susceptible to distribution shifts in non-stationary environments. Existing approaches construct prediction sets via dynamically estimated conditional quantiles but suffer from two key limitations. First, most algorithms passively correct for coverage deviations, which often results in excessively conservative prediction intervals. Second, prevailing update strategies rely on the standard quantile loss, whose inherent drawbacks undermine coverage adaptability. To address these issues, we introduce an ε-insensitive Hybrid Quantile (ε-HQ) loss with a bounded quadratic component, formally establish its Lipschitz continuity and generalized weak pseudo-convexity (GWPC), and design a gain-balancing mechanism to match gain magnitudes between the two components. Building upon this loss, we propose a **D**ual-**C**alibrated **O**nline **C**onformal **I**nference framework, ε-**HQ-DCOCI**. The first calibration refines gradient signals via adaptive residual mapping and insensitive-zone tuning, enabling continuous feedback correction. The second calibration accelerates threshold updates under persistent drift via momentum-based smoothing of historical subgradient directions. Theoretically, we prove that ε-**HQ-DCOCI** achieves sublinear regret bounds and approximately valid coverage in non-stationary settings. Experiments on real-world datasets, including financial time series and medical image classification, demonstrate that our framework consistently yields tighter and more efficient prediction sets compared to state-of-the-art baselines.

1 INTRODUCTION

Uncertainty quantification is a crucial issue in machine learning. In high-stakes decision-making scenarios such as financial analysis and medical diagnosis, predictors must provide not only accurate predictions but also reliable confidence (Abdar et al., 2021; Chakraborti et al., 2025). Conformal inference (Romano et al., 2019), as a model-agnostic and distribution-free framework, constructs prediction sets with finite-sample coverage guarantees for any black-box model, and is fundamentally different from Bayesian inference (Magris & Iosifidis, 2023) and deep ensembles (Lakshminarayanan et al., 2017). Relying only on the exchangeability, it achieves rigorous marginal coverage via post-processing, driving its widespread adoption (Lee & Barber, 2021; Zhou et al., 2025).

In non-stationary environments, however, distribution shifts violate exchangeability and break standard coverage guarantees. Gibbs & Candès (2021) introduced adaptive conformal inference (ACI), which first formulated threshold adaptation as online gradient descent (OGD) to maintain long-run coverage by tuning confidence levels. Subsequent work has advanced online conformal inference (OCI) along several lines: sequential predictive conformal inference (SPCI) aims to estimate conditional quantiles (Xu & Xie, 2023); error-quantified conformal inference (ECI) attempts to incorporate an error-quantified term (Wu et al., 2023); linear quantile tracking (LQT) method was proposed to approximate quantiles (Areces et al., 2025); and conformal optimistic prediction (COP) minimizes the expected conditional quantile errors to exploit distributional information (Hu et al., 2026).

Despite varied emphases, most OCI methods still adopt the *standard quantile loss* (Q loss) (Koenker & Bassett, 1978) as the update signal, attending only to instantaneous coverage error. This paradigm has three inherent bottlenecks. (1) *Excessively coarse feedback*: the subgradient of Q loss is a binary coverage indicator that ignores error magnitude. (2) *Unattenuated outlier response*: Q loss lacks

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

an active mechanism to suppress anomalous nonconformity scores, leaving threshold estimation vulnerable to abrupt noise in non-stationary settings. (3) *Lack of trend awareness*: most methods react passively to coverage deviations, disregarding the shift direction and persistence, which causes over-expansion under sustained drift. Consequently, existing methods show clear deficiencies in *information efficiency*, *anomaly robustness*, and *trend adaptability*. Although some methods, such as ECI and COP, incorporate gradient correction terms to mitigate these deficiencies, their correction signals lack magnitude coordination with the original binary gradients, often causing gain mismatch. Overcoming these bottlenecks requires designing a novel loss with fine-grained gradient feedback, and leveraging historical information to detect drift patterns for more efficient coverage correction.

The coarse gradient and unboundedness of Q loss have motivated smoothing and bounding strategies. Smoothing approaches include the asymmetric least squares (ALS) loss (Newey & Powell, 1987) and its insensitive variant (Wu et al., 2023); bounding approaches comprise the truncated pinball loss (Shen et al., 2017), bounded quantile (BQ) loss (Zhang & Yang, 2024), and bounded asymmetric least squares (BALS) loss (Zhang & Yang, 2025). However, these losses bring new challenges in OCI: boundedness destroys global convexity and complicates regret analysis (Wang et al., 2024; Xu & Zhang, 2024); gradient saturation away from zero impedes sustained threshold updates; and fully smoothed losses can degrade early-stage coverage efficiency (Wu et al., 2025).

To address these challenges, we propose an ε-insensitive hybrid quantile loss with a bounded quadratic term (ε-**HQ** loss) and develop a Dual-Calibrated Online Conformal Inference framework (ε-**HQ-DCOCI**). The loss comprises three components: (1) a bounded quadratic term that enriches gradient signals and suppresses anomalous perturbations, while its generalized weak pseudo-convexity facilitates regret analysis; (2) a linear component that maintains a constant gradient magnitude away from the origin to mitigate saturation, while the hybrid structure induces gradient jumps at the insensitive boundary to preserve responses to small signals; (3) an ε-insensitive zone that filters out minor fluctuations, thereby improving prediction stability. We further introduce a gain-balancing mechanism to match gain magnitudes between the two loss components, upon which we design a dual calibration mechanism: the *first calibration* delivers precise continuous gradient feedback via adaptive residual mapping and insensitive-zone tuning, while the *second calibration* smooths historical subgradient directions with momentum to perceive drift trends actively. Together, they overcome the three principal deficiencies of existing methods in information efficiency, anomaly robustness, and trend adaptability. Theoretically, we prove that ε-HQ-DCOCI achieves sublinear dynamic regret and approximately valid coverage under non-stationarity. Our main contributions are as follows.

- We propose the ε-HQ loss function and systematically prove its Lipschitz continuity and generalized weak pseudo-convexity.
- We design the ε-HQ-DCOCI framework, where the first calibration provides precise gradient feedback and the second perceives drift trends via a momentum smoothing term.
- We establish rigorous theoretical guarantees for ε-HQ-DCOCI, including the sublinear base and joint dynamic regret bounds and approximately valid coverage in the worst case.
- Evaluations on six real-world datasets covering financial time series and medical images demonstrate that ε-HQ-DCOCI consistently yields tighter and more valid prediction sets.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 defines the ε-HQ loss and analyzes its properties, Section 4 presents the ε-**HQ-DCOCI** algorithm framework, Section 5 provides the theoretical analysis, Section 6 reports experimental results, and Section 7 concludes the paper.

2 Background

2.1 Problem Setup

Let $\{(X_t, Y_t)\}_{t\geq 1}$ be a data sequence streaming over time, where $X_t \in \mathcal{X}$ denotes the input features and $Y_t \in \mathcal{Y}$ the response variable. The goal of OCI is to construct, at each time step t, a prediction set $\mathcal{C}_t(X_t) \subseteq \mathcal{Y}$ such that the long-run empirical coverage converges to the target level $1-\alpha$, i.e.,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{Y_t \in \mathcal{C}_t(X_t)\} = 1-\alpha,$$

where $\alpha \in (0,1)$ is the target miscoverage rate. For regression, given a predictor $f : \mathcal{X} \to \mathbb{R}$, the prediction set takes the symmetric interval $\mathcal{C}_t(X_t) = [f(X_t) - \hat{q}_t, f(X_t) + \hat{q}_t]$, with dynamic threshold $\hat{q}_t \geq 0$ and nonconformity score $s_t = |Y_t - f(X_t)|$. For classification, $\mathcal{C}_t(X_t) = \{y \in \mathcal{Y} : s(X_t, y) \leq \hat{q}_t\}$ is a label subset obtained by thresholding class-conditional scores. We follow the assumptions of (Angelopoulos et al., 2024): **(1)** *nonconformity scores are bounded*, i.e., there exists $R > 0$ such that $s_t \in [0, R]$; and **(2)** *the threshold is initialized reasonably*, i.e., $\hat{q}_1 \in [0, R]$.

The core challenge of OCI is updating the threshold online under distribution shift. The general update is $\hat{q}_{t+1} = \hat{q}_t - \eta \nabla \ell_t(\hat{q}_t)$, with loss ℓ_t computed from the current observation. The manner in which historical information is used to form the feedback signal $\nabla \ell_t(\hat{q}_t)$ directly governs the algorithm's accuracy in sensing coverage deviations and its efficiency in adapting to non-stationarity.

2.2 Online Conformal Inference

ACI (Gibbs & Candès, 2021) pioneered OGD-based threshold updating using the Q loss, but relies on exchangeability via a fixed calibration set and a step size sensitive to environmental changes. Under distribution shift, the confidence threshold α_t may exit the admissible range, degenerating the prediction set. Dynamically-tuned adaptive conformal inference (DtACI) (Gibbs & Candès, 2024) adapts the step size via online expert aggregation with forgetting, but still updates the interval radius through the intermediate α_t. SPCI (Xu & Xie, 2023) addresses this degeneration by estimating conditional quantiles directly via offline quantile regression, yet its coverage accuracy is constrained by estimation error. SF-OGD (Bhatnagar et al., 2023) and Decay-OGD (Angelopoulos et al., 2024) instead update $\hat{q}_t$ directly via OGD, respectively applying gradient normalization and step decay to reduce learning-rate sensitivity, but leave the coarse gradient of the Q loss unaddressed.

To remedy the binary feedback gap, several works smooth the gradient or inject extra information. PIDCP (Angelopoulos et al., 2023) injects cumulative gradient and a score predictor via PID control, but relies heavily on external component accuracy. ECI (Wu et al., 2025) smooths the binary indicator via a Sigmoid function and adds an error term, locally approximating the Q loss gradient. LQT (Areces et al., 2025) extends threshold updates to OGD on a parametric predictor, with accuracy tied to the external predictor. COP (Hu et al., 2026) converts coverage errors to probability-scale deviations by minimizing the expected threshold deviation, then applies an optimistic correction via the estimated cumulative distribution. These methods embed extra estimation modules, adding hyperparameters and estimation errors. More critically, for those with auxiliary corrections, the correction signals are hyperparameter-controlled and scale-mismatched with the binary gradient: overly large corrections disrupt unbiased quantile tracking, while overly small ones are negligible.

While these methods improve non-stationary adaptability via gradient adjustment, loss smoothing, or external components, none fundamentally alter the threshold-update loss, which may compromise the stability and efficiency of threshold estimation under rapid shifts or anomalous scores.

2.3 Related Loss Functions

As outlined in Section 1, the coarse gradient and unboundedness of the standard quantile (Q) loss (Koenker & Bassett, 1978) have motivated a series of smoothing and bounding strategies. The asymmetric least squares (ALS) loss (Newey & Powell, 1987) enriches gradient information via a squared formulation but amplifies outlier sensitivity. Its ε-insensitive variant (ε-ALS loss) (Wu et al., 2023) mitigates minor-deviation oscillations. To address unboundedness, Fu et al. (2023) proposed a general framework for bounded loss functions in regression (BLFR) inspired by the bounded linear exponential (BLinex) loss (Wen & Levy, 2001), subsequently extended to the bounded quantile (BQ) loss (Zhang & Yang, 2024) and the bounded asymmetric least squares (BALS) loss (Zhang & Yang, 2025). Although BALS addresses simultaneously gradient coarseness and unboundedness, its global smoothness can degrade early-stage coverage efficiency in OCI (Wu et al., 2025). Moreover, bounded losses sacrifice convexity, complicating dynamic regret analysis (Wang et al., 2024; Xu & Zhang, 2024), and suffer gradient saturation away from zero, impeding sustained threshold updates.

In summary, existing loss functions do not natively accommodate the requirements of OCI. To bridge this gap, we start from this core component and propose an ε-insensitive hybrid quantile loss with a bounded quadratic term (ε-HQ loss), upon which the ε-HQ-DCOCI framework is built. The relevant design, theory, and empirical validation are detailed in the following sections.

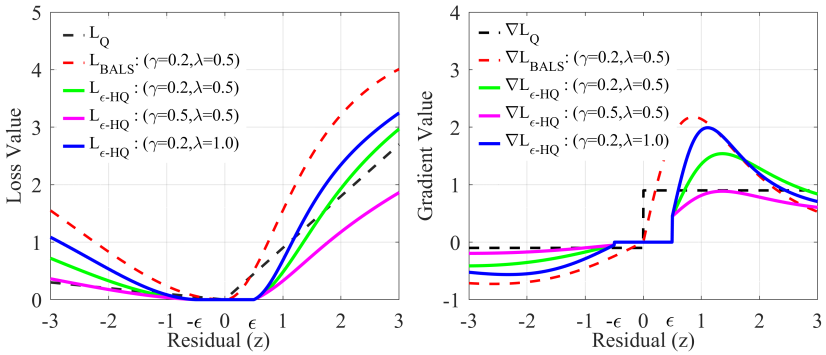

Figure 1: Visual comparison of the ε-HQ loss ($\alpha = 0.1$, $\varepsilon = 0.5$, $\tau = 0.5$) with the Q loss and BALS loss (both corresponding to $\varepsilon = 0$), together with their (sub)gradients.

3 Hybrid Quantile Loss

3.1 Definition

Let $z = s_t - q_t$ be the residual between the nonconformity score s_t and the dynamic threshold q_t at time t. We define the ε-HQ loss as a convex combination of the insensitive quantile loss $L_{\varepsilon\text{-Q}}$ and the insensitive bounded asymmetric least squares loss $L_{\varepsilon\text{-BALS}}$:

$$L_{\varepsilon\text{-HQ}}(z) = \tau\, L_{\varepsilon\text{-Q}}(z;\alpha,\varepsilon) + (1-\tau)\, L_{\varepsilon\text{-BALS}}(z;\alpha,\varepsilon,\gamma,\lambda), \tag{1}$$

where $\alpha \in (0,1)$ is target miscoverage rate, ε is the insensitive zone width, and $\tau \in (0,1)$ balances the linear and bounded quadratic terms. The latter is defined as $L_{\varepsilon\text{-BALS}}(z) = \frac{1}{\gamma}\big[1 - \frac{1}{1+\lambda L_{\varepsilon\text{-ALS}}(z)}\big]$, with $\gamma, \lambda > 0$ controlling the upper bound and near-origin sensitivity. The two base losses are

$$L_{\varepsilon\text{-Q}}(z) = \begin{cases} -\alpha(z+\varepsilon), & z < -\varepsilon, \\ 0, & |z| \le \varepsilon, \\ (1-\alpha)(z-\varepsilon), & z > \varepsilon, \end{cases} \quad L_{\varepsilon\text{-ALS}}(z) = \begin{cases} \alpha(z+\varepsilon)^2, & z < -\varepsilon, \\ 0, & |z| \le \varepsilon, \\ (1-\alpha)(z-\varepsilon)^2, & z > \varepsilon. \end{cases} \tag{2}$$

Figure 1 illustrates the parameter dependence of the ε-HQ loss and its gradient. **Left**: increasing γ weakens the penalty, larger λ sharpens near-origin sensitivity; τ interpolates between the linear-dominated regime ($\tau = 1$, i.e., $L_{\varepsilon\text{-Q}}$) and fully smooth bounded regime ($\tau = 0$, i.e., $L_{\varepsilon\text{-BALS}}$); dashed black and red curves mark the Q and BALS losses at $\varepsilon = 0$. **Right**: the ε-HQ gradient is continuous and magnitude-dependent, with a near-origin jump preserving a minimal penalty for sensitive signals; the insensitive zone filters out minor fluctuations, stabilizing prediction.

3.2 Mathematical Properties

The proposed $L_{\varepsilon\text{-HQ}}$ loss possesses several favorable mathematical properties that are essential for designing efficient online learning algorithms and establishing rigorous theoretical guarantees.

Theorem 3.1 (Basic Properties)**.** *For any $z \in \mathbb{R}$, $L_{\varepsilon\text{-HQ}}(z)$ satisfies the following properties:*

1. ***Nonnegativity:*** *$L_{\varepsilon\text{-HQ}}(z) \ge 0$. If $\tau = 0$, then $0 \le L_{\varepsilon\text{-HQ}}(z) < 1/\gamma$; otherwise, the loss is unbounded above and exhibits linear growth.*
2. ***Asymmetry:*** *If $\alpha \ne 0.5$, there exists z such that $L_{\varepsilon\text{-HQ}}(z) \ne L_{\varepsilon\text{-HQ}}(-z)$.*
3. ***Insensitivity:*** *$L_{\varepsilon\text{-HQ}}(z) = 0$ if and only if $|z| \le \varepsilon$.*
4. ***Monotonicity:*** *$L_{\varepsilon\text{-HQ}}(z)$ is strictly decreasing on $(-\infty, -\varepsilon)$ and strictly increasing on $(\varepsilon, +\infty)$.*
5. ***Lipschitz continuity:*** *There exists a constant $M > 0$ such that $|L_{\varepsilon\text{-HQ}}(z_1) - L_{\varepsilon\text{-HQ}}(z_2)| \le M|z_1 - z_2|$ for all $z_1, z_2 \in \mathbb{R}$. A valid Lipschitz constant is*

$$M = \tau \max(\alpha, 1-\alpha) + (1-\tau)\frac{3\sqrt{3\lambda}}{8\gamma}\max(\sqrt{\alpha}, \sqrt{1-\alpha}). \tag{3}$$

Theorem 3.2 (Generalized Weak Pseudo-convexity)**.** *$L_{\varepsilon\text{-HQ}}(z)$ is* generalized weakly pseudo-convex *(GWPC) (Bazaraa et al., 2006) on $\mathbb{R}$: there exists $K > 0$ such that for any $z \in \mathbb{R}$ and global minimizer $z^* \in [-\varepsilon, \varepsilon]$, there exists a Clarke subgradient $g \in \partial L_{\varepsilon\text{-HQ}}(z)$ satisfying*

$$L_{\varepsilon\text{-HQ}}(z) - L_{\varepsilon\text{-HQ}}(z^*) \leq K \cdot \frac{g\,(z - z^*)}{\|g\|},$$

where the right-hand side is taken as 0 when $g = 0$. A feasible GWPC constant is

$$K = \tau \max(\alpha, 1-\alpha) + \frac{(1-\tau)\sqrt{\lambda}}{2\gamma} \max(\sqrt{\alpha}, \sqrt{1-\alpha}). \tag{4}$$

Theorem 3.1 elaborates the basic properties of the ε-HQ loss, among which Lipschitz continuity guarantees stable gradient updates and is essential for convergence analysis. Theorem 3.2 further proves GWPC, a condition weaker than convexity that still ensures a descent direction, thereby enabling sublinear dynamic regret bounds in nonconvex settings (Gao et al., 2018). Complete proofs of Theorems 3.1 and 3.2 are given in Appendices A.1 and A.2, respectively. Moreover, a systematic comparison with Q, ALS, BQ, and BALS is provided in Appendix A.3.

4 Dual-Calibrated Online Conformal Inference

Building on the ε-HQ loss, we propose the Dual-Calibrated Online Conformal Inference framework (ε-**HQ-DCOCI**). It consists of three sequential stages: a *gain-balancing mechanism* aligns the peak gains of the linear and bounded quadratic components; a *dynamic fine-tuning mechanism* adapts the residual scale and the insensitive zone online; and a *dual-calibration update* combines a base proximal subgradient step with a momentum-based correction. Together, these stages configure the loss, adapt it at runtime, and evolve the threshold for drift-aware, stable prediction sets.

4.1 Gain-Balancing Mechanism

The effectiveness of the ε-HQ loss hinges on matched gain magnitudes: a dominant linear component suppresses the quadratic component's continuous signal, whereas a dominant quadratic component causes updates to overreact. We therefore impose an equal-peak alignment constraint.

By Theorem 3.1, the linear component has constant gradient magnitude τA and the bounded quadratic component attains its peak gradient magnitude $(1-\tau)\frac{3\sqrt{3\lambda}}{8\gamma}B$, where $A = \max(\alpha, 1-\alpha)$ and $B = \max(\sqrt{\alpha}, \sqrt{1-\alpha})$. Equal-peak alignment requires $\tau A = (1-\tau)\frac{3\sqrt{3\lambda}}{8\gamma}B$, which prevents either gain from dominating. Setting $\gamma = 1$ without loss of generality (fixing λ instead would drive $\gamma(\tau) \to 0$ as $\tau \to 1$, causing the bounded quadratic gradient to diverge) yields

$$\lambda(\tau) = \left(\frac{8A}{3B} \cdot \frac{\tau}{1-\tau}\right)^2, \tag{5}$$

reducing the parameter space (τ, λ, γ) to a single mixing coefficient $\tau \in (0, 1)$. Under this constraint, the Lipschitz constant, gradient lower bound, and GWPC constant reduce to

$$M = 2\tau A, \qquad \mu = \tau \min(\alpha, 1-\alpha), \qquad K = \tau A \left(1 + \frac{4}{3\sqrt{3}}\right). \tag{6}$$

Here μ provides the gradient lower bound from the linear component, used for the truncation criterion in Section 4.2; M supports the bounded-perturbation analysis of momentum step in Section 4.3.

4.2 Dynamic Fine-Tuning Mechanism

With gain balancing in place, the online performance of the ε-HQ loss depends on whether residuals fall within a gradient-sensitive range and whether the insensitive zone matches the current noise level. Large residuals saturate the bounded quadratic component, weakening the correction of large deviations, while a fixed insensitive zone cannot balance sensitivity with robustness. To address this, we introduce two coupled mechanisms: *adaptive residual mapping* affinely rescales raw residuals into $[Z_{\min}, Z_{\max}]$ to prevent saturation, and *insensitive-zone tuning* adjusts ε_t based on the mapped residual distribution and recent gradient-direction consistency. Both mechanisms operate on the same update schedule, preserving the loss's gradient responsiveness under non-stationarity.

Adaptive Residual Mapping We maintain a sliding window $\mathcal{H}_r = \{z_{t-W}, \dots, z_{t-1}\}$ of the W most recent raw residuals, partitioned into $\mathcal{H}_r^- = \{z \in \mathcal{H}_r : z < 0\}$ and $\mathcal{H}_r^+ = \{z \in \mathcal{H}_r : z > 0\}$. The endpoints $Z_{\min} < 0 < Z_{\max}$ are derived analytically from the gradient decay of the ε-HQ loss (Appendix A.4). For $t < W$, $z'_t = \mathrm{clip}(z_t, Z_{\min}, Z_{\max})$; for $t \geq W$, the mapping is sign-dependent.

For each sign $s \in \{-, +\}$, define $z^s_{\min} = \min(\mathcal{H}_r^s \cup \{0\})$ and $z^s_{\max} = \max(\mathcal{H}_r^s \cup \{0\})$. When $\mathcal{H}_r^s \neq \emptyset$ and $z^s_{\max} - z^s_{\min} \geq \delta$ ($\delta = 10^{-6}$), the mapping ϕ_t is affine, $z'_t = \phi_t(z_t) = a^s_t z_t + b^s_t$, with

$$a^s_t = \frac{\Delta^s}{z^s_{\max} - z^s_{\min}}, \quad b^s_t = o^s - a^s_t z^s_{\min}, \quad (\Delta^s, o^s) = \begin{cases} (-Z_{\min},\, Z_{\min}), & s = -, \\ (Z_{\max},\, 0), & s = +. \end{cases} \tag{7}$$

The slope satisfies $a^s_t \in [a_{\min}, a_{\max}]$ with $a_{\min} > 0$. Otherwise, z'_t is projected onto $[Z_{\min}, 0]$ or $[0, Z_{\max}]$ by the sign of z_t. This mapping confines residuals to $[Z_{\min}, Z_{\max}]$, preventing gradient saturation; its slope increases as the spread narrows, avoiding over-compression in calm periods.

Insensitive-Zone Tuning We also maintain a window $\mathcal{H}_n = \{z'_{t-W}, \dots, z'_{t-1}\}$ of normalized residuals. During warm-up ($t < W$), we set $\varepsilon_t = \varepsilon_{\min}$ with $\varepsilon_{\min} = 10^{-6}$. For $t \geq W$, define $\mathcal{A}^- = \{|z'| : z' \in \mathcal{H}_n,\, z' < 0\} \cup \{0\}$ and $\mathcal{A}^+ = \{z' \in \mathcal{H}_n : z' > 0\} \cup \{0\}$, and compute the Tukey lower fences (Hoaglin et al., 1986) for $\mathcal{A}^-$ and $\mathcal{A}^+$, respectively:

$$d_t^- = \max\big(0,\, Q_1^- - k \cdot \mathrm{IQR}^-\big), \qquad d_t^+ = \max\big(0,\, Q_1^+ - k \cdot \mathrm{IQR}^+\big), \tag{8}$$

where $Q_1^\pm, Q_3^\pm$ are the first and third quartiles of $\mathcal{A}^\pm$, $\mathrm{IQR}^\pm = Q_3^\pm - Q_1^\pm$, k is the Tukey fence multiplier (default $k = 1.5$), and $d_t = \min(d_t^-, d_t^+)$ is a robust estimate of the current noise level.

To incorporate drift perception, let $\{g_i\}_{i=1}^t$ be the historical subgradients at the thresholds $\hat{q}_i$, and define the direction-consistency norm

$$\nu_t = \frac{1}{W} \sum_{i=t-W}^{t-1} \frac{g_i}{\|g_i\|}, \tag{9}$$

with $g_i = 0$ treated as zero vectors and, for $t < W$, the denominator set to the actual number of terms. By the triangle inequality, $0 \leq \|\nu_t\| \leq 1$. At time t, the insensitive-zone width is defined as

$$\varepsilon_t = \max\big(d_t \cdot (1 - \|\nu_t\|),\, \varepsilon_{\min}\big). \tag{10}$$

As $\|\nu_t\| \to 1$, ε_t shrinks to $\varepsilon_{\min}$, preserving drift sensitivity; as $\|\nu_t\| \to 0$, d_t dominates, expanding under noise and shrinking otherwise. This distribution-free mechanism is robust to skewness and heavy tails; adding 0 in each $\mathcal{A}^\pm$ ensures well-defined quartiles and zero factors for empty sets.

4.3 Threshold Dual-Calibration Update

Base Calibration At time t, define the scaled loss $\tilde{\ell}_t(q) = (1/a^s_t) \cdot L_{\varepsilon\text{-HQ}}\big(\phi_t(s_t - q); \alpha, \varepsilon_t, \tau, \gamma, \lambda\big)$, where ϕ_t is the sign-preserving residual mapping in (7) with slope $a^s_t > 0$, s_t is the nonconformity score and ε_t is given by (10). The base calibration performs online proximal subgradient descent:

$$\hat{q}_{t+1} = \arg\min_{q \in \mathcal{Q}} \Big\{ \tilde{\ell}_t(\hat{q}_t) + \langle g_t, q - \hat{q}_t \rangle + \frac{1}{2\eta} \|q - \hat{q}_t\|^2 \Big\}, \tag{11}$$

where $g_t = -\frac{\partial L_{\varepsilon\text{-HQ}}}{\partial z}(z'_t)$ is the subgradient at $\hat{q}_t$, $\eta > 0$ the base learning rate, $\mathcal{Q} = [0, R]$ the feasible set; explicit forms are given in Section 3.2 and Appendix A.1. The closed-form solution is

$$\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta g_t). \tag{12}$$

Momentum Calibration To accelerate threshold updates under persistent drift, the momentum calibration uses the direction-consistency norm ν_t from (9) with its window shifted to include the current gradient g_t, i.e., $\nu_{t+1} = \frac{1}{W} \sum_{i=t-W+1}^{t} \frac{g_i}{\|g_i\|}$, and solves the proximal problem

$$q_{t+1} = \arg\min_{q \in \mathcal{Q}} \Big\{ \langle \nu_{t+1}, q \rangle + \frac{1}{2\eta'} \|q - \hat{q}_{t+1}\|^2 \Big\}, \tag{13}$$

with momentum learning rate $\eta' > 0$. Strong convexity yields the closed-form solution

$$q_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_{t+1} - \eta' \nu_{t+1}). \tag{14}$$

Since $\|\nu_{t+1}\| \leq 1$, the momentum step is a bounded perturbation of the base threshold. Algorithm 1 in Appendix C.1 details the full procedure, which integrates the ε-HQ gradient and momentum calibration as a history-smoothing module to overcome the three bottlenecks. Variants DCOGD, ε-HQ-OGD, and ε-HQ-DCOGD are also investigated, with other extensions provided in Appendix C.2.

5 THEORETICAL ANALYSIS

We establish three levels of guarantees for ε-**HQ-DCOCI**: (i) base dynamic regret (Theorem 5.1), (ii) joint regret showing that momentum preserves the regret order (Theorem 5.2), and (iii) coverage validity linking optimization to statistical coverage (Theorem 5.3, Corollary 5.4). Let $\mathcal{Q} = [0, R]$ with projection $\Pi_{\mathcal{Q}}$, and define the instantaneous loss $\ell_t(q) = L_{\varepsilon\text{-HQ}}(s_t - q; \alpha, \varepsilon_t, \tau, \gamma, \lambda)$, instantaneous optimum $q_t^{\text{opt}} \in \arg\min_{q\in\mathcal{Q}} \ell_t(q)$, and path length $V_T = \sum_{t=1}^{T-1} |q_t^{\text{opt}} - q_{t+1}^{\text{opt}}|$. The conditional $(1-\alpha)$-quantile q_t^{qt} serves as the statistical comparator. Throughout, we work under the gain-balancing configuration of Section 4.1, under which ℓ_t is M-Lipschitz and K-GWPC with constants given by Eq. (6). Algorithm 1 optimizes the scaled loss $\tilde{\ell}_t$ defined in Section 4.3, which shares the same minimizers and inherits the M-Lipschitz, K-GWPC constants of ℓ_t, and is optimized with the constant learning rate η; hence the theorems below are stated for $\tilde{\ell}_t$ and hold for the implemented algorithm (see Appendix B.1).

5.1 DYNAMIC REGRET ANALYSIS

We first state a gradient lower bound that is naturally satisfied by the linear component of ε-HQ loss.

Assumption 5.1 (Gradient Lower Bound)**.** *There exists $\mu > 0$ such that, whenever $\tilde{\ell}_t(q) > \tilde{\ell}_t(q_t^{\text{opt}})$, any subgradient $g \in \partial\tilde{\ell}_t(q)$ satisfies $\|g\| \geq \mu$. In particular, $\mu = \tau\min(\alpha, 1-\alpha)$ is strictly positive.*

Theorem 5.1 (Base Dynamic Regret)**.** *Under Assumption 5.1 and with $\tilde{\ell}_t$ being M-Lipschitz and K-GWPC, choose $\eta = \sqrt{\frac{4R^2+3RV_T}{TM^2}}$, the intermediate thresholds $\{\hat{q}_t\}$ from base calibration satisfy*

$$\sum_{t=1}^{T}\big[\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\text{opt}})\big] \leq \frac{KM}{\mu}\sqrt{T(4R^2+3RV_T)} = \mathcal{O}\big(\sqrt{T(1+V_T)}\big). \tag{15}$$

Theorem 5.1 provides an $\mathcal{O}(\sqrt{T(1+V_T)})$ dynamic regret bound for the base calibration, extending the standard optimal rate for dynamic online convex optimization to the GWPC non-convex setting.

Theorem 5.2 (Joint Dynamic Regret)**.** *Under the assumptions of Theorem 5.1, choose $\eta' = \frac{K}{\mu}\sqrt{\frac{4R^2+3RV_T}{T}}$. Then the final thresholds $\{q_t\}$ of ε-**HQ-DCOCI** satisfy*

$$\tilde{R}_T^{\text{dyn}} = \sum_{t=1}^{T}\big[\tilde{\ell}_t(q_t) - \tilde{\ell}_t(q_t^{\text{opt}})\big] \leq \frac{2KM}{\mu}\sqrt{T(4R^2+3RV_T)} = \mathcal{O}\big(\sqrt{T(1+V_T)}\big). \tag{16}$$

Theorem 5.2 demonstrates that momentum calibration preserves the regret order, since the bounded perturbation $\|\nu_{t+1}\| \leq 1$ adds at most $\eta' MT$, which is $\mathcal{O}(\sqrt{T})$ when $\eta' = \mathcal{O}(1/\sqrt{T})$. The complete proofs of Theorems 5.1 and 5.2 are given in Appendices B.2 and B.3, respectively.

5.2 COVERAGE VALIDITY ANALYSIS

To connect optimization regret to statistical coverage, we introduce the following standard regularity condition, which holds for any non-degenerate score distribution with bounded support.

Assumption 5.2 (Bounded Conditional Density)**.** *There exists $C > 0$ such that $\sup_{t,s} p_t(s \mid \mathcal{F}_{t-1}) \leq C$, where $p_t(\cdot \mid \mathcal{F}_{t-1})$ is the conditional density of s_t.*

Theorem 5.3 (Coverage Error Bound)**.** *Under Assumption 5.2, let q_t^{qt} be the conditional $(1-\alpha)$-quantile of s_t, i.e., $\mathbb{P}(s_t \leq q_t^{\text{qt}} \mid \mathcal{F}_{t-1}) = 1-\alpha$. With probability at least $1-\delta$, the empirical coverage $\hat{C}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\{Y_t \in \hat{C}_t\}$ of ε-**HQ-DCOCI** satisfies*

$$\big|\hat{C}_T - (1-\alpha)\big| \leq \frac{C}{\mu}\cdot\Big(\frac{\tilde{R}_T^{\text{dyn}}}{T} + \frac{2\mu}{a_{\min}}\bar{\varepsilon}_T + \Delta_\infty\Big) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{T}}\right), \tag{17}$$

where $\tilde{R}_T^{\text{dyn}}$ is the regret bound from Theorem 5.2, $a_{\min} > 0$ is the lower bound of the slope a_t^s in Eq. (7), $\bar{\varepsilon}_T = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t$, and $\Delta_\infty = \frac{CMZ_{\text{abs}}^2}{2a_{\min}^2} + \frac{1-\tau}{a_{\min}}$ with $Z_{\text{abs}} = \max(|Z_{\min}|, Z_{\max})$. The $\mathcal{O}(\cdot)$ term hides a constant depending only on $(\tau, \alpha, C, Z_{\text{abs}})$.

Corollary 5.4 (Asymptotic Coverage Bias). *Under Theorem 5.3, as $T \to \infty$ with $V_T = O(T)$,*

$$\limsup_{T\to\infty} \left|\hat{C}_T - (1-\alpha)\right| \leq \frac{2C}{a_{\min}}\varepsilon_\infty + \frac{C}{\mu}\Delta_\infty, \tag{18}$$

where $\varepsilon_\infty = \limsup_{T\to\infty} \bar{\varepsilon}_T$. The first term is controlled by the adaptive insensitive zone, reducing to $\varepsilon_{\min}$ (arbitrarily small) under persistent drift with $\|\nu_t\| \to 1$. The second term is an intrinsic bias of the ε-HQ loss; since $M = 2\tau A$ and $\mu = \tau \min(\alpha, 1-\alpha)$, both Δ_∞ and C/μ depend on τ, and their product is jointly minimized over $\tau \in (0,1)$. Its value is typically small for small C and Z_{abs}.

Theorem 5.3 provides a finite-sample decomposition of coverage error: the regret term vanishes as $T \to \infty$; the insensitive-zone bias $\frac{2C}{a_{\min}}\bar{\varepsilon}_T$ is adaptively regulated; and the intrinsic bias $\frac{C}{\mu}\Delta_\infty$ remains controllable through τ. Corollary 5.4 thus establishes *approximately valid* coverage with both bias sources adjustable. Complete proofs appear in Appendices B.4 and B.5.

6 Experimental Analysis

We evaluate ε-**HQ-DCOCI** on time series prediction and medical image classification against nine baselines with $\alpha = 0.1$: ACI (Gibbs & Candès, 2021), DtACI (Gibbs & Candès, 2024), OGD, SF-OGD (Bhatnagar et al., 2023), Decay-OGD (Angelopoulos et al., 2024), PIDCP (Angelopoulos et al., 2023), ECI (Wu et al., 2025), LQT (Areces et al., 2025), and COP (Hu et al., 2026). Following (Angelopoulos et al., 2023), the Tanh function and Theta are used as integrator and scorecaster. We report three metrics: (1) Cov (coverage); (2) PI-Width (regression) / PI-Size (classification) for efficiency; (3) Mean Interval Score (MIS) (Gneiting & Raftery, 2007; Allen et al., 2025), penalizing coverage deviation and interval width. For medical image classification we additionally report the Mean DSC (Dice Similarity Coefficient) (Dice, 1945; Taha & Hanbury, 2015), denoted as MDSC. All experiments are run in Python 3.10 on a workstation (Intel Core i7-8700, 16 GB RAM, NVIDIA Quadro P400). A sliding window $W = 100$ is used for updating, prediction, and conformal inference; reported results are averages over 5 runs. Detailed settings appear in Appendix D.1.

6.1 Financial Time Series Prediction

Datasets and Models. We use daily opening prices of four global assets—Bitcoin, U.S. Dollar Index Futures (DX), Nasdaq Index (IXIC), and Gold Spot (XAU) (Zhang et al., 2022)—from January 1, 2014 to October 31, 2025. They cover cryptocurrencies, FX, equities, and commodities, exhibiting pronounced distribution shifts ideal for stress-testing adaptability. Point predictors are LightGBM (Ke et al., 2017), Theta (Assimakopoulos & Nikolopoulos, 2000), and Prophet (Taylor & Letham, 2018). The nonconformity score is the p-order normalized residual (Cordier et al., 2023) with $p = 1$ (reduces to absolute residual when $p = 0$). Data and model details are in Appendix D.2.

Results and Analysis. Table 1 reports LightGBM results for 12 OCI algorithms; Theta and Prophet results (Appendix D.4) are consistent. The proposed methods substantially outperform baselines in MIS. Specifically, ε-HQ-OGD leads on Bitcoin, IXIC, and XAU, while ε-HQ-DCOGD leads on DX (second on Bitcoin and XAU), confirming that the ε-HQ loss improves coverage, with momentum adding further gains. Among baselines, OGD ranks second on IXIC, whereas LQT attains second-best MIS only on DX and degrades under moderate-to-high volatility. Across all metrics, both methods yield narrower intervals (lower PI-Width) on Bitcoin and DX and coverage closer to 90% on IXIC and XAU, demonstrating a joint trade-off rather than single-metric optimization.

Ablation analysis. Table 18 reports the MIS of ε-HQ-DCOGD under 10 configurations of mixing coefficient τ, insensitive-zone width ε, and gradient decay factor Γ. The full $L_{\varepsilon\text{-HQ}}$ (mapping + tuning) achieves top-two MIS in 10/12 cases (3 best, 7 second), whereas the mapping-only variant ($\varepsilon = 0$) attains 8/12 (5 best, 3 second). This aligns with Corollary 5.4: $\varepsilon = 0$ removes the insensitive-zone bias, reducing coverage error and yielding more best-case results. Nevertheless, the full algorithm secures more top-two rankings overall, indicating a superior accuracy–stability trade-off. The tuning-only variant obtains 2 second-best, while L_{Q} achieves 1 second-best and $L_{\varepsilon\text{-Q}}$ achieves 3 best; smooth losses lag. Thus, dynamic adjustment mechanisms, not merely continuous gradients, drive the coverage gains. **Sensitivity analysis** (Appendix E.3) further evaluates $\Delta\text{Cov} = |\text{Cov} - (1-\alpha)|$ and MIS with respect to τ (Tables 20), decay factor Γ (Tables 21), W (Tables 22), and Tukey multiplier k (Tables 23): smaller τ and Γ improve coverage ($\tau = 0.3$, $\Gamma = 0.1$ for smaller ΔCov; $\tau = 0.1$, $\Gamma = 0.2$ for smaller MIS), while W and k are largely insensitive.

Table 1: Performance comparison of 12 OCI algorithms on four financial time series datasets with LightGBM. The **best** and second-best results for the MIS metric are bolded and underlined in red.

Algorithm	Bitcoin			DX			IXIC			XAU		
	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
ACI	87.467	2521.977	3580.518	89.618	1.415	1.863	88.282	444.405	627.866	87.441	50.820	76.760
DtACI	88.818	2475.914	3565.269	88.265	1.352	1.842	88.282	435.600	628.180	88.012	51.599	75.610
OGD	89.647	2431.156	3512.292	89.550	1.375	1.823	89.291	427.392	624.772	88.684	50.503	74.136
SF-OGD	89.766	2450.432	3522.821	89.686	1.371	1.840	89.465	429.062	628.796	88.516	49.000	74.985
Decay-OGD	89.860	2433.238	3522.170	89.888	1.373	1.845	89.569	434.387	632.212	88.885	49.903	74.595
PIDCP	94.030	2808.332	3984.351	89.956	1.535	2.140	91.551	507.286	741.140	87.979	51.955	87.130
ECI	86.283	2433.251	3553.253	88.468	1.351	1.832	86.579	428.416	632.712	85.359	44.973	75.432
LQT	43.449	465.852	8272.316	89.449	1.363	1.823	63.804	186.570	962.323	88.482	49.703	75.169
COP	87.894	2386.968	3535.199	87.826	1.331	1.834	88.491	424.527	625.898	88.382	50.793	74.292
DCOGD	89.718	2405.738	3513.560	86.879	1.278	1.830	89.326	435.338	624.854	88.079	49.383	74.207
ε-HQ-OGD	89.078	2368.859	**3510.987**	89.280	1.352	1.824	90.021	444.509	**624.004**	89.490	52.211	**73.633**
ε-HQ-DCOGD	89.173	2377.241	3512.139	88.739	1.341	**1.822**	90.230	447.119	624.794	88.919	50.076	73.937

Table 2: Performance comparison of 12 OCI algorithms on two medical image datasets with ResNet-50. The **best** and second-best results for the MDSC metric are bolded and underlined in red.

Algorithm	DERM12345			MRI		
	Cov(%)	PI-Size ↓	MDSC ↑	Cov(%)	PI-Size ↓	MDSC ↑
ACI	89.839 ± 0.347	2.308 ± 0.020	0.663 ± 0.002	89.909 ± 0.314	1.039 ± 0.004	0.860 ± 0.003
DtACI	89.629 ± 0.401	2.214 ± 0.021	0.668 ± 0.003	89.743 ± 0.348	1.037 ± 0.005	0.858 ± 0.003
OGD	89.760 ± 0.314	2.315 ± 0.041	0.661 ± 0.001	89.811 ± 0.322	1.038 ± 0.004	0.859 ± 0.003
SF-OGD	89.666 ± 0.430	2.147 ± 0.029	0.667 ± 0.002	89.872 ± 0.261	1.038 ± 0.003	0.859 ± 0.002
Decay-OGD	89.785 ± 0.400	2.198 ± 0.035	0.666 ± 0.001	89.895 ± 0.248	1.039 ± 0.003	0.859 ± 0.002
PIDCP	76.213 ± 0.516	1.821 ± 0.023	0.627 ± 0.004	76.274 ± 0.334	1.081 ± 0.018	0.724 ± 0.003
ECI	87.562 ± 0.492	2.154 ± 0.034	0.666 ± 0.003	88.480 ± 0.315	1.018 ± 0.004	0.848 ± 0.003
LQT	89.201 ± 0.444	2.231 ± 0.030	0.660 ± 0.002	89.721 ± 0.322	1.081 ± 0.013	0.854 ± 0.004
COP	89.604 ± 0.402	2.255 ± 0.026	0.664 ± 0.002	89.750 ± 0.285	1.037 ± 0.004	0.858 ± 0.002
DCOGD	87.427 ± 0.379	2.041 ± 0.019	0.671 ± 0.003	89.632 ± 0.324	1.047 ± 0.005	0.856 ± 0.003
ε-HQ-OGD	89.613 ± 0.471	2.272 ± 0.030	0.663 ± 0.002	92.489 ± 0.183	1.083 ± 0.002	$\mathbf{0.879 \pm 0.002}$
ε-HQ-DCOGD	87.271 ± 0.328	2.003 ± 0.027	$\mathbf{0.673 \pm 0.002}$	91.977 ± 0.233	1.074 ± 0.003	0.876 ± 0.002

6.2 Medical Image Classification

Datasets and Models. We use two medical image datasets: DERM12345 (Yilmaz et al., 2024) (12,345 dermoscopic images, 40 subcategories) and a benchmark diagnostic MRI dataset (Ajmain et al., 2024) (34,192 brain, neurological, and spinal MRI images, 40 disease classes). Their scale and imbalance facilitate drift simulation. Predictors are fine-tuned ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), and ResNeXt-50 (Xie et al., 2017). We use RAPS (Angelopoulos et al., 2021) (parameters tuned via grid search on validation). Stratified sampling splits data into train (calibration) and test sets; rolling prediction across disease classes simulates streams. Detailed settings are in Appendix D.3. RAPS has a random term, and thus we report mean $\pm$ std. over 5 runs.

Results and Analysis. Table 2 reports results with ResNet-50; DenseNet-121 and ResNeXt-50 results are provided in Appendix D.5 and are consistent. On DERM12345, ε-HQ-DCOGD achieves the best MDSC and DCOGD the second-best; on MRI, ε-HQ-OGD attains the best and ε-HQ-DCOGD the second-best, validating the ε-HQ loss and momentum calibration combined in ε-HQ-DCOGD. PIDCP obtains high efficiency on DERM12345 but at the cost of accuracy, whereas our methods achieve a superior accuracy–efficiency trade-off with smaller prediction sets than baselines.

7 Conclusion

We propose the ε-HQ loss (Lipschitz continuous, GWPC) and ε-**HQ-DCOCI** to remedy three Q-loss OCI bottlenecks: *coarse binary feedback*, *unattenuated outlier response*, and *drift ignorance*. A hybrid linear–bounded structure with adaptive mapping and insensitive-zone tuning (first calibration) yields fine-grained feedback and robustness; momentum-smoothed historical gradients (second calibration) track drift. The framework guarantees sublinear dynamic regret and approximately valid coverage under non-stationarity, with plug-in modules for other OCI methods. Experiments confirm a superior accuracy–efficiency trade-off over baselines; ablation verifies each component. Future work extends to multivariate forecasting and missing-data decision-making in high-stakes settings.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.) Generative AI tools were used solely for manuscript polishing, grammar refinement, and formula formatting. They were not used for research ideation, algorithm design, theoretical proof derivation, or experimental analysis. No new human-subject research was conducted. All AI-assisted content was reviewed, revised, and verified by the authors, who take full responsibility for the final text, technical claims, and experimental results.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.) This work focuses on online conformal prediction algorithm design and empirical evaluation using publicly available, de-identified benchmark datasets. No new human subjects were recruited, and no private or sensitive information was collected. The experiments pose no potential discrimination, bias, privacy risks, or harmful societal impacts. The authors declare no conflicts of interest or sponsorship-related issues. All experiments comply with standard research ethics and legal requirements.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.) All algorithm descriptions and pseudocode are provided in the main text and appendix. Complete theoretical proofs, hyperparameter configurations, and experimental details—including data-processing pipelines—are provided in the appendix. The experimental results can be fully reproduced using the provided pseudocode and documented settings. Source code will be released upon paper acceptance.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. doi: 10.1016/j.inffus.2021.05.008.

Moshfiqur Rahman Ajmain, Mst. Farhana Khatun, and Sharun Akter Khushbu. Benchmark diagnostic MRI and medical imaging dataset. Mendeley Data, V1, 2024.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330701.

Sam Allen, Julia Burnello, and Johanna Ziegel. Assessing the conditional calibration of interval forecasts using decompositions of the interval score. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2508.18034.

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021)*. OpenReview, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23047–23074, 2023. URL https://dl.acm.org/doi/10.5555/3666122.3667122.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1616–1630. PMLR, 2024. URL https://proceedings.mlr.press/v235/angelopoulos24a.html.

Felipe Areces, Christopher Mohri, Tatsunori Hashimoto, and John Duchi. Online conformal prediction via online optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 1604–1649. PMLR, 2025. URL https://proceedings.mlr.press/v267/areces25a.html.

V. Assimakopoulos and K. Nikolopoulos. The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16(4):521–530, 2000. doi: 10.1016/S0169-2070(00)00066-2.

Mokhtar S. Bazaraa, Hanif D. Sherali, and Chitharanjan M. Shetty. *Nonlinear programming: Theory and algorithms*. Wiley, 3rd edition, 2006.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2337–2363. PMLR, 2023. URL https://proceedings.mlr.press/v202/bhatnagar23a.html.

Tapabrata Chakraborti, Christopher R. S. Banerji, Ariane Marandon, Vicky Hellon, Robin Mitra, Brieuc Lehmann, Leandra Bräuninger, Sarah McGough, Cagatay Turkay, Alejandro F. Frangi, Ginestra Bianconi, Wei Li, Owen Rackham, Deeksha Parashar, Christopher Harbron, and Ben MacArthur. Personalized uncertainty quantification in artificial intelligence. *Nature Machine Intelligence*, 7:522–530, 2025. doi: 10.1038/s42256-025-01024-8.

Thibault Cordier, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, and Nicolas Brunel. Flexible and systematic uncertainty estimation with conformal prediction via the MAPIE library. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pp. 549–581. PMLR, 2023. URL https://proceedings.mlr.press/v204/cordier23a.html.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL http://jmlr.org/papers/v7/demsar06a.html.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. doi: 10.2307/1932409.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, Cambridge, UK, 5th edition, 2019.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522.

Saiji Fu, Yingjie Tian, and Long Tang. Robust regression under the general framework of bounded loss functions. *European Journal of Operational Research*, 310(3):1325–1339, 2023. doi: 10.1016/j.ejor.2023.04.025.

Xiang Gao, Xiaobo Li, and Shuzhong Zhang. Online learning with non-convex losses and non-stationary regret. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 235–243. PMLR, 2018. URL https://proceedings.mlr.press/v84/gao18a.html.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024. URL https://jmlr.org/papers/v25/22-1218.html.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021. URL https://dl.acm.org/doi/10.5555/3540261.3540389.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90.

David C. Hoaglin, Boris Iglewicz, and John W. Tukey. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991–999, 1986. doi: 10.1080/01621459.1986.10478363.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.2307/2282952.

Dongjian Hu, Junxi Wu, Shu-Tao Xia, and Changliang Zou. Distribution-informed online conformal prediction. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026. URL https://openreview.net/forum?id=I69SaLbwqZ.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. IEEE, 2017. doi: 10.1109/CVPR.2017.243.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154, 2017. URL https://dl.acm.org/doi/10.5555/3294996.3295074.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 6402–6413. Curran Associates, Inc., 2017. URL https://dl.acm.org/doi/10.5555/3295222.3295387.

Yonghoon Lee and Rina F. Barber. Distribution-free inference for regression: discrete, continuous, and in between. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7448–7459, 2021. URL https://dlnext.acm.org/doi/10.5555/3540261.3540831.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Martin Magris and Alexandros Iosifidis. Bayesian learning for neural networks: An algorithmic survey. *Artificial Intelligence Review*, 56:11773–11823, 2023. doi: 10.1007/s10462-023-10443-1.

Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman (eds.), *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Whitney K. Newey and James L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847, 1987. doi: 10.2307/1911031.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, volume 37, pp. 8026–8037, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553, 2019. URL https://dl.acm.org/doi/10.5555/3454287.3454605.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pp. 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

Xin Shen, Lingfeng Niu, Zhiquan Qi, and Yingjie Tian. Support vector machine classifier with truncated pinball loss. *Pattern Recognition*, 68:199–210, 2017. doi: 10.1016/j.patcog.2017.03.011.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15:29, 2015. doi: 10.1186/s12880-015-0068-x.

Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi: 10.1080/00031305.2017.1380080.

Yibo Wang, Wenhao Yang, Wei Jiang, Shiyin Lu, Binghui Wang, Haiyang Tang, Yuanyu Wan, and Lijun Zhang. Non-stationary projection-free online learning with dynamic and adaptive regret guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15671–15679, 2024. doi: 10.1609/aaai.v38i14.29495.

Daling Wen and Martin S. Levy. BLINEX: A bounded asymmetric loss function with application to Bayesian estimation. *Communications in Statistics - Theory and Methods*, 30(1):147–153, 2001. doi: 10.1081/STA-100001564.

Junxi Wu, Dongjian Hu, Yajie Bao, Shu-Tao Xia, and Changliang Zou. Error-quantified conformal inference for time series. In *International Conference on Learning Representations*, pp. 81566–81593, 2025. URL https://openreview.net/forum?id=RD9q5vEe1Q.

Qing Wu, Heng-Chang Zhang, and Yi-Jui Chiu. Application of asymmetric proximal support vector regression based on multitask learning in the stock market. *Expert Systems with Applications*, 227:120208, 2023. doi: 10.1016/j.eswa.2023.120208.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500. IEEE, 2017. doi: 10.1109/CVPR.2017.634.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38707–38727. PMLR, 2023. URL https://proceedings.mlr.press/v202/xu23r.

Zhipan Xu and Lijun Zhang. Online non-convex learning in dynamic environments. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, volume 37, pp. 51930–51962, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d1233f819202ade06023346df80a6d2-Abstract-Conference.html.

Abdurrahim Yilmaz, Sirin Pekcan Yasar, Gulsum Gencoglan, and Burak Temelkuran. DERM12345: A large, multisource dermatoscopic skin lesion dataset with 40 subclasses. *Scientific Data*, 11:1302, 2024. doi: 10.1038/s41597-024-04104-3.

Heng-Chang Zhang, Qing Wu, and Fei-Yan Li. Application of online multitask learning based on least squares support vector regression in the financial market. *Applied Soft Computing*, 121: 108754, 2022. doi: 10.1016/j.asoc.2022.108754.

Jiaqi Zhang and Hu Yang. Bounded quantile loss for robust support vector machines-based classification and regression. *Expert Systems with Applications*, 242:122759, 2024. doi: 10.1016/j.eswa.2023.122759.

Jiaqi Zhang and Hu Yang. Robust support vector machine based on the bounded asymmetric least squares loss function and its applications in noise corrupted data. *Advanced Engineering Informatics*, 65:103371, 2025. doi: 10.1016/j.aei.2025.103371.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. Conformal prediction: A data perspective. *ACM Computing Surveys*, 58(2):49:1–49:37, 2025. doi: 10.1145/3736575.

A PROOFS AND TECHNICAL DETAILS FOR THE ε-HQ LOSS

This appendix provides the complete proofs of Theorems 3.1 and 3.2 (Appendices A.1 and A.2), a systematic comparison of loss function properties (Appendix A.3), and the analytical determination of the target interval $[Z_{\min}, Z_{\max}]$ required by the adaptive residual mapping mechanism (Appendix A.4).

A.1 PROOF OF THEOREM 3.1

Theorem A.1 (Theorem 3.1 restated. Basic Properties)**.** *For any $z \in \mathbb{R}$, $L_{\varepsilon\text{-}HQ}(z)$ satisfies the following properties:*

1. ***Nonnegativity:*** *$L_{\varepsilon\text{-}HQ}(z) \geq 0$. If $\tau = 0$, then $0 \leq L_{\varepsilon\text{-}HQ}(z) < 1/\gamma$; otherwise, the loss is unbounded above and exhibits linear growth.*
2. ***Asymmetry:*** *If $\alpha \neq 0.5$, there exists z such that $L_{\varepsilon\text{-}HQ}(z) \neq L_{\varepsilon\text{-}HQ}(-z)$.*
3. ***Insensitivity:*** *$L_{\varepsilon\text{-}HQ}(z) = 0$ if and only if $|z| \leq \varepsilon$.*
4. ***Monotonicity:*** *$L_{\varepsilon\text{-}HQ}(z)$ is strictly decreasing on $(-\infty, -\varepsilon)$ and strictly increasing on $(\varepsilon, +\infty)$.*
5. ***Lipschitz continuity:*** *There exists a constant $M > 0$ such that $|L_{\varepsilon\text{-}HQ}(z_1) - L_{\varepsilon\text{-}HQ}(z_2)| \leq M|z_1 - z_2|$ for all $z_1, z_2 \in \mathbb{R}$. A valid Lipschitz constant is*

$$M = \tau \max(\alpha, 1-\alpha) + (1-\tau)\frac{3\sqrt{3\lambda}}{8\gamma}\max(\sqrt{\alpha}, \sqrt{1-\alpha}). \tag{19}$$

Proof. For brevity, write $L(z) = L_{\varepsilon\text{-HQ}}(z)$. Recall that $L(z)$ is the convex combination $L(z) = \tau L_{\varepsilon\text{-Q}}(z) + (1-\tau)L_{\varepsilon\text{-BALS}}(z)$, with $L_{\varepsilon\text{-BALS}}(z) = \frac{1}{\gamma}\big[1 - \frac{1}{1+\lambda L_{\varepsilon\text{-ALS}}(z)}\big]$, where the base losses $L_{\varepsilon\text{-Q}}$ and $L_{\varepsilon\text{-ALS}}$ are defined in Section 3.1. We prove the five properties in order.

1. Nonnegativity. Since $L_{\varepsilon\text{-Q}}(z) \geq 0$, $L_{\varepsilon\text{-ALS}}(z) \geq 0$, and $\lambda > 0$, we have $1 + \lambda L_{\varepsilon\text{-ALS}}(z) \geq 1$, from which

$$0 \leq 1 - \frac{1}{1+\lambda L_{\varepsilon\text{-ALS}}(z)} < 1,$$

and consequently $0 \leq L_{\varepsilon\text{-BALS}}(z) < 1/\gamma$. Thus $L(z) = \tau L_{\varepsilon\text{-Q}}(z) + (1-\tau)L_{\varepsilon\text{-BALS}}(z) \geq 0$. If $\tau = 0$, $L(z) = L_{\varepsilon\text{-BALS}}(z) \in [0, 1/\gamma)$. If $\tau > 0$, the term $L_{\varepsilon\text{-Q}}(z)$ grows linearly as $|z| \to \infty$ (specifically, it behaves like $\max(\alpha, 1-\alpha)|z|$), making $L(z)$ unbounded above.

2. Asymmetry. Take $z = c > \varepsilon$. From the definitions, we have

$$L_{\varepsilon\text{-Q}}(c) = (1-\alpha)(c-\varepsilon), \qquad L_{\varepsilon\text{-Q}}(-c) = \alpha(c-\varepsilon),$$
$$L_{\varepsilon\text{-ALS}}(c) = (1-\alpha)(c-\varepsilon)^2, \qquad L_{\varepsilon\text{-ALS}}(-c) = \alpha(c-\varepsilon)^2.$$

Hence

$$L_{\varepsilon\text{-BALS}}(c) = \frac{1}{\gamma}\Big[1 - \frac{1}{1+\lambda(1-\alpha)(c-\varepsilon)^2}\Big], \quad L_{\varepsilon\text{-BALS}}(-c) = \frac{1}{\gamma}\Big[1 - \frac{1}{1+\lambda\alpha(c-\varepsilon)^2}\Big].$$

If $\alpha \neq 0.5$, then $\alpha \neq 1-\alpha$, so the pairs $(L_{\varepsilon\text{-Q}}(c), L_{\varepsilon\text{-Q}}(-c))$ and $(L_{\varepsilon\text{-BALS}}(c), L_{\varepsilon\text{-BALS}}(-c))$ both differ; consequently $L(c) \neq L(-c)$, which establishes asymmetry.

3. Insensitivity. If $|z| \leq \varepsilon$, then directly $L_{\varepsilon\text{-Q}}(z) = 0$ and $L_{\varepsilon\text{-ALS}}(z) = 0$, which forces $L_{\varepsilon\text{-BALS}}(z) = 0$ and hence $L(z) = 0$. Conversely, suppose $L(z) = 0$. By nonnegativity of the components, we must have $\tau L_{\varepsilon\text{-Q}}(z) = 0$ and $(1-\tau)L_{\varepsilon\text{-BALS}}(z) = 0$. If $\tau > 0$, the first equality yields $L_{\varepsilon\text{-Q}}(z) = 0$. If $\tau < 1$, the second equality forces $L_{\varepsilon\text{-BALS}}(z) = 0$, which implies $L_{\varepsilon\text{-ALS}}(z) = 0$ (because $L_{\varepsilon\text{-BALS}}$ vanishes only when $L_{\varepsilon\text{-ALS}}$ does). In either case we obtain $|z| \leq \varepsilon$. Hence $L(z) = 0 \iff |z| \leq \varepsilon$.

4. Monotonicity. Where L is differentiable, its derivative is

$$L'(z) = \tau L'_{\varepsilon\text{-Q}}(z) + (1-\tau)L'_{\varepsilon\text{-BALS}}(z), \quad L'_{\varepsilon\text{-BALS}}(z) = \frac{\lambda}{\gamma} \cdot \frac{L'_{\varepsilon\text{-ALS}}(z)}{[1+\lambda L_{\varepsilon\text{-ALS}}(z)]^2}, \tag{20}$$

with the base derivatives given by

$$L'_{\varepsilon\text{-Q}}(z) = \begin{cases} -\alpha, & z < -\varepsilon, \\ 0, & |z| < \varepsilon, \\ 1-\alpha, & z > \varepsilon, \end{cases} \qquad L'_{\varepsilon\text{-ALS}}(z) = \begin{cases} 2\alpha(z+\varepsilon), & z < -\varepsilon, \\ 0, & |z| < \varepsilon, \\ 2(1-\alpha)(z-\varepsilon), & z > \varepsilon. \end{cases} \tag{21}$$

For $z < -\varepsilon$, we have $L'_{\varepsilon\text{-Q}}(z) = -\alpha < 0$ and $L'_{\varepsilon\text{-ALS}}(z) = 2\alpha(z+\varepsilon) < 0$, which implies $L'_{\varepsilon\text{-BALS}}(z) < 0$ and therefore $L'(z) < 0$. Hence L is strictly decreasing on $(-\infty, -\varepsilon)$. For $z > \varepsilon$, we have $L'_{\varepsilon\text{-Q}}(z) = 1-\alpha > 0$ and $L'_{\varepsilon\text{-ALS}}(z) = 2(1-\alpha)(z-\varepsilon) > 0$, whence $L'(z) > 0$ and L is strictly increasing on $(\varepsilon, +\infty)$. At the non-differentiable points $z = \pm\varepsilon$, the left and right derivatives retain the monotonic behaviour.

5. Lipschitz continuity. Since L is continuous and piecewise differentiable, it suffices to bound $|L'(z)|$ at all differentiable points by a uniform constant M.

Region $|z| < \varepsilon$: $L'(z) = 0$.

Region $z < -\varepsilon$: Set $u = -(z+\varepsilon) > 0$. Then

$$L'_{\varepsilon\text{-Q}}(z) = -\alpha, \quad L_{\varepsilon\text{-ALS}}(z) = \alpha u^2, \quad L'_{\varepsilon\text{-ALS}}(z) = -2\alpha u.$$

Therefore,

$$|L'(z)| = \tau\alpha + (1-\tau)\frac{2\lambda\alpha}{\gamma} \cdot \frac{u}{(1+\lambda\alpha u^2)^2}. \tag{22}$$

Let $D = \lambda\alpha$ and define $h(u) = \frac{u}{(1+Du^2)^2}$ for $u > 0$. From $h'(u) = \frac{1-3Du^2}{(1+Du^2)^3}$, the maximum is attained at $u = 1/\sqrt{3D}$ and equals

$$h_{\max} = \frac{1/\sqrt{3D}}{(1+1/3)^2} = \frac{3\sqrt{3}}{16\sqrt{D}}. \tag{23}$$

Consequently,

$$|L'(z)| \le \tau\alpha + (1-\tau)\frac{2\lambda\alpha}{\gamma} \cdot \frac{3\sqrt{3}}{16\sqrt{\lambda\alpha}} = \tau\alpha + (1-\tau)\frac{3\sqrt{3\lambda\alpha}}{8\gamma}. \tag{24}$$

Region $z > \varepsilon$: Put $v = z - \varepsilon > 0$. By symmetry,

$$|L'(z)| \le \tau(1-\alpha) + (1-\tau)\frac{3\sqrt{3\lambda(1-\alpha)}}{8\gamma}. \tag{25}$$

Taking the supremum over all three regions, we obtain, for every differentiable point,

$$|L'(z)| \le \tau\max(\alpha, 1-\alpha) + (1-\tau)\frac{3\sqrt{3\lambda}}{8\gamma}\max\big(\sqrt{\alpha}, \sqrt{1-\alpha}\big) \equiv M, \tag{26}$$

which is exactly the constant announced in Eq. (19). Because L is continuous and almost everywhere differentiable with derivative bounded by M, the mean-value theorem yields

$$|L(z_1) - L(z_2)| \le M|z_1 - z_2| \quad \forall\, z_1, z_2 \in \mathbb{R}.$$

This proves M-Lipschitz continuity and completes the proof. □

A.2 Proof of Theorem 3.2

Theorem A.2 (Theorem 3.2 restated. Generalized Weak Pseudo-convexity)**.** *$L_{\varepsilon\text{-}HQ}(z)$ is* generalized weakly pseudo-convex *(GWPC) (Bazaraa et al., 2006) on $\mathbb{R}$: there exists a constant $K > 0$ such that for any $z \in \mathbb{R}$ and any global minimizer $z^* \in [-\varepsilon, \varepsilon]$, there exists a Clarke subgradient $g \in \partial L_{\varepsilon\text{-}HQ}(z)$ satisfying*

$$L_{\varepsilon\text{-}HQ}(z) - L_{\varepsilon\text{-}HQ}(z^*) \le K \cdot \frac{g\,(z - z^*)}{\|g\|}, \tag{27}$$

where the right-hand side is taken as 0 when $g = 0$. A feasible GWPC constant is

$$K = \tau \max(\alpha, 1-\alpha) + \frac{(1-\tau)\sqrt{\lambda}}{2\gamma} \max(\sqrt{\alpha}, \sqrt{1-\alpha}). \tag{28}$$

Proof. From Theorem 3.1, the set of global minimizers of $L_{\varepsilon\text{-HQ}}$ is $[-\varepsilon, \varepsilon]$ and $L_{\varepsilon\text{-HQ}}(z^*) = 0$ for every $z^* \in [-\varepsilon, \varepsilon]$. Hence it suffices to prove that for each $z \in \mathbb{R}$ there exist $z^* \in [-\varepsilon, \varepsilon]$ and $g \in \partial L_{\varepsilon\text{-HQ}}(z)$ satisfying

$$L_{\varepsilon\text{-HQ}}(z) \le K \frac{g\,(z - z^*)}{\|g\|},$$

with K as defined in Eq. (28), where the right-hand side is interpreted as 0 when $g = 0$. We distinguish several cases.

Case 1: $|z| < \varepsilon$. Here $L_{\varepsilon\text{-HQ}}(z) = 0$ and the loss is differentiable with $L'_{\varepsilon\text{-HQ}}(z) = 0$. Taking $g = 0$ makes both sides of Eq. (27) vanish, so the inequality holds trivially.

Case 2: $z < -\varepsilon$. On this interval $L_{\varepsilon\text{-HQ}}$ is differentiable; set $g = L'_{\varepsilon\text{-HQ}}(z) < 0$, where the strict negativity follows from the monotonicity proved in Theorem 3.1. Choose $z^* = -\varepsilon$ and let $u = -(z + \varepsilon) > 0$, so that $z - z^* = -u$. Since $g < 0$,

$$\frac{g\,(z - z^*)}{\|g\|} = \frac{g \cdot (-u)}{-g} = u.$$

Expanding the loss value yields

$$\begin{aligned} L_{\varepsilon\text{-HQ}}(z) &= \tau L_{\varepsilon\text{-Q}}(z) + (1-\tau) L_{\varepsilon\text{-BALS}}(z) \\ &= \tau\alpha u + \frac{1-\tau}{\gamma}\Big[1 - \frac{1}{1+\lambda\alpha u^2}\Big] \\ &= \tau\alpha u + \frac{1-\tau}{\gamma} \cdot \frac{\lambda\alpha u^2}{1+\lambda\alpha u^2} = \tau\alpha u + \frac{1-\tau}{\gamma}\, u\, \frac{\lambda\alpha u}{1+\lambda\alpha u^2}. \end{aligned} \tag{29}$$

Define $\phi(u) = \dfrac{\lambda\alpha u}{1+\lambda\alpha u^2}$ for $u > 0$. One readily checks that ϕ attains its maximum at $u = 1/\sqrt{\lambda\alpha}$, with value $\phi_{\max} = \frac{\sqrt{\lambda\alpha}}{2}$. Consequently,

$$L_{\varepsilon\text{-HQ}}(z) \le \tau\alpha u + \frac{1-\tau}{\gamma}\, u \cdot \frac{\sqrt{\lambda\alpha}}{2} = \Big(\tau\alpha + \frac{(1-\tau)\sqrt{\lambda\alpha}}{2\gamma}\Big)\, u. \tag{30}$$

Denote $K_1 = \tau\alpha + \frac{(1-\tau)\sqrt{\lambda\alpha}}{2\gamma}$. Then

$$L_{\varepsilon\text{-HQ}}(z) \le K_1 u = K_1 \frac{g\,(z - z^*)}{\|g\|} \le K \frac{g\,(z - z^*)}{\|g\|}, \tag{31}$$

since $K_1 \le K$ by definition of K.

Case 3: $z > \varepsilon$. Symmetrically, $L_{\varepsilon\text{-HQ}}$ is differentiable with $g = L'_{\varepsilon\text{-HQ}}(z) > 0$. Pick $z^* = \varepsilon$ and set $v = z - \varepsilon > 0$, giving $\frac{g\,(z - z^*)}{\|g\|} = v$. An analogous computation yields

$$L_{\varepsilon\text{-HQ}}(z) \le \Big(\tau(1-\alpha) + \frac{(1-\tau)\sqrt{\lambda(1-\alpha)}}{2\gamma}\Big)\, v. \tag{32}$$

Set $K_2 = \tau(1-\alpha) + \frac{(1-\tau)\sqrt{\lambda(1-\alpha)}}{2\gamma} \le K$. Then

$$L_{\varepsilon\text{-HQ}}(z) \le K_2 v = K_2 \frac{g\,(z - z^*)}{\|g\|} \le K \frac{g\,(z - z^*)}{\|g\|}. \tag{33}$$

Table 3: Comparison of loss function properties.

Loss function	Asymmetry	Insensitivity	(Sub)gradient property	Boundedness	Convexity
Q loss	✓	×	Binary	Unbounded	✓
ALS loss	✓	×	Continuous	Unbounded	✓
ε-ALS loss	✓	✓	Continuous	Unbounded	✓
BQ loss	✓	×	Continuous, except at 0	Fully bounded	×
BALS loss	✓	×	Continuous	Fully bounded	×
ε-HQ loss (Ours)	✓	✓	Continuous, except at $\pm\varepsilon$	Hybrid	GWPC (Theorem. 3.2)

Case 4: Non-differentiable points $z = \pm\varepsilon$. At these points $L_{\varepsilon\text{-HQ}}(z) = 0$. The Clarke subdifferential is

$$\partial L_{\varepsilon\text{-HQ}}(-\varepsilon) = [-\tau\alpha,\, 0], \qquad \partial L_{\varepsilon\text{-HQ}}(\varepsilon) = [0,\, \tau(1-\alpha)]. \tag{34}$$

If we choose $g = 0$, Eq. (27) holds trivially. For any non-zero g in the subdifferential, g and $z - z^*$ have the same sign for every $z^* \in [-\varepsilon, \varepsilon]$ (since $z = \varepsilon$ implies $g \geq 0$ and $z - z^* \geq 0$; $z = -\varepsilon$ implies $g \leq 0$ and $z - z^* \leq 0$). Hence the right-hand side of Eq. (27) is non-negative, while the left-hand side is 0, so the inequality holds as equality.

We have exhibited, for every $z \in \mathbb{R}$, a global minimizer $z^* \in [-\varepsilon, \varepsilon]$ and a subgradient $g \in \partial L_{\varepsilon\text{-HQ}}(z)$ that satisfy Eq. (27) with the constant K defined in Eq. (28). Hence $L_{\varepsilon\text{-HQ}}$ is generalized weakly pseudo-convex on $\mathbb{R}$. □

A.3 Detailed Loss Comparison

Table 3 systematically compares the ε-HQ loss with five representative loss functions from the main text across five properties: asymmetry, insensitivity, (sub)gradient type, boundedness, and convexity.

Asymmetry. All compared losses are asymmetric when $\alpha \neq 0.5$, weighting positive and negative residuals by α and $1 - \alpha$, respectively. This asymmetry is essential for meeting a prescribed miscoverage rate.

Insensitivity. Only ε-ALS and ε-HQ possess an explicit insensitive zone. For $|z| \leq \varepsilon$, both the loss and its gradient are identically zero, filtering minor fluctuations and preventing unnecessary updates. The remaining losses respond to arbitrarily small residuals, which can be detrimental under high-frequency noise in non-stationary settings.

(Sub)gradient property. ALS, ε-ALS, and BALS are fully smooth and deliver a continuous, magnitude-aware gradient. In contrast, the Q loss has a binary subgradient that only indicates the sign of the error. BQ has a continuous gradient almost everywhere, except at the origin. The ε-HQ loss has a continuous gradient except at the insensitive-zone boundaries $\pm\varepsilon$; within $[-\varepsilon, \varepsilon]$, the gradient is zero, suppressing spurious updates while preserving fine-grained feedback outside.

Boundedness. Q loss, ALS, and ε-ALS are unbounded, meaning extreme residuals can disproportionately affect the update. BQ and BALS are fully bounded, limiting outlier influence but causing gradient saturation for moderately large residuals. The ε-HQ loss adopts a hybrid structure: its quadratic component is bounded to suppress outliers, while its linear component remains unbounded to avoid saturation and keep gradients informative for large coverage errors.

Convexity. Q loss, ALS, and ε-ALS are globally convex, supporting standard gradient descent analysis. BQ and BALS sacrifice convexity for boundedness. The ε-HQ loss is also non-convex, yet it satisfies generalized weak pseudo-convexity (GWPC; Theorem 3.2), which guarantees a descent direction toward any global minimizer. GWPC is sufficient for the sublinear dynamic regret bounds established in Section 5 despite the absence of convexity (Gao et al., 2018).

Collectively, these five properties directly underpin the practical advantages of the ε-HQ loss: the continuous gradient with an insensitive zone provides fine-grained feedback; the hybrid boundedness ensures robustness with sustained updates; and the insensitive zone yields noise tolerance. Together with Lipschitz continuity (Theorem 3.1), they constitute the principled theoretical foundation for the dual-calibrated online conformal inference framework developed in Section 4.

A.4 Determining the Target Interval

The adaptive residual mapping mechanism introduced in Section 4.2 requires compressing raw residuals into a predefined interval $[Z_{\min}, Z_{\max}]$. The endpoints are determined analytically from the gradient decay characteristics of the ε-HQ loss. We detail the computation below.

Recall that $\varepsilon_{\min}$ is the lower bound of the insensitive-zone width defined in Section 4.2. At the boundaries $\pm\varepsilon_{\min}$, the quadratic contribution to the gradient vanishes, so the subgradient magnitudes take the simple forms

$$\rho_+ = \left|\frac{\partial L_{\varepsilon\text{-HQ}}}{\partial z}(\varepsilon_{\min})\right| = \tau(1-\alpha), \qquad \rho_- = \left|\frac{\partial L_{\varepsilon\text{-HQ}}}{\partial z}(-\varepsilon_{\min})\right| = \tau\alpha. \tag{35}$$

The theoretically maximal gradient magnitudes on the positive and negative half-axes are obtained in the Lipschitz continuity proof of Theorem 3.1:

$$g^+_{\max} = \tau(1-\alpha) + \frac{3\sqrt{3\lambda(1-\alpha)}}{8\gamma}(1-\tau), \qquad g^-_{\max} = \tau\alpha + \frac{3\sqrt{3\lambda\alpha}}{8\gamma}(1-\tau). \tag{36}$$

To avoid hard thresholding, we introduce a gradient decay factor $\Gamma \in (0,1)$ (typically 0.05) and define effective gradient bounds that interpolate between the boundary and maximal magnitudes:

$$g_{\text{pos}} = \rho_+ + \Gamma \cdot (g^+_{\max} - \rho_+), \qquad g_{\text{neg}} = \rho_- + \Gamma \cdot (g^-_{\max} - \rho_-). \tag{37}$$

The endpoint $Z_{\max}$ is set to $\varepsilon_{\min} + u^*$, where u^* is the smallest $u > 0$ satisfying

$$\frac{\partial L_{\varepsilon\text{-HQ}}}{\partial z}(\varepsilon_{\min} + u) = g_{\text{pos}}.$$

Inserting the explicit gradient expression from Section 3.2 and defining

$$A_{\text{pos}} = \tau(1-\alpha), \qquad B_{\text{pos}} = \frac{2\lambda(1-\tau)(1-\alpha)}{\gamma},$$

reduces this condition to

$$A_{\text{pos}} + \frac{B_{\text{pos}}\, u}{(1+\lambda(1-\alpha)u^2)^2} = g_{\text{pos}}. \tag{38}$$

Rearranging Eq. (38) yields a quartic equation in u, whose smallest positive real root is taken as u^*; if no positive real root exists, a fallback value of 10 is adopted.

The negative endpoint $Z_{\min}$ is obtained symmetrically. Let

$$A_{\text{neg}} = \tau\alpha, \qquad B_{\text{neg}} = \frac{2\lambda(1-\tau)\alpha}{\gamma}.$$

Solving

$$\left|\frac{\partial L_{\varepsilon\text{-HQ}}}{\partial z}(-\varepsilon_{\min} - u)\right| = A_{\text{neg}} + \frac{B_{\text{neg}}\, u}{(1+\lambda\alpha u^2)^2} = g_{\text{neg}} \tag{39}$$

for the smallest $u > 0$ gives the corresponding u^* (with the same fallback), and $Z_{\min} = -u^* - \varepsilon_{\min}$.

The linear interpolation in Eq. (37) smoothly adapts the effective gradient bound to the actual gradient decay profile, thereby avoiding the numerical instability that hard clipping can introduce. The resulting endpoints $Z_{\min}$ and $Z_{\max}$ are then used in the adaptive residual mapping of Section 4.2.

B Proofs for Section 5

This appendix contains the complete proofs of the theoretical guarantees presented in Section 5. Appendix B.1 clarifies the relations among the original, composite, and scaled losses and verifies that the scaled loss inherits the properties required by the analysis. Appendix B.2 establishes the dynamic regret bound for the base calibration (Theorem 5.1), Appendix B.3 proves that the momentum calibration does not degrade the regret order (Theorem 5.2), Appendix B.4 bounds the coverage error (Theorem 5.3), and Appendix B.5 proves the asymptotic coverage bias (Corollary 5.4).

B.1 Relations among the Original, Composite, and Scaled Losses

We clarify the three losses used in the analysis and verify that the scaled loss inherits the properties required by the regret and coverage analyses.

Original loss. The original loss $\ell_t(q) = L_{\varepsilon\text{-HQ}}(s_t - q; \alpha, \varepsilon_t, \tau, \gamma, \lambda)$ defines the instantaneous optimum $q_t^{\mathrm{opt}} \in \arg\min_{q\in\mathcal{Q}} \ell_t(q)$ and the statistical comparator q_t^{qt}. By Theorems 3.1 and 3.2, it is M-Lipschitz and K-GWPC; the gradient lower bound μ of Assumption 5.1 follows from the linear component of $L_{\varepsilon\text{-HQ}}$ and is given by Eq. (6).

Composite loss. To avoid gradient saturation of the bounded quadratic component (Section 4.2), the implemented algorithm replaces the raw residual $s_t - q$ by the adaptively mapped residual $z_t' = \phi_t(s_t - q)$, where $\phi_t(z) = a_t^s z + b_t^s$ is the sign-preserving affine mapping in Eq. (7) with $a_t^s \in [a_{\min}, a_{\max}]$ and $a_{\min} > 0$. This mapping, together with the insensitive-zone width ε_t, provides the algorithm's adaptivity. The composite loss is $\bar{\ell}_t(q) = L_{\varepsilon\text{-HQ}}\big(\phi_t(s_t - q)\big)$, with gradient $\partial\bar{\ell}_t/\partial q(\hat{q}_t) = a_t^s g_t$, where $g_t = -\partial L_{\varepsilon\text{-HQ}}/\partial z(z_t')$.

Gradient descent on $\bar{\ell}_t$ with a fixed learning rate η yields the update $\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta a_t^s g_t)$. Since $a_t^s \le a_{\max}$, the Lipschitz and GWPC constants are amplified by up to $a_{\max}$ relative to the original loss, which weakens the regret bound and, in turn, the coverage error bound. The empirical behavior of this variant is examined in Appendix D.9.

Scaled loss. To avoid these complications, rescaling the composite loss by $1/a_t^s$ yields the scaled loss

$$\tilde{\ell}_t(q) = \frac{1}{a_t^s}\,\bar{\ell}_t(q) = \frac{1}{a_t^s}\,L_{\varepsilon\text{-HQ}}\big(\phi_t(s_t - q)\big).$$

Its gradient at $\hat{q}_t$ is $\partial\tilde{\ell}_t/\partial q(\hat{q}_t) = -L'_{\varepsilon\text{-HQ}}(z_t') = g_t$, which coincides with the subgradient used by Algorithm 1. Hence, the base calibration is exactly online proximal gradient descent on $\tilde{\ell}_t$ with the fixed learning rate η, without amplifying the constants or introducing a time-varying learning rate. The adaptive behavior of the algorithm—carried by z_t' and ε_t—is preserved, while the analysis remains standard and the constants unscaled.

Property inheritance. For theoretical completeness, we verify that $\tilde{\ell}_t$ inherits the three key properties of ℓ_t with identical constants.

Lipschitz continuity. Since ϕ_t has slope a_t^s and $L_{\varepsilon\text{-HQ}}$ is M-Lipschitz, $|\partial\tilde{\ell}_t/\partial q| = |L'_{\varepsilon\text{-HQ}}(z_t')| \le M$, and thus $\tilde{\ell}_t$ is M-Lipschitz.

GWPC. Let $u = \phi_t(s_t - q)$ and $u^{\mathrm{opt}} = \phi_t(s_t - q_t^{\mathrm{opt}})$. Since $u - u^{\mathrm{opt}} = -a_t^s(q - q_t^{\mathrm{opt}})$ and $\partial\tilde{\ell}_t/\partial q = -\partial L_{\varepsilon\text{-HQ}}/\partial u$, the GWPC inequality for $L_{\varepsilon\text{-HQ}}$ with constant K gives

$$\tilde{\ell}_t(q) - \tilde{\ell}_t(q_t^{\mathrm{opt}}) = \frac{L(u) - L(u^{\mathrm{opt}})}{a_t^s} \le \frac{K}{a_t^s}\cdot\frac{g_u(u - u^{\mathrm{opt}})}{\|g_u\|} = K\cdot\frac{g_q(q - q_t^{\mathrm{opt}})}{\|g_q\|}, \tag{40}$$

where $g_q = -g_u$. Hence $\tilde{\ell}_t$ is K-GWPC.

Gradient lower bound. When $\tilde{\ell}_t(q) > \tilde{\ell}_t(q_t^{\mathrm{opt}})$, the corresponding u satisfies $L(u) > L(u^{\mathrm{opt}})$, and thus $|g_u| \ge \mu$ implies $|g_q| \ge \mu$. Hence Assumption 5.1 holds for $\tilde{\ell}_t$ with the same constant μ.

In summary, the scaled loss $\tilde{\ell}_t$ satisfies the conditions required by Theorems 5.1 and 5.2—namely, M-Lipschitz continuity, K-GWPC, and the gradient lower bound μ—with the same constants as the original loss. Hence the subsequent regret analysis applies directly to the implemented algorithm.

B.2 Proof of Theorem 5.1

Theorem B.1 (Theorem 5.1 restated. Base Dynamic Regret Bound)**.** *Suppose Assumption 5.1 holds, and let $\tilde{\ell}_t$ be M-Lipschitz and K-GWPC as verified in Appendix B.1. Choose the base learning rate*

as $\eta = \sqrt{\frac{4R^2+3RV_T}{TM^2}}$. *Then the intermediate thresholds* $\{\hat{q}_t\}$ *from the base calibration satisfy*

$$\sum_{t=1}^{T}\big[\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big] \le \frac{KM}{\mu}\sqrt{T(4R^2+3RV_T)} = \mathcal{O}\big(\sqrt{T(1+V_T)}\big). \tag{41}$$

Proof. Let $z_t = \hat{q}_t - q_t^{\mathrm{opt}}$ and $\rho_t = |q_t^{\mathrm{opt}} - q_{t+1}^{\mathrm{opt}}|$; the path length is $V_T = \sum_{t=1}^{T-1}\rho_t$. The proof is organized into five steps.

Step 1: One-step proximal inequality. The base calibration update is $\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta g_t)$. By non-expansiveness of the projection onto the convex set $\mathcal{Q}$, for any $q \in \mathcal{Q}$,

$$(\hat{q}_{t+1} - q)^2 \le (\hat{q}_t - \eta g_t - q)^2.$$

Taking $q = q_t^{\mathrm{opt}}$ and expanding the square gives

$$(\hat{q}_{t+1} - q_t^{\mathrm{opt}})^2 \le z_t^2 - 2\eta g_t(\hat{q}_t - q_t^{\mathrm{opt}}) + \eta^2 g_t^2. \tag{42}$$

Step 2: Switching to the dynamic comparator q_{t+1}^{opt}. By the triangle inequality,

$$|z_{t+1}| = |\hat{q}_{t+1} - q_{t+1}^{\mathrm{opt}}| \le |\hat{q}_{t+1} - q_t^{\mathrm{opt}}| + |q_t^{\mathrm{opt}} - q_{t+1}^{\mathrm{opt}}| = |\hat{q}_{t+1} - q_t^{\mathrm{opt}}| + \rho_t. \tag{43}$$

Squaring both sides and using $(a+b)^2 \le a^2 + 2|a|b + b^2$ together with the bounds $|\hat{q}_{t+1} - q_t^{\mathrm{opt}}| \le R$ and $\rho_t \le R$ yields

$$z_{t+1}^2 \le (\hat{q}_{t+1} - q_t^{\mathrm{opt}})^2 + 3R\rho_t. \tag{44}$$

Substituting Eq. (42) into Eq. (44) provides the key recursion

$$z_{t+1}^2 \le z_t^2 - 2\eta g_t(\hat{q}_t - q_t^{\mathrm{opt}}) + \eta^2 g_t^2 + 3R\rho_t. \tag{45}$$

Step 3: Applying generalized weak pseudo-convexity. By Theorem 3.2 and the inheritance proved in Appendix B.1, there exists a subgradient $g_t \in \partial\tilde{\ell}_t(\hat{q}_t)$ such that

$$\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}}) \le K \cdot \frac{g_t(\hat{q}_t - q_t^{\mathrm{opt}})}{\|g_t\|}.$$

Rearranging gives

$$g_t(\hat{q}_t - q_t^{\mathrm{opt}}) \ge \frac{\|g_t\|}{K}\big(\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big). \tag{46}$$

Step 4: Substituting and summing. Inserting Eq. (46) into Eq. (45) leads to

$$z_{t+1}^2 \le z_t^2 - \frac{2\eta}{K}\|g_t\|\big(\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big) + \eta^2 g_t^2 + 3R\rho_t, \tag{47}$$

and rearranging,

$$\frac{2\eta}{K}\|g_t\|\big(\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big) \le z_t^2 - z_{t+1}^2 + \eta^2 g_t^2 + 3R\rho_t. \tag{48}$$

Since $\Delta_t = \tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}}) \ge 0$, Assumption 5.1 gives $\|g_t\| \ge \mu$ whenever $\Delta_t > 0$, while $\Delta_t = 0$ makes both sides zero. Hence $\|g_t\|\Delta_t \ge \mu\Delta_t$ for all t; replacing $\|g_t\|\Delta_t$ by its lower bound $\mu\Delta_t$ in Eq. (48) yields

$$\frac{2\eta\mu}{K}\big(\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big) \le z_t^2 - z_{t+1}^2 + \eta^2 g_t^2 + 3R\rho_t. \tag{49}$$

Summing Eq. (49) over $t = 1, \ldots, T$ telescopes the z_t^2 terms:

$$\sum_{t=1}^{T}(z_t^2 - z_{t+1}^2) = z_1^2 - z_{T+1}^2 \le R^2.$$

Using the Lipschitz bound $\|g_t\| \le M$ and $\sum_{t=1}^{T} \rho_t \le V_T + R$, we obtain

$$\frac{2\eta\mu}{K} \sum_{t=1}^{T} \big(\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big) \le R^2 + \eta^2 T M^2 + 3R(V_T + R). \tag{50}$$

Simplifying the right-hand side to $4R^2 + \eta^2 T M^2 + 3RV_T$ and dividing by $\frac{2\eta\mu}{K}$ yields

$$\sum_{t=1}^{T} \big[\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big] \le \frac{K}{2\eta\mu}\big(4R^2 + \eta^2 T M^2 + 3RV_T\big). \tag{51}$$

Step 5: Optimizing the learning rate. The right-hand side of Eq. (51) is minimized by balancing the terms $4R^2$ and $\eta^2 T M^2$. Choosing

$$\eta = \sqrt{\frac{4R^2 + 3RV_T}{TM^2}},$$

and substituting into Eq. (51) gives

$$\begin{aligned} \sum_{t=1}^{T} \big[\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big] &\le \frac{K}{2\mu}\sqrt{\frac{TM^2}{4R^2 + 3RV_T}} \cdot 2(4R^2 + 3RV_T) \\ &= \frac{KM}{\mu}\sqrt{T(4R^2 + 3RV_T)} = \mathcal{O}\big(\sqrt{T(1+V_T)}\big), \end{aligned} \tag{52}$$

which completes the proof. □

B.3 Proof of Theorem 5.2

Theorem B.2 (Theorem 5.2 restated. Joint Dynamic Regret Bound)**.** *Under Theorem 5.1's assumptions, choose the momentum learning rate as* $\eta' = \frac{K}{\mu}\sqrt{\frac{4R^2+3RV_T}{T}}$. *Then the final thresholds* $\{q_t\}$ *of* ε-***HQ-DCOCI*** *satisfy*

$$\tilde{R}_T^{dyn} = \sum_{t=1}^{T} \big[\tilde{\ell}_t(q_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big] \le \frac{2KM}{\mu}\sqrt{T(4R^2 + 3RV_T)} = \mathcal{O}\big(\sqrt{T(1+V_T)}\big). \tag{53}$$

Proof. The momentum calibration update is given in Section 4.3 as $q_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_{t+1} - \eta'\nu_{t+1})$, where ν_{t+1} is the averaged normalized gradient direction with $\|\nu_{t+1}\| \le 1$. For the regret analysis, we adopt the equivalent time-shifted form $q_t = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta'\nu_t)$, where ν_t is the direction-consistency norm defined in Eq. (9). By non-expansiveness of the projection,

$$|q_t - \hat{q}_t| \le \eta'\|\nu_t\| \le \eta'. \tag{54}$$

Using the M-Lipschitz continuity of $\tilde{\ell}_t$ (Appendix B.1) and the fact that $\tilde{\ell}_t(q_t) - \tilde{\ell}_t(\hat{q}_t) \le |\tilde{\ell}_t(q_t) - \tilde{\ell}_t(\hat{q}_t)|$, we obtain

$$\tilde{\ell}_t(q_t) - \tilde{\ell}_t(\hat{q}_t) \le M|q_t - \hat{q}_t| \le \eta' M. \tag{55}$$

Summing over $t = 1, \dots, T$ yields

$$\sum_{t=1}^{T} \big[\tilde{\ell}_t(q_t) - \tilde{\ell}_t(\hat{q}_t)\big] \le \eta' M T. \tag{56}$$

The overall dynamic regret decomposes as

$$\tilde{R}_T^{\mathrm{dyn}} = \underbrace{\sum_{t=1}^{T} \big[\tilde{\ell}_t(q_t) - \tilde{\ell}_t(\hat{q}_t)\big]}_{\text{Momentum perturbation}} + \underbrace{\sum_{t=1}^{T} \big[\tilde{\ell}_t(\hat{q}_t) - \tilde{\ell}_t(q_t^{\mathrm{opt}})\big]}_{\text{Base regret } \tilde{\mathcal{R}}_T^{\mathrm{base}}}. \tag{57}$$

By Theorem 5.1 (Eq. (41) in Appendix B.2), the base regret satisfies

$$\tilde{\mathcal{R}}_T^{\text{base}} \le \frac{KM}{\mu}\sqrt{T(4R^2+3RV_T)}. \tag{58}$$

Substituting this bound and Eq. (56) into Eq. (57) gives

$$\tilde{R}_T^{\text{dyn}} \le \eta' MT + \frac{KM}{\mu}\sqrt{T(4R^2+3RV_T)}. \tag{59}$$

Choosing the momentum learning rate as specified,

$$\eta' = \frac{K}{\mu}\sqrt{\frac{4R^2+3RV_T}{T}},$$

makes the two terms equal and results in

$$\tilde{R}_T^{\text{dyn}} \le \frac{2KM}{\mu}\sqrt{T(4R^2+3RV_T)} = \mathcal{O}\big(\sqrt{T(1+V_T)}\big), \tag{60}$$

which is exactly Eq. (53). More generally, any choice $\eta' = \mathcal{O}(1/\sqrt{T})$ gives $\eta' MT = \mathcal{O}(\sqrt{T})$, preserving the regret order. Thus the momentum calibration does not degrade the worst-case dynamic regret. □

B.4 Proof of Theorem 5.3

Theorem B.3 (Theorem 5.3 restated. Coverage Error Bound)**.** *Under Assumption 5.2, let q_t^{qt} be the conditional $(1-\alpha)$-quantile of s_t, i.e., $\mathbb{P}(s_t \le q_t^{\text{qt}} \mid \mathcal{F}_{t-1}) = 1-\alpha$. With probability at least $1-\delta$, the empirical coverage $\hat{C}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{1}\{Y_t \in \hat{C}_t\}$ of ε-**HQ-DCOCI** satisfies*

$$\big|\hat{C}_T - (1-\alpha)\big| \le \frac{C}{\mu}\Big(\frac{\tilde{R}_T^{\text{dyn}}}{T} + \frac{2\mu}{a_{\min}}\bar{\varepsilon}_T + \Delta_\infty\Big) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{T}}\right), \tag{61}$$

where $\mu = \tau\min(\alpha, 1-\alpha)$, $\tilde{R}_T^{\text{dyn}}$ is the regret bound from Theorem 5.2, $a_{\min} > 0$ is the lower bound of the slope a_t^s in Eq. (7), $\bar{\varepsilon}_T = \frac{1}{T}\sum_{t=1}^T \varepsilon_t$, and $\Delta_\infty = \frac{CMZ_{\text{abs}}^2}{2a_{\min}^2} + \frac{1-\tau}{a_{\min}}$ with $Z_{\text{abs}} = \max(|Z_{\min}|, Z_{\max})$. The $\mathcal{O}(\cdot)$ term hides a constant depending only on $(\tau, \alpha, C, Z_{\text{abs}})$.

Proof. The proof uses the regret bound $\tilde{R}_T^{\text{dyn}}$ for the scaled loss $\tilde{\ell}_t$ (Theorem 5.2) and proceeds in four steps. We use $\gamma = 1$ and $M = 2\tau A$ with $A = \max(\alpha, 1-\alpha)$, as specified in Eq. (6).

Step 1: Threshold deviation via loss differences. Let $m = \min(\alpha, 1-\alpha)$, $d_t = |q_t - q_t^{\text{qt}}|$, and $D_t = \mathbf{1}\{s_t \le q_t\} - \mathbf{1}\{s_t \le q_t^{\text{qt}}\}$. When $D_t \ne 0$, s_t lies strictly between q_t and q_t^{qt}, so the raw residuals $s_t - q_t$ and $s_t - q_t^{\text{qt}}$ have opposite signs. Let $u_1 = \phi_t(s_t - q_t)$ and $u_2 = \phi_t(s_t - q_t^{\text{qt}})$. Since ϕ_t is fixed by the past window at time t, both share the same slope a_t^s and intercept b_t^s; the intercept cancels, and sign preservation gives that u_1 and u_2 are also of opposite signs. Hence

$$|u_1| + |u_2| = |u_1 - u_2| = a_t^s|q_t - q_t^{\text{qt}}| = a_t^s d_t. \tag{62}$$

By the linear-component lower bound $L_{\varepsilon\text{-HQ}}(u) \ge \tau m\,(|u| - \varepsilon_t)_+$ and the elementary inequality $(x-\varepsilon)_+ + (y-\varepsilon)_+ \ge (x+y-2\varepsilon)_+$, we obtain

$$\tilde{\ell}_t(q_t) + \tilde{\ell}_t(q_t^{\text{qt}}) = \frac{L(u_1)+L(u_2)}{a_t^s} \ge \frac{\tau m}{a_t^s}(a_t^s d_t - 2\varepsilon_t)_+. \tag{63}$$

Since $a_t^s \in [a_{\min}, a_{\max}]$ and $a_{\min} > 0$, and thus

$$d_t \le \frac{2\varepsilon_t}{a_{\min}} + \frac{\tilde{\ell}_t(q_t) + \tilde{\ell}_t(q_t^{\text{qt}})}{\tau m}. \tag{64}$$

Averaging over t and using $\frac{1}{T}\sum_t \tilde{\ell}_t(q_t) = \tilde{R}_T^{\text{dyn}}/T$ (since $\tilde{\ell}_t(q_t^{\text{opt}}) = 0$), we obtain

$$\frac{1}{T}\sum_{t=1}^T d_t \le \frac{2}{a_{\min}}\bar{\varepsilon}_T + \frac{1}{\tau m}\left(\frac{\tilde{R}_T^{\text{dyn}}}{T} + \Delta_T\right), \tag{65}$$

where $\Delta_T = \frac{1}{T}\sum_{t=1}^T \tilde{\ell}_t(q_t^{\text{qt}})$.

Step 2: Concentrating the bridge term Δ_T. Under Assumption 5.2, the mapped residual $u = \phi_t(s_t - q_t^{\mathrm{qt}})$ lies in $[Z_{\min}, Z_{\max}]$, and its conditional density is bounded by $C/a_{\min}$, since the map $s_t \mapsto u$ is linear with slope $a_t^s \geq a_{\min}$. Therefore

$$\mathbb{E}\big[|u| \mid \mathcal{F}_{t-1}\big] \leq \frac{C}{a_{\min}} \int_{Z_{\min}}^{Z_{\max}} |u|\, du \leq \frac{C Z_{\mathrm{abs}}^2}{a_{\min}}, \tag{66}$$

where $Z_{\mathrm{abs}} = \max(|Z_{\min}|, Z_{\max})$. Since $L_{\varepsilon\text{-HQ}}(u) \leq \tau A|u| + (1-\tau)/\gamma$, it follows that

$$\mathbb{E}\big[\tilde{\ell}_t(q_t^{\mathrm{qt}}) \mid \mathcal{F}_{t-1}\big] \leq \frac{1}{a_{\min}} \Big(\tau A \cdot \frac{C Z_{\mathrm{abs}}^2}{a_{\min}} + \frac{1-\tau}{\gamma}\Big). \tag{67}$$

Using $\gamma = 1$ and $\tau A = M/2$, we obtain

$$\Delta_\infty = \frac{M C Z_{\mathrm{abs}}^2}{2 a_{\min}^2} + \frac{1-\tau}{a_{\min}}. \tag{68}$$

Define the martingale sequence $\xi_t = \tilde{\ell}_t(q_t^{\mathrm{qt}}) - \mathbb{E}[\tilde{\ell}_t(q_t^{\mathrm{qt}}) \mid \mathcal{F}_{t-1}]$. Since $0 \leq \tilde{\ell}_t(q_t^{\mathrm{qt}}) \leq \bar{L}$ with $\bar{L} = \big(\tau A Z_{\mathrm{abs}} + (1-\tau)/\gamma\big)/a_{\min}$, we have $|\xi_t| \leq \bar{L}$. By Azuma–Hoeffding (Hoeffding, 1963), with probability at least $1 - \delta/3$,

$$\Delta_T \leq \Delta_\infty + \bar{L}\sqrt{\frac{2\log(3/\delta)}{T}}. \tag{69}$$

Step 3: Converting threshold deviation to coverage-indicator difference. By Assumption 5.2, when $D_t \neq 0$, s_t lies in an interval of length d_t, so $\mathbb{E}[|D_t| \mid \mathcal{F}_{t-1}] \leq C d_t$. Let $\zeta_t = |D_t| - \mathbb{E}[|D_t| \mid \mathcal{F}_{t-1}]$. Since $|D_t| \leq 1$, we have $|\zeta_t| \leq 1$. By Azuma–Hoeffding, with probability at least $1 - \delta/3$,

$$\frac{1}{T}\sum_{t=1}^{T} |D_t| \leq \frac{C}{T}\sum_{t=1}^{T} d_t + \sqrt{\frac{\log(3/\delta)}{2T}}. \tag{70}$$

Substituting (65) and (69) into (70) yields

$$\frac{1}{T}\sum_{t=1}^{T} |D_t| \leq \frac{2C}{a_{\min}}\bar{\varepsilon}_T + \frac{C}{\tau m}\Big(\frac{\tilde{R}_T^{\mathrm{dyn}}}{T} + \Delta_\infty\Big) + C_2\sqrt{\frac{\log(3/\delta)}{T}}, \tag{71}$$

where C_2 absorbs all martingale terms.

Step 4: Quantile deviation and combination. Let $M_t = \mathbf{1}\{s_t \leq q_t^{\mathrm{qt}}\} - (1-\alpha)$. Since $\{M_t\}$ is a martingale difference sequence with $|M_t| \leq 1$, Azuma–Hoeffding gives, with probability at least $1 - \delta/3$,

$$\Big|\frac{1}{T}\sum_{t=1}^{T} M_t\Big| \leq \sqrt{\frac{2\log(3/\delta)}{T}}. \tag{72}$$

Decompose the coverage deviation as

$$\big|\hat{C}_T - (1-\alpha)\big| \leq \Big|\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{s_t \leq q_t\} - \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{s_t \leq q_t^{\mathrm{qt}}\}\Big| + \Big|\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{s_t \leq q_t^{\mathrm{qt}}\} - (1-\alpha)\Big|. \tag{73}$$

Combining (71) and (72), and merging the concentration terms (each holding with probability at least $1 - \delta/3$, and thus all hold simultaneously with probability at least $1 - \delta$), and noting $\mu = \tau m$, we obtain

$$\big|\hat{C}_T - (1-\alpha)\big| \leq \frac{C}{\mu}\Big(\frac{\tilde{R}_T^{\mathrm{dyn}}}{T} + \frac{2\mu}{a_{\min}}\bar{\varepsilon}_T + \Delta_\infty\Big) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{T}}\right), \tag{74}$$

where the $\mathcal{O}(\cdot)$ term hides a constant depending only on $\tau, \alpha, C, Z_{\mathrm{abs}}$. This completes the proof. □

B.5 Proof of Corollary 5.4

Corollary B.4 (Corollary 5.4 restated. Asymptotic Coverage Bias)**.** *Under Theorem 5.3's assumptions, as $T \to \infty$ with $V_T = O(T)$,*

$$\limsup_{T\to\infty} \left|\hat{C}_T - (1-\alpha)\right| \le \frac{2C}{a_{\min}}\varepsilon_\infty + \frac{C}{\mu}\Delta_\infty, \tag{75}$$

where $\varepsilon_\infty = \limsup_{T\to\infty} \bar{\varepsilon}_T$, $\bar{\varepsilon}_T = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t$, and $\Delta_\infty = \frac{MCZ_{\text{abs}}^2}{2a_{\min}^2} + \frac{1-\tau}{a_{\min}}$. The first term is controlled by the adaptive insensitive zone and reduces to $\varepsilon_{\min}$ under persistent drift with $\|\nu_t\| \to 1$. The second term is an intrinsic bias of the ε-HQ loss; since $M = 2\tau A$ and $\mu = \tau\min(\alpha, 1-\alpha)$, both Δ_∞ and C/μ depend on τ, so the product $\frac{C}{\mu}\Delta_\infty$ is jointly minimized over $\tau \in (0,1)$ and is small for small C and Z_{abs}.

Proof. From Theorem 5.3, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\hat{C}_T - (1-\alpha)\right| \le \frac{C}{\mu}\Big(\frac{\tilde{R}_T^{\text{dyn}}}{T} + \frac{2\mu}{a_{\min}}\bar{\varepsilon}_T + \Delta_\infty\Big) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{T}}\right). \tag{76}$$

Set $\delta_T = T^{-2}$. Since $\sum_{T=1}^{\infty} T^{-2} < \infty$, the Borel–Cantelli lemma (Durrett, 2019) implies that Eq. (76) holds for all sufficiently large T, almost surely. On this event we take the limit superior as $T \to \infty$ under the condition $V_T = O(T)$.

By Theorem 5.2, $\tilde{R}_T^{\text{dyn}}/T = \mathcal{O}\big(\sqrt{(1+V_T)/T}\big) \to 0$. With $\delta_T = T^{-2}$, the concentration term satisfies $\mathcal{O}\big(\sqrt{\log T/T}\big) \to 0$. By definition, $\limsup_{T\to\infty} \bar{\varepsilon}_T = \varepsilon_\infty$. Collecting the remaining terms,

$$\limsup_{T\to\infty} \left|\hat{C}_T - (1-\alpha)\right| \le \frac{2C}{a_{\min}}\varepsilon_\infty + \frac{C}{\mu}\Delta_\infty. \tag{77}$$

When the insensitive zone is disabled ($\varepsilon_{\min} \to 0$), the first term vanishes asymptotically, leaving only the intrinsic bias $\frac{C}{\mu}\Delta_\infty$. This bias is bounded and jointly tunable via τ, and is small for moderate C and Z_{abs}. Hence the algorithm attains approximately valid coverage with a controllable, bounded asymptotic bias. □

C Details of the ε-**HQ-DCOCI** Framework

This appendix supplements Section 4 with the complete algorithm pseudocode (Appendix C.1) and discusses how the two core modules—the ε-HQ gradient and the momentum calibration—can be integrated into existing OCI algorithms (Appendix C.2).

C.1 Complete Algorithm Procedure

For completeness, Algorithm 1 presents the full pseudocode of the ε-**HQ-DCOCI** framework. The procedure integrates the equal-peak alignment (Eq. 5), adaptive residual mapping (Eq. 7), insensitive-zone tuning (Eqs. 8–10), base calibration (Eq. 12), and momentum calibration (Eq. 14) into a single online loop. This algorithm is referenced in Section 4, and its modular design allows the ε-HQ gradient and momentum calibration to be transferred to existing OCI methods as discussed.

Algorithm 1 produces a symmetric interval from a single threshold q_t. For asymmetric intervals, run two independent instances of ε-HQ-DCOCI, one for each bound with target miscoverage $\alpha/2$. They share the residual mapping and insensitive-zone tuning mechanisms but maintain separate thresholds (q_t^- and q_t^+) and momentum buffers, requiring no algorithmic changes.

C.2 Integration with Other OCI Algorithms

The proposed ε-**HQ-DCOCI** framework comprises two functionally independent modules:

Algorithm 1: Dual-Calibrated Online Conformal Inference with ε-HQ Loss (ε-**HQ-DCOCI**)

Input: Target miscoverage rate $\alpha \in (0,1)$; base learning rate $\eta > 0$; momentum learning rate $\eta' > 0$; mixing coefficient $\tau \in (0,1)$; gradient decay factor Γ; insensitive-zone lower bound $\varepsilon_{\min} > 0$; base predictor $\hat{f}$; threshold upper bound R (e.g., the maximum historical score).

Output: Prediction interval sequence $\{\hat{C}_t\}_{t=1}^{T}$

// Initialization

Fix $\gamma = 1$ and compute $\lambda(\tau)$ via Eq. (5);

Compute initial threshold $\hat{q}_1 \in [0, R]$ (e.g., as the $(1-\alpha)$-quantile of calibration scores);

Initialize residual window $\mathcal{H}_r = \emptyset$ and normalized residual window $\mathcal{H}_n = \emptyset$;

Compute $\rho_{\pm} = |\nabla L_{\varepsilon\text{-HQ}}(\pm\varepsilon_{\min})|$ and set $Z_{\min}, Z_{\max}$ using Γ (see Appendix A.4);

Initialize the history gradient list $\mathcal{G} = \emptyset$;

Define the projection operator $\Pi_{\mathcal{Q}}(x) = \min(\max(x, 0), R)$;

for $t = 1$ **to** T **do**

  Construct prediction interval $\hat{C}_t = [\hat{f}(X_t) - q_t,\ \hat{f}(X_t) + q_t]$;

  Observe the true label Y_t, compute the nonconformity score $s_t = |Y_t - \hat{f}(X_t)|$ and the coverage residual $z_t = s_t - q_t$;

  // Adaptive residual mapping

  Update the window $\mathcal{H}_r \leftarrow \mathcal{H}_r \cup \{z_t\}$, keep latest W values; map z_t to z'_t via Eq. (7);

  // Insensitive zone fine-tuning

  Update the window $\mathcal{H}_n \leftarrow \mathcal{H}_n \cup \{z'_t\}$, keep latest W values; compute ε_t via Eqs. (8)–(10);

  // Base Calibration

  Compute the subgradient $g_t = -\frac{\partial L_{\varepsilon\text{-HQ}}}{\partial z}(z'_t)$ using ε_t;

    Append g_t to $\mathcal{G}$;

    Update the intermediate threshold $\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta g_t)$;

  // Momentum Calibration

  Retrieve latest W gradients from $\mathcal{G}$ (including current g_t; all if $t < W$);

    Compute $\nu_{t+1} = \frac{1}{W}\sum_{i=t-W+1}^{t} g_i/\|g_i\|$;

    Update the final threshold $q_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_{t+1} - \eta'\nu_{t+1})$;

end

- **Module I (ε-HQ gradient)** replaces the binary subgradient of the standard quantile loss with the continuous subgradient $g_t = \nabla \ell_t^{\varepsilon\text{-HQ}}(\hat{q}_t)$, which encodes error magnitude and incorporates an insensitive zone via the dynamic width ε_t.
- **Module II (momentum calibration)** applies a secondary correction based on the smoothed normalized gradient direction $\nu_{t+1} = \frac{1}{W}\sum_{i=t-W+1}^{t} g_i/\|g_i\|$ to the base threshold, suppressing oscillations and stabilizing the update trajectory.

Their combination gives rise to three integration levels: (1) *Module I only*, yielding "ε-HQ-OCI" variants; (2) *Module II only*, yielding "DCOCI" variants with the standard quantile loss; and (3) *both modules*, yielding the full "ε-HQ-DCOCI" variants. In this appendix we instantiate these three levels on the OGD-type family, which consists of three representative update rules: OGD, SF-OGD, and Decay-OGD. Extensions to others are deferred to future work, as those algorithms already embed additional correction terms whose interaction with Modules I–II requires a separate gain-balancing treatment.

Module I Only: ε-HQ-OCI Variants

Replacing the standard quantile subgradient by g_t yields the basic variant

$$\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta g_t), \tag{78}$$

which we refer to as ε-HQ-OGD.

The scale-free variant incorporates an adaptive step-size normalization. Denoting $G_t = \sum_{i=1}^{t} \|g_i\|^2$, its update is

$$\hat{q}_{t+1} = \Pi_{\mathcal{Q}}\Big(\hat{q}_t - \frac{\eta}{\sqrt{G_t + \epsilon}}\, g_t\Big), \tag{79}$$

where $\epsilon > 0$ prevents division by zero. We denote this version as ε-HQ-SF-OGD.

The decaying learning rate variant uses a time-varying step size $\eta_t = \eta \cdot t^{-0.5-\beta}$ with decay offset $\beta > 0$, leading to

$$\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta_t g_t), \tag{80}$$

which we denote as ε-HQ-Decay-OGD.

MODULE II ONLY: DCOCI VARIANTS

Retaining the standard quantile subgradient $g_t^{\mathrm{Q}} = \alpha - \mathbf{1}\{s_t > \hat{q}_t\}$ and appending the momentum calibration to the three OGD-type updates yields

$$q_{t+1} = \Pi_{\mathcal{Q}}\big(\hat{q}_{t+1}^{\mathrm{Q}} - \eta' \nu_{t+1}^{\mathrm{Q}}\big), \tag{81}$$

where $\hat{q}_{t+1}^{\mathrm{Q}}$ denotes the intermediate threshold produced by the corresponding OGD-type update with g_t^{Q}, and

$$\nu_{t+1}^{\mathrm{Q}} = \frac{1}{W} \sum_{i=t-W+1}^{t} \frac{g_i^{\mathrm{Q}}}{\|g_i^{\mathrm{Q}}\|}, \tag{82}$$

is computed from the normalized quantile subgradients. We denote these variants as DCOGD, DCSF-OGD, and DCDecay-OGD.

MODULES I AND II: ε-HQ-DCOCI VARIANTS

Combining both modules, the full update for the three OGD-type variants reads

$$\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta g_t), \qquad q_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_{t+1} - \eta' \nu_{t+1}), \tag{83}$$

with the corresponding modifications for SF-OGD and Decay-OGD. We denote these as ε-HQ-DCOGD, ε-HQ-DCSF-OGD, and ε-HQ-DCDecay-OGD.

In summary, the ε-HQ gradient module can be seamlessly plugged into any OGD-type update rule, and the momentum calibration can be optionally stacked on top. This modular design provides a flexible, reusable paradigm for systematically upgrading online conformal inference algorithms. The three-level comparison (Module I only, Module II only, and both) isolates the contribution of each module and is reported in Appendix D.8.

D ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

D.1 DETAILED EXPERIMENTAL SETTINGS

Hardware and software. Financial time-series predictions were conducted on a workstation (Intel Core i7-8700, 16 GB RAM, NVIDIA Quadro P400). Medical image models were fine-tuned on Kaggle using dual NVIDIA T4 GPUs (16 GB, automatic mixed precision); all subsequent online conformal inference steps ran on the same workstation. The software stack comprised Python 3.10, NumPy (Harris et al., 2020), Pandas (McKinney, 2010), LightGBM (Ke et al., 2017), statsmodels (Seabold & Perktold, 2010), Prophet (Taylor & Letham, 2018), PyTorch (Paszke et al., 2019), and Optuna (Akiba et al., 2019).

General protocol. For both benchmark tasks, the target miscoverage rate is set to $\alpha = 0.1$ (i.e., 90% nominal coverage). Point predictions were produced in a rolling-window fashion (window size $W = 100$), after which all OCI algorithms processed the resulting score sequences online. Hyperparameters were tuned via Bayesian optimization with Optuna (Akiba et al., 2019), using MIS for financial time series and MDSC for medical images as the validation objective. Reported metrics are averages over 5 independent runs, where the financial series are deterministic evaluations. *Details of the original datasets, processed score sequences, and optimal hyperparameters for all compared algorithms are provided in the supplementary material.*

Table 4: Hyperparameter search ranges for all compared algorithms.

Parameter	Category	Range	Algorithms
τ	float	$(0, 1)$	All ε-HQ extensions
Γ	float	$(0.05, 0.95)$	All ε-HQ extensions
η	log-float	$(10^{-3}, 10^{3})$	All baselines and ε-HQ extensions
η'	log-float	$(10^{-4}, 10^{2})$	COP, ε-HQ-DCOCI extensions
K_I	log-float	$(10^{-4}, 10^{2})$	PIDCP
C_{sat}	log-float	$(10^{-2}, 10^{4})$	PIDCP
c	log-float	$(2^{-3}, 2^{3})$	ECI
β	float	$(0, 0.5)$	Decay-OGD, DCDecay-OGD, ε-HQ-Decay-OGD, ε-HQ-DCDecay-OGD

Evaluation metrics. Let $\hat{C}_t$ be the prediction set at time t, $\alpha \in (0, 1)$ the target miscoverage rate, and T the test length. For regression, $\hat{C}_t = [\hat{Q}_t^l, \hat{Q}_t^u]$. Four metrics are reported:

- **Coverage** $\text{Cov} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_t \in \hat{C}_t\}$.
- **Efficiency:** PI-Width $= \frac{1}{T}\sum_{t=1}^{T}(\hat{Q}_t^u - \hat{Q}_t^l)$ (regression) or PI-Size $= \frac{1}{T}\sum_{t=1}^{T}|\hat{C}_t|$ (classification).
- **Mean Interval Score (MIS)** (Gneiting & Raftery, 2007; Allen et al., 2025), the primary composite metric for regression:

$$\text{MIS} = \frac{1}{T}\sum_{t=1}^{T}\big[(\hat{Q}_t^u - \hat{Q}_t^l) + \tfrac{2}{\alpha}(Y_t - \hat{Q}_t^u)_+ + \tfrac{2}{\alpha}(\hat{Q}_t^l - Y_t)_+\big],\ (x)_+ = \max(0, x). \quad (84)$$

  MIS jointly penalizes coverage violations and excessive width; lower is better.
- **Mean Dice Similarity Coefficient (MDSC)** (Dice, 1945; Taha & Hanbury, 2015), the primary composite metric for classification:

$$\text{MDSC} = \frac{1}{T}\sum_{t=1}^{T}\frac{2|\hat{C}_t \cap \{Y_t\}|}{|\hat{C}_t| + |\{Y_t\}|}. \quad (85)$$

  MDSC $\in [0, 1]$ simultaneously rewards coverage accuracy and set compactness; higher is better.

Hyperparameter search ranges. Table 4 lists the search spaces for all OCI algorithms and their extensions, following Optuna conventions (Akiba et al., 2019).

D.2 Details for Financial Time Series Prediction

Datasets. We use daily opening prices of four global assets: Bitcoin[1], U.S. Dollar Index Futures (DX)[2], NASDAQ Composite (IXIC)[3], and Gold Spot (XAU) (Zhang et al., 2022). The sample period spans from January 1, 2014 to October 31, 2025, yielding 4322, 3058, 2977, and 3079 observations, respectively. Bitcoin and IXIC are highly volatile with severe distribution shifts; XAU exhibits moderate, persistent drift; DX is the calmest series. This diversity stress-tests OCI algorithms across three distinct non-stationarity regimes.

Point predictors. Three models are trained on rolling windows of log-returns $r_t = \log(P_t/P_{t-1})$. Predicted log-returns $\hat{r}_t$ are converted back to prices via $\hat{P}_t = P_{t-1}\exp(\hat{r}_t)$.

- **LightGBM** (Ke et al., 2017): gradient boosting (10 estimators, max depth 5) using 30-day lagged prices, moving average, standard deviation, minimum, maximum, and rate-of-change as features; retrained every 100 steps.

[1]Bitcoin data source: `https://cn.investing.com/crypto/bitcoin`
[2]DX data source: `https://cn.investing.com/currencies/us-dollar-index`
[3]IXIC data source: `https://cn.investing.com/indices/nasdaq-composite`

- **Theta** (Assimakopoulos & Nikolopoulos, 2000): seasonal decomposition with a period of 7 (Bitcoin) or 5 (others); re-estimated at every step.
- **Prophet** (Taylor & Letham, 2018): additive Bayesian model incorporating weekly and yearly seasonality; retrained every 100 steps.

Nonconformity scores. The p-order normalized residual score (Cordier et al., 2023) is adopted for constructing asymmetric prediction intervals:

$$s_{t,p} = \left| \frac{Y_t - \hat{P}_t}{\hat{P}_t^{p/2}} \right|, \quad p \in \mathbb{R}, \tag{86}$$

which yields scale-adaptive intervals with separate bounds:

$$\hat{C}_{t,p} = \left[\hat{P}_t - Q_t^- \cdot \hat{P}_t^{p/2},\ \hat{P}_t + Q_t^+ \cdot \hat{P}_t^{p/2} \right], \tag{87}$$

where Q_t^- and Q_t^+ are dynamic thresholds updated via the OCI algorithm. Setting $p = 0$ recovers the absolute residual, while the present work fixes $p = 1$ throughout all financial experiments.

Dataset split. Each series is split at January 1, 2025 into a calibration set (before) and a test set (after), with the calibration set reserved for hyperparameter tuning. After tuning, all the compared algorithms are re-initialized on the full calibration set for warm-up and evaluated online on the test set under an identical protocol.

D.3 Details for Medical Image Classification

Datasets. We use two large-scale medical image collections: DERM12345 (Yilmaz et al., 2024) (12,345 dermoscopic images across 40 skin lesion classes) and a benchmark MRI dataset (Ajmain et al., 2024) (34,192 images covering 40 disease classes, including brain, neurological, and spinal conditions). Both datasets are highly class-imbalanced, enabling realistic simulation of distribution drift in online medical diagnosis.

Base predictors. Three convolutional architectures pretrained on ImageNet are fine-tuned using dual NVIDIA T4 GPUs with automatic mixed precision: ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), and ResNeXt-50 (32×4d) (Xie et al., 2017). For each network, the classification head is replaced and trained for 50 epochs with the Adam optimizer (batch size 64, initial learning rate 10^{-4}, weight decay 10^{-4}), a cosine annealing scheduler, and early stopping (patience 8). Focal Loss (Lin et al., 2017) ($\gamma = 2$) with class-balanced α weights mitigates the severe class imbalance. After training, temperature scaling (Guo et al., 2017) is applied on a validation subset to calibrate the predicted probabilities.

Nonconformity scores. The Regularized Adaptive Prediction Set (RAPS) score (Angelopoulos et al., 2021) is used. For an input $\boldsymbol{x}$ with softmax probabilities $\hat{\boldsymbol{\pi}}_x$ and true label y, let $\rho_x(y) = \sum_{y':\hat{\pi}_x(y')>\hat{\pi}_x(y)} \hat{\pi}_x(y')$ denote the cumulative probability of classes ranked above y, and let $o_x(y)$ be the rank of y. The RAPS score is

$$s(\boldsymbol{x}, y, u) = \rho_x(y) + \hat{\pi}_x(y) \cdot u + \lambda_{\text{raps}} \cdot \max(0,\ o_x(y) - k_{\text{reg}}), \tag{88}$$

where $u \sim \text{Uniform}(0, 1)$ handles discreteness, and $\lambda_{\text{raps}}, k_{\text{reg}} \geq 0$ are regularization parameters. These parameters are tuned via grid search over $\lambda_{\text{raps}} \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5\}$ and $k_{\text{reg}} \in \{1, 2, 5, 10, 20, 50, 100\}$ to minimize the average prediction set size on a parameter-selection subset while maintaining the target coverage; Table 5 reports the optimal values.

Data split and tuning procedure. Stratified sampling partitions each dataset into 80% training and 20% test subsets. For RAPS parameter tuning, the training data is further split using stratified sampling into a parameter-selection set (20%) and a calibration set (80%). The grid search is performed exclusively on the parameter-selection set; the resulting optimal λ_{raps} and k_{reg} are then used to compute calibration scores and conformal thresholds on the full calibration set.

Table 5: Optimal RAPS regularization parameters.

Dataset	Architecture	λ_{raps}	k_{reg}
DERM12345	ResNet-50	0.1	50
	DenseNet-121	0.05	5
	ResNeXt-50	0.001	10
MRI	ResNet-50	0.05	10
	DenseNet-121	0.01	2
	ResNeXt-50	0.1	5

Drift simulation. To emulate non-stationary clinical workflows, we construct a fixed test stream for each dataset. For DERM12345, common and rare skin lesion classes are presented in an alternating sequence; for the MRI dataset, brain-related and non-brain disease classes alternate sequentially. The identical test streams are used across all algorithms, guaranteeing a fair and reproducible evaluation.

Reproducibility. Due to the stochastic nature of the RAPS score, all reported results are the mean and standard deviation over 5 independent runs. Within each run, the test stream order is fixed and all random seeds (data splitting, model training, and score computation) are set to 42. The grid search itself uses a dedicated base random state to ensure consistent parameter comparisons, while seeds vary across the 5 runs to honestly capture performance variability.

D.4 Financial Time Series Results

Tables 6 and 7 report the MIS performance of the 12 OCI algorithms on the four financial series using Theta and Prophet predictors, respectively.

Results with Theta. As shown in Table 6, DCOGD achieves two best results (DX and IXIC), ε-HQ-OGD ranks top-two in three cases (best on Bitcoin, second-best on DX and XAU), and ε-HQ-DCOGD obtains the best on XAU and second-best on Bitcoin and IXIC. Among the nine baselines, DtACI improves upon ACI but remains inferior to direct threshold-update methods. PIDCP attains high coverage on the highly volatile Bitcoin and IXIC at the cost of the largest interval widths. In contrast, ECI produces narrow intervals on Bitcoin and IXIC and the narrowest on XAU, yet its coverage accuracy degrades substantially, indicating that the error-quantification term improves interval efficiency but is limited by gradient gain mismatch. LQT suffers clear coverage failure on Bitcoin and IXIC, revealing its limitation under non-stationary fluctuations. The remaining OGD-type baselines (OGD, SF-OGD, Decay-OGD) perform comparably and none enters the top-two.

Results with Prophet. Table 7 reports that ε-HQ-OGD achieves the best MIS on Bitcoin, DX, and XAU, while ε-HQ-DCOGD ranks second on Bitcoin and IXIC, further confirming that the continuous gradient from the ε-HQ loss substantially improves coverage. Correspondingly, ACI attains the best MIS on IXIC, LQT the second-best on low-volatility DX, and COP the second-best on XAU. PIDCP generally yields wider intervals across all four datasets, and ECI performs

Table 6: Performance comparison of 12 OCI algorithms on four financial time series datasets with Theta. The **best** and second-best results for the MIS metric are bolded and underlined in red.

Algorithm	Bitcoin			DX			IXIC			XAU		
	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
ACI	87.136	2476.018	3648.764	89.686	1.441	1.904	89.499	480.773	663.510	87.676	52.868	78.625
DtACI	88.842	2514.705	3607.181	88.570	1.376	1.884	88.526	448.241	655.836	88.281	52.749	77.028
OGD	89.671	2412.164	3560.482	89.686	1.423	1.882	89.395	454.606	647.710	88.684	50.693	75.613
SF-OGD	89.671	2431.707	3558.664	89.618	1.380	1.892	89.430	448.610	651.020	88.314	48.921	76.494
Decay-OGD	89.813	2439.852	3569.320	89.855	1.400	1.899	89.638	448.733	653.125	88.415	48.920	76.573
PIDCP	93.556	2655.187	4026.875	89.246	1.517	2.168	91.168	506.289	757.159	88.684	52.203	86.569
ECI	85.454	2357.167	3577.430	88.502	1.395	1.890	85.709	438.779	652.344	84.990	45.881	76.534
LQT	44.800	515.122	8396.062	89.686	1.427	1.885	63.665	185.202	1008.428	88.314	49.487	76.295
COP	88.510	2383.340	3574.439	88.164	1.366	1.892	88.352	450.723	644.697	88.516	52.417	75.619
DCOGD	88.699	2336.967	3587.314	87.454	1.306	**1.854**	88.804	436.706	**630.776**	88.852	51.650	75.760
ε-HQ-OGD	88.297	2320.906	**3553.846**	89.716	1.400	1.882	89.534	454.960	644.797	88.751	51.116	75.526
ε-HQ-DCOGD	89.600	2439.278	3558.182	88.367	1.357	1.883	89.673	460.297	644.337	89.154	51.526	**75.523**

Table 7: Performance comparison of 12 OCI algorithms on four financial time series datasets with Prophet. The **best** and second-best results for the MIS metric are bolded and underlined in red.

Algorithm	Bitcoin			DX			IXIC			XAU		
	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
ACI	87.752	2562.683	3590.868	89.652	1.447	1.880	88.700	445.300	**630.787**	88.247	53.315	76.644
DtACI	88.605	2452.127	3533.588	88.570	1.362	1.862	88.526	437.081	639.450	88.281	52.019	74.914
OGD	89.623	2465.468	3498.349	89.550	1.402	1.850	89.395	440.586	632.903	89.187	51.746	73.839
SF-OGD	89.671	2448.005	3498.929	89.787	1.385	1.855	89.430	433.873	634.738	88.852	49.893	74.345
Decay-OGD	89.860	2475.087	3507.901	90.024	1.411	1.865	89.638	429.173	636.424	88.919	49.891	74.534
PIDCP	93.556	2683.140	3964.375	88.874	1.476	2.153	87.239	473.757	778.424	88.583	52.989	86.706
ECI	86.164	2457.517	3514.806	88.299	1.352	1.853	86.335	433.418	638.491	85.594	46.271	74.888
LQT	43.592	486.242	8278.322	89.550	1.401	1.850	63.561	188.669	969.553	88.146	49.311	75.178
COP	87.870	2364.969	3513.358	87.792	1.317	1.856	87.448	421.936	634.522	88.751	52.129	73.805
DCOGD	89.552	2421.824	3557.778	88.062	1.363	1.883	89.082	444.754	645.775	88.314	49.863	75.826
ε-HQ-OGD	89.315	2393.302	**3486.548**	89.618	1.399	**1.850**	89.534	444.661	631.639	89.255	52.334	**73.656**
ε-HQ-DCOGD	89.315	2388.768	3486.916	88.705	1.352	1.851	89.360	442.671	631.028	88.818	50.119	73.977

Table 8: Performance comparison of 12 OCI algorithms on two medical image datasets with DenseNet-121. The **best** and second-best results for the MDSC metric are bolded and underlined in red.

Algorithm	DERM12345			MRI		
	Cov(%)	PI-Size ↓	MDSC ↑	Cov(%)	PI-Size ↓	MDSC ↑
ACI	89.839 ± 0.619	2.267 ± 0.027	0.657 ± 0.003	89.906 ± 0.300	1.022 ± 0.004	0.864 ± 0.003
DtACI	89.640 ± 0.573	2.275 ± 0.020	0.660 ± 0.003	89.727 ± 0.293	1.021 ± 0.004	0.862 ± 0.002
OGD	89.662 ± 0.553	2.277 ± 0.025	0.656 ± 0.003	89.849 ± 0.273	1.021 ± 0.004	0.863 ± 0.002
SF-OGD	89.638 ± 0.665	2.201 ± 0.033	0.658 ± 0.003	89.946 ± 0.226	1.021 ± 0.003	0.864 ± 0.002
Decay-OGD	89.463 ± 0.635	2.165 ± 0.034	0.658 ± 0.002	89.789 ± 0.214	1.018 ± 0.002	0.863 ± 0.002
PIDCP	77.064 ± 0.662	1.913 ± 0.034	0.617 ± 0.003	76.258 ± 0.322	1.038 ± 0.010	0.729 ± 0.003
ECI	87.691 ± 0.595	2.173 ± 0.032	0.659 ± 0.004	88.605 ± 0.272	1.001 ± 0.003	0.853 ± 0.002
LQT	89.407 ± 0.652	2.302 ± 0.023	0.653 ± 0.003	89.635 ± 0.321	1.034 ± 0.011	0.860 ± 0.003
COP	89.675 ± 0.612	2.337 ± 0.026	0.655 ± 0.002	89.789 ± 0.250	1.021 ± 0.004	0.862 ± 0.002
DCOGD	89.054 ± 0.521	2.279 ± 0.032	0.656 ± 0.002	89.534 ± 0.301	1.023 ± 0.005	0.860 ± 0.003
ε-HQ-OGD	87.471 ± 0.585	2.075 ± 0.016	**0.663 ± 0.004**	92.584 ± 0.239	1.067 ± 0.004	**0.884 ± 0.002**
ε-HQ-DCOGD	87.346 ± 0.560	2.059 ± 0.018	0.663 ± 0.004	92.190 ± 0.276	1.075 ± 0.006	0.880 ± 0.002

consistently with its Theta-based results. The remaining baselines show no notable advantage over the proposed methods.

Overall, the proposed algorithms achieve superior coverage performance compared with the baselines, confirming the findings reported in Section 6.1.

D.5 Medical Image Classification Results

Tables 8 and 9 report the MDSC performance on the two medical image datasets with DenseNet-121 and ResNeXt-50 as base predictors, respectively.

Results with DenseNet-121. Table 8 reveals that ε-HQ-OGD achieves the best MDSC on both datasets, while ε-HQ-DCOGD obtains the second-best on both. Among the baselines, PIDCP again exhibits a clear trade-off, attaining high set efficiency at the cost of noticeably lower coverage accuracy; in contrast, the proposed methods maintain a superior balance between the two objectives.

Results with ResNeXt-50. As presented in Table 9, ε-HQ-DCOGD attains the best MDSC on the MRI dataset, and DCOGD and ε-HQ-OGD exhibit stable coverage performance. Among the baselines, SF-OGD achieves two second-best MDSC results, and Decay-OGD obtains the best MDSC on DERM12345; PIDCP remains the least competitive, and ECI's coverage accuracy is slightly below the average. These supplementary results are fully consistent with the main findings reported in Section 6.2.

Table 9: Performance comparison of 12 OCI algorithms on two medical image datasets with ResNext-50. The **best** and second-best results for the MDSC metric are bolded and underlined in red.

Algorithm	DERM12345			MRI		
	Cov(%)	PI-Size ↓	MDSC ↑	Cov(%)	PI-Size ↓	MDSC ↑
ACI	89.976 ± 0.402	2.085 ± 0.011	0.722 ± 0.003	89.885 ± 0.363	1.025 ± 0.004	0.863 ± 0.003
DtACI	89.750 ± 0.508	2.052 ± 0.012	0.722 ± 0.004	89.745 ± 0.319	1.025 ± 0.003	0.862 ± 0.003
OGD	89.669 ± 0.597	2.028 ± 0.016	0.719 ± 0.004	89.845 ± 0.341	1.029 ± 0.005	0.862 ± 0.003
SF-OGD	89.758 ± 0.406	1.914 ± 0.015	0.726 ± 0.003	89.890 ± 0.278	1.026 ± 0.004	0.863 ± 0.002
Decay-OGD	89.675 ± 0.504	1.883 ± 0.015	**0.727 ± 0.003**	89.796 ± 0.246	1.026 ± 0.003	0.862 ± 0.002
PIDCP	76.491 ± 0.700	1.731 ± 0.036	0.655 ± 0.004	76.501 ± 0.346	1.088 ± 0.011	0.727 ± 0.003
ECI	88.281 ± 0.569	1.922 ± 0.018	0.722 ± 0.004	88.648 ± 0.347	1.007 ± 0.004	0.853 ± 0.003
LQT	89.473 ± 0.512	2.047 ± 0.021	0.715 ± 0.003	89.629 ± 0.324	1.047 ± 0.012	0.859 ± 0.003
COP	89.802 ± 0.436	1.977 ± 0.019	0.725 ± 0.002	89.634 ± 0.295	1.021 ± 0.003	0.861 ± 0.003
DCOGD	88.723 ± 0.581	1.953 ± 0.019	0.721 ± 0.004	89.549 ± 0.327	1.024 ± 0.004	0.860 ± 0.003
ε-HQ-OGD	89.527 ± 0.533	2.004 ± 0.024	0.720 ± 0.003	89.753 ± 0.341	1.024 ± 0.003	0.862 ± 0.003
ε-HQ-DCOGD	89.644 ± 0.540	1.968 ± 0.015	0.723 ± 0.003	92.079 ± 0.234	1.068 ± 0.003	**0.880 ± 0.002**

Table 10: MIS of 12 OCI algorithms on four financial time series datasets with different α values. For each dataset-predictor pair, the **best** and second-best results are bolded and underlined in red.

Miscoverage rate	Algorithm	Bitcoin			DX			IXIC			XAU		
		LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet
$\alpha = 0.05$	ACI	4522.899	4621.058	4517.581	2.261	2.300	2.261	762.385	793.832	768.625	94.706	96.122	91.741
	DtACI	4455.069	4580.910	4417.599	2.245	2.289	2.239	775.128	803.585	779.357	92.991	95.093	92.601
	OGD	4354.934	4429.448	4273.401	2.193	2.248	2.194	**753.818**	786.641	774.740	89.697	91.570	89.672
	SF-OGD	4327.072	4437.644	4286.864	2.203	2.274	2.205	785.204	807.538	781.976	91.069	92.755	91.033
	Decay-OGD	4380.710	4451.070	4346.352	2.231	2.301	2.239	794.346	820.872	795.257	91.814	93.274	90.780
	PIDCP	4881.382	4888.632	4826.055	2.615	2.641	2.653	923.925	936.857	988.501	109.782	108.247	108.979
	ECI	4362.856	4437.638	4316.308	2.208	2.260	2.208	777.015	795.473	766.118	95.477	96.013	93.348
	LQT	15839.274	16085.877	15728.696	**2.190**	2.253	2.199	1730.739	1823.549	1744.651	95.214	96.951	95.400
	COP	4350.953	4448.566	4311.775	2.212	2.266	2.215	754.569	789.718	**758.948**	89.729	91.797	89.357
	DCOGD	4314.747	4442.983	**4253.897**	2.193	2.252	2.197	755.822	786.553	767.656	**89.472**	**91.256**	89.544
	ε-HQ-OGD	4312.567	**4406.801**	4267.432	2.199	**2.248**	**2.194**	763.182	795.352	762.526	89.643	91.724	88.628
	ε-HQ-DCOGD	**4303.529**	4412.515	4266.793	2.201	2.248	2.194	765.914	**785.703**	763.761	90.110	91.278	**88.437**
$\alpha = 0.20$	ACI	2790.543	2811.716	2798.624	1.533	1.571	1.545	507.106	519.208	510.640	60.614	61.321	60.945
	DtACI	2786.024	2799.775	2764.314	1.529	1.568	1.533	495.962	511.094	500.676	59.919	61.040	59.960
	OGD	2756.902	2777.877	2751.189	**1.512**	1.551	1.517	494.227	507.808	498.793	59.455	60.110	59.497
	SF-OGD	2764.412	2776.458	2756.577	1.518	1.555	1.521	496.424	509.880	499.650	59.740	60.554	59.574
	Decay-OGD	2760.486	2777.368	2752.266	1.517	1.556	1.520	497.100	510.170	499.249	59.731	60.716	59.539
	PIDCP	4790.765	4863.583	4766.230	1.696	1.738	1.718	573.904	652.337	595.271	67.466	67.186	67.286
	ECI	2771.420	2789.365	2749.096	1.517	1.556	1.521	499.720	514.191	504.691	59.692	60.568	59.538
	LQT	4481.973	4538.567	4426.817	1.512	1.549	1.516	575.981	597.413	581.334	59.467	60.278	59.469
	COP	2764.587	2790.289	2756.879	1.520	1.557	1.519	493.938	508.173	500.211	59.412	60.083	59.459
	DCOGD	2757.117	**2769.820**	2744.958	1.516	1.556	1.520	494.545	507.445	498.239	59.439	60.192	59.444
	ε-HQ-OGD	2750.469	2786.214	**2737.903**	1.513	1.550	1.516	**493.340**	506.652	497.973	59.346	60.059	**59.202**
	ε-HQ-DCOGD	**2749.398**	2776.278	2738.414	1.512	**1.550**	**1.514**	493.784	**506.356**	**497.837**	**59.346**	**59.995**	59.220

D.6 Multi-quantile Comparison

Tables 10 and 11 report the MIS and MDSC results of 12 OCI algorithms under $\alpha = 0.05$ and $\alpha = 0.2$ on financial time series and medical image datasets, respectively. Main results for $\alpha = 0.1$ are reported in Tables 1–2 of the main text and Tables 6–9 of Appendix D.

Financial time series. For the 12 financial tasks (4 datasets $\times$ 3 base predictors), each task contributes one best and one second-best placement, giving 24 top-two placements in total. Under $\alpha = 0.05$, the proposed algorithms (DCOGD, ε-HQ-OGD, and ε-HQ-DCOGD) collectively obtain 18 of these placements, while the baselines obtain 6. Under $\alpha = 0.20$, the proposed algorithms obtain 22 placements and the baselines 2. In both settings, the proposed algorithms account for the majority of top-two placements, with ε-HQ-OGD and ε-HQ-DCOGD contributing most of them, particularly on the highly volatile Bitcoin and IXIC datasets. Among the baselines, OGD, LQT, and COP appear in the top two, each with at most four placements across both settings; LQT performs well on the low-volatility DX but ranks lower on Bitcoin and IXIC. PIDCP obtains no top-two placement in either setting and consistently ranks below the average baseline performance.

Medical image classification. For the 6 medical tasks (2 datasets $\times$ 3 base predictors), there are 12 top-two placements in total. Under $\alpha = 0.05$, the proposed algorithms obtain 7 placements and the baselines 5. Among the nine baselines, Decay-OGD achieves two best results on MRI (ResNet-50, DenseNet-121) and one second-best on ResNeXt-50, and SF-OGD attains two second-best results

Table 11: MDSC of 12 OCI algorithms on two medical image datasets with different α values. For each dataset-predictor pair, the **best** and second-best results are bolded and underlined in red.

Miscoverage rate	Algorithm	DERM12345			MRI		
		ResNet-50	DenseNet-121	ResNeXt-50	ResNet-50	DenseNet-121	ResNeXt-50
$\alpha = 0.05$	ACI	0.623 ± 0.001	0.631 ± 0.002	0.691 ± 0.002	0.895 ± 0.001	0.899 ± 0.001	0.900 ± 0.001
	DtACI	0.632 ± 0.001	0.627 ± 0.001	0.694 ± 0.001	0.892 ± 0.002	0.896 ± 0.001	0.897 ± 0.001
	OGD	0.623 ± 0.001	0.620 ± 0.002	0.689 ± 0.001	0.894 ± 0.001	0.899 ± 0.001	0.899 ± 0.002
	SF-OGD	0.631 ± 0.001	0.625 ± 0.001	0.694 ± 0.002	0.896 ± 0.001	0.901 ± 0.001	0.900 ± 0.002
	Decay-OGD	0.665 ± 0.002	0.659 ± 0.002	0.693 ± 0.002	**0.897 ± 0.002**	**0.902 ± 0.001**	0.901 ± 0.002
	PIDCP	0.627 ± 0.004	0.618 ± 0.004	0.656 ± 0.004	0.725 ± 0.003	0.730 ± 0.003	0.729 ± 0.004
	ECI	0.662 ± 0.003	0.659 ± 0.004	0.718 ± 0.002	0.885 ± 0.002	0.890 ± 0.001	0.890 ± 0.002
	LQT	0.611 ± 0.002	0.610 ± 0.003	0.669 ± 0.005	0.878 ± 0.003	0.883 ± 0.004	0.882 ± 0.003
	COP	0.634 ± 0.001	0.628 ± 0.001	0.697 ± 0.001	0.894 ± 0.001	0.899 ± 0.001	0.898 ± 0.002
	DCOGD	**0.668 ± 0.003**	**0.661 ± 0.004**	**0.723 ± 0.003**	0.890 ± 0.001	0.895 ± 0.001	0.894 ± 0.001
	ε-HQ-OGD	0.665 ± 0.004	0.655 ± 0.003	0.721 ± 0.004	0.895 ± 0.001	0.899 ± 0.001	**0.906 ± 0.001**
	ε-HQ-DCOGD	0.668 ± 0.003	0.657 ± 0.005	0.722 ± 0.003	0.891 ± 0.001	0.895 ± 0.002	0.895 ± 0.001
$\alpha = 0.20$	ACI	0.665 ± 0.003	0.651 ± 0.004	0.700 ± 0.004	0.775 ± 0.002	0.778 ± 0.003	0.777 ± 0.003
	DtACI	0.668 ± 0.004	0.652 ± 0.005	0.700 ± 0.004	0.775 ± 0.002	0.778 ± 0.003	0.777 ± 0.002
	OGD	0.666 ± 0.003	0.651 ± 0.005	0.701 ± 0.004	0.774 ± 0.002	0.777 ± 0.001	0.776 ± 0.003
	SF-OGD	0.668 ± 0.003	0.652 ± 0.005	0.705 ± 0.004	0.773 ± 0.002	0.778 ± 0.003	0.776 ± 0.003
	Decay-OGD	0.669 ± 0.003	0.652 ± 0.005	0.705 ± 0.005	0.773 ± 0.002	0.778 ± 0.003	0.776 ± 0.003
	PIDCP	0.627 ± 0.004	0.618 ± 0.004	0.654 ± 0.004	0.722 ± 0.003	0.726 ± 0.004	0.725 ± 0.004
	ECI	0.661 ± 0.004	0.646 ± 0.006	0.695 ± 0.004	0.766 ± 0.002	0.769 ± 0.002	0.769 ± 0.002
	LQT	0.664 ± 0.004	0.650 ± 0.006	0.699 ± 0.005	0.774 ± 0.002	0.775 ± 0.003	0.776 ± 0.002
	COP	0.668 ± 0.004	0.649 ± 0.006	0.701 ± 0.005	0.774 ± 0.003	0.777 ± 0.003	0.775 ± 0.003
	DCOGD	0.666 ± 0.003	0.650 ± 0.004	0.700 ± 0.004	0.773 ± 0.001	0.777 ± 0.002	0.776 ± 0.002
	ε-HQ-OGD	0.673 ± 0.003	**0.664 ± 0.004**	**0.721 ± 0.005**	**0.823 ± 0.003**	**0.825 ± 0.003**	**0.824 ± 0.002**
	ε-HQ-DCOGD	**0.675 ± 0.003**	0.662 ± 0.005	0.719 ± 0.005	0.816 ± 0.002	0.820 ± 0.003	0.821 ± 0.003

on MRI (ResNet-50, DenseNet-121). Under $\alpha = 0.20$, the proposed algorithms obtain all 12 placements: ε-HQ-OGD ranks best on five tasks and second-best on one, while ε-HQ-DCOGD ranks best on one and second-best on five; no baseline enters the top two. PIDCP remains ranked below the average level, reflecting its heavy dependence on the accuracy of the external score predictor.

In summary, across both coverage levels and both task types, the proposed algorithms account for the majority of top-two placements, demonstrating a superior trade-off between coverage accuracy and efficiency, and this advantage further expands under moderately relaxed target settings ($\alpha = 0.2$).

D.7 Efficiency Comparison

Table 12 reports the running time (in seconds) of the 12 OCI algorithms over 1,000 prediction rounds across six real-world datasets. The OGD family (OGD, SF-OGD, Decay-OGD) is the fastest, averaging 0.049–0.052 s. COP, ECI, and DCOGD form the next tier (0.056–0.068 s), with DCOGD incurring only 0.019 s additional overhead over OGD for momentum smoothing. LQT (0.090 s) and ACI (0.169 s) follow, the latter requiring an additional quantile–threshold transformation. The proposed ε-HQ-OGD and ε-HQ-DCOGD respectively take 0.649 s and 0.673 s, due to the added continuous-gradient and momentum-smoothing computations. DtACI (1.050 s) and PIDCP (1.824 s) are the slowest, owing to online expert aggregation and the scorecaster component, respectively.

Compared with OGD, the ε-HQ extensions increase runtime by about 0.6 s per 1,000 rounds (i.e., roughly 0.6 ms per prediction, negligible for real-time use), with ε-HQ-DCOGD adding only 0.024 s more than ε-HQ-OGD for momentum calibration. Despite being slower than the OGD family, the proposed methods remain much faster than DtACI and PIDCP while delivering markedly better coverage efficiency (Section 6.1 and Appendices D.4–D.5), thus achieving a favorable accuracy–efficiency trade-off for practical deployment.

D.8 Extended OCI Comparison

Tables 13–15 report the MIS of three base update rules (OGD, SF-OGD, Decay-OGD) under the four OCI schemes defined in Appendix C.2—the OCI baseline, DCOCI, ε-HQ-OCI, and ε-HQ-DCOCI. For each dataset and base predictor, we compare the three base update rules under each scheme and mark the best and second-best MIS within each base update rule across the four schemes, yielding $4 \times 3 \times 3 = 36$ trials in total. The baseline results are also provided in Table 1 and Tables 6–7.

Table 12: Running time comparison (in seconds) of 12 OCI algorithms over 1000 prediction rounds.

DataSet	ACI	DtACI	OGD	SF-OGD	Decay-OGD	PIDCP	ECI	LQT	COP	DCOGD	ε-HQ-OGD	ε-HQ-DCOGD
Bitcoin	0.207	1.356	0.046	0.052	0.052	2.483	0.060	0.114	0.051	0.076	0.772	0.803
DX	0.209	1.410	0.047	0.044	0.048	2.251	0.059	0.105	0.050	0.075	0.760	0.792
IXIC	0.209	1.284	0.042	0.047	0.051	2.337	0.076	0.100	0.058	0.075	0.765	0.801
XAU	0.209	1.305	0.039	0.051	0.051	2.298	0.067	0.109	0.053	0.077	0.758	0.795
DERM12345	0.087	0.481	0.060	0.057	0.055	0.800	0.064	0.060	0.062	0.054	0.419	0.425
MRI	0.092	0.466	0.059	0.055	0.056	0.772	0.065	0.053	0.064	0.053	0.422	0.423
Average time	0.169	1.050	0.049	0.051	0.052	1.824	0.065	0.090	0.056	0.068	0.649	0.673

Table 13: Performance comparison of 3 OCI algorithms under four schemes on four financial time series datasets with the LightGBM base predictor. The **best** and second-best results for the MIS metric are bolded and underlined in red.

Scheme	Algorithm	Bitcoin			DX			IXIC			XAU		
		Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
OCI	OGD	89.647	2431.156	3512.292	89.550	1.375	1.823	89.291	427.392	624.772	88.684	50.503	74.136
	SF-OGD	89.766	2450.432	3522.821	89.686	1.371	1.840	89.465	429.062	628.796	88.516	49.000	74.985
	Decay-OGD	89.860	2433.238	**3522.170**	89.888	1.373	1.845	89.569	434.387	632.212	88.885	49.903	74.595
DCOCI	OGD	89.718	2405.738	3513.560	86.879	1.278	1.830	89.326	435.338	624.854	88.079	49.383	74.207
	SF-OGD	89.576	2426.544	3519.081	87.758	1.330	1.845	89.465	434.554	629.119	88.012	48.212	75.003
	Decay-OGD	89.860	2449.783	3522.737	86.676	1.270	1.852	89.117	426.299	631.438	88.079	48.760	74.915
ε-HQ-OCI	OGD	89.078	2368.859	**3510.987**	89.280	1.352	1.824	90.021	444.509	**624.004**	89.490	52.211	**73.633**
	SF-OGD	88.865	2353.750	**3516.960**	88.502	1.329	1.839	89.778	434.992	628.588	89.355	50.900	**74.131**
	Decay-OGD	89.505	2420.703	3530.447	87.521	1.299	1.839	88.526	413.810	630.243	88.617	50.209	**74.190**
ε-HQ-DCOCI	OGD	89.173	2377.241	3512.139	88.739	1.341	**1.822**	90.230	447.119	624.794	88.919	50.076	73.937
	SF-OGD	89.505	2441.740	3564.566	88.671	1.335	**1.834**	90.056	439.920	**628.029**	88.449	49.304	74.561
	Decay-OGD	88.960	2364.134	3524.450	88.840	1.331	**1.839**	89.013	422.944	**628.845**	88.382	49.086	75.146

LightGBM. With LightGBM (Table 13), the OCI baseline achieves 4 top-two placements: OGD second-best on DX and IXIC, and Decay-OGD best on Bitcoin and second-best on XAU. The DCOCI variants achieve 2 placements: SF-OGD and Decay-OGD second-best on Bitcoin. The ε-HQ-OCI variants achieve 10 placements (6 best, 4 second-best): OGD best on Bitcoin, IXIC, and XAU; SF-OGD best on Bitcoin and XAU and second-best on DX and IXIC; Decay-OGD best on XAU and second-best on DX and IXIC. The full ε-HQ-DCOCI variants achieve 8 placements (5 best, 3 second-best): OGD best on DX and second-best on Bitcoin and XAU; SF-OGD best on DX and IXIC and second-best on XAU; Decay-OGD best on DX and IXIC. At the algorithm level, OGD attains 2, 0, 3, and 3 top-two placements across the four schemes, respectively; SF-OGD attains 0, 1, 4, and 3; and Decay-OGD attains 2, 1, 3, and 2. These results indicate that the continuous gradient from the ε-HQ loss substantially improves coverage, and the momentum correction provides further gains across datasets of different volatility.

Theta. With Theta (Table 14), the OCI baseline achieves only 1 top-two placement (SF-OGD best on Bitcoin). The DCOCI variants achieve 6 (3 best, 3 second-best): OGD best on DX and IXIC; SF-OGD second-best on Bitcoin; Decay-OGD best on Bitcoin and second-best on IXIC and XAU. The ε-HQ-OCI variants achieve 8 (4 best, 4 second-best): OGD best on Bitcoin and second-best on DX and XAU; SF-OGD best on IXIC and XAU and second-best on DX; Decay-OGD best on XAU and second-best on DX. The full ε-HQ-DCOCI variants achieve 9 (4 best, 5 second-best): OGD second-best on Bitcoin, IXIC, and XAU; SF-OGD best on DX and second-best on IXIC and XAU; Decay-OGD best on DX and IXIC and second-best on Bitcoin. This confirms that adding momentum correction to the baseline can improve coverage to some extent, while the coarse gradient of the standard quantile loss remains the core bottleneck.

Prophet. With Prophet (Table 15), the OCI baseline achieves 2 top-two placements (OGD and SF-OGD second-best on DX). The DCOCI variants obtain no top-two placements. The ε-HQ-OCI variants achieve 5 best and 6 second-best (11 total), and the full ε-HQ-DCOCI variants achieve 7 best and 4 second-best (11 total). This again demonstrates that both the continuous gradient and the momentum correction substantially enhance coverage performance, yielding a better trade-off between coverage accuracy and efficiency.

Collectively, across the 36 trials, the OCI baseline obtains 7 top-two placements (2 best, 5 second-best), the DCOCI variants obtain 8 (3 best, 5 second-best), the ε-HQ-OCI variants attain 29 (15 best, 14 second-best), and the full ε-HQ-DCOCI variants achieve 28 (16 best, 12 second-best). The OCI baseline performs relatively well on the calm DX dataset (3 placements), the DCOCI

Table 14: Performance comparison of 3 OCI algorithms under four schemes on four financial time series datasets with the Theta base predictor. The **best** and second-best results for the MIS metric are bolded and underlined in red.

Scheme	Algorithm	Bitcoin			DX			IXIC			XAU		
		Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
OCI	OGD	89.671	2412.164	3560.482	89.686	1.423	1.882	89.395	454.606	647.710	88.684	50.693	75.613
	SF-OGD	89.671	2431.707	**3558.664**	89.618	1.380	1.892	89.430	448.610	651.020	88.314	48.921	76.494
	Decay-OGD	89.813	2439.852	3569.320	89.855	1.400	1.899	89.638	448.733	653.125	88.415	48.920	76.573
DCOCI	OGD	88.699	2336.967	3587.314	87.454	1.306	**1.854**	88.804	436.706	**630.776**	88.852	51.650	75.760
	SF-OGD	89.671	2420.989	3562.299	87.183	1.303	1.899	89.430	453.462	650.149	88.046	48.983	76.177
	Decay-OGD	89.505	2408.872	**3565.982**	86.405	1.279	1.910	88.873	432.962	652.651	88.650	50.715	76.098
ε-HQ-OCI	OGD	88.297	2320.906	**3553.846**	89.716	1.400	1.882	89.534	454.960	644.797	88.751	51.116	75.526
	SF-OGD	88.865	2352.115	3563.125	88.536	1.352	1.890	88.213	423.273	**648.130**	89.691	51.342	**75.908**
	Decay-OGD	88.699	2320.887	3568.759	88.299	1.350	1.892	89.326	447.021	656.006	89.590	51.655	**75.779**
ε-HQ-DCOCI	OGD	89.600	2439.278	3558.182	88.367	1.357	1.883	89.673	460.297	644.337	89.154	51.526	**75.523**
	SF-OGD	89.339	2390.176	3565.378	88.434	1.337	**1.888**	88.004	423.422	648.293	88.617	50.006	76.071
	Decay-OGD	89.150	2362.444	3566.276	88.603	1.353	**1.889**	88.491	428.573	**648.521**	88.247	48.636	76.627

Table 15: Performance comparison of 3 OCI algorithms under four schemes on four financial time series datasets with the Prophet base predictor. The **best** and second-best results for the MIS metric are bolded and underlined in red.

Scheme	Algorithm	Bitcoin			DX			IXIC			XAU		
		Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
OCI	OGD	89.623	2465.468	3498.349	89.550	1.402	1.850	89.395	440.586	632.903	89.187	51.746	73.839
	SF-OGD	89.671	2448.005	3498.929	89.787	1.385	1.855	89.430	433.873	634.738	88.852	49.893	74.345
	Decay-OGD	89.860	2475.087	3507.901	90.024	1.411	1.865	89.638	429.173	636.424	88.919	49.891	74.534
DCOCI	OGD	89.552	2421.824	3557.778	88.062	1.363	1.883	89.082	444.754	645.775	88.314	49.863	75.826
	SF-OGD	89.694	2446.432	3501.852	88.367	1.341	1.861	89.221	434.066	634.021	88.482	49.348	74.324
	Decay-OGD	89.031	2361.442	3481.659	87.318	1.299	1.868	89.430	432.768	635.053	88.650	49.898	74.481
ε-HQ-OCI	OGD	89.315	2393.302	**3486.548**	89.618	1.399	**1.850**	89.534	444.661	631.639	89.255	52.334	**73.656**
	SF-OGD	89.600	2427.715	3491.344	87.859	1.320	1.872	88.352	416.746	632.218	88.751	49.506	**74.180**
	Decay-OGD	89.837	2435.246	3500.739	88.739	1.350	1.856	88.734	418.611	**632.851**	88.684	49.573	74.238
ε-HQ-DCOCI	OGD	89.315	2388.768	3486.916	88.705	1.352	1.851	89.360	442.671	**631.028**	88.818	50.119	73.977
	SF-OGD	89.031	2380.725	**3489.771**	88.637	1.346	**1.853**	88.317	415.128	**631.773**	88.247	48.928	74.188
	Decay-OGD	89.244	2380.415	**3500.240**	88.434	1.321	**1.854**	88.908	420.032	633.048	88.549	49.193	**74.023**

variants show clear advantages on the highly volatile Bitcoin and IXIC (6 placements), and the ε-HQ-OCI variants exhibit substantial gains across datasets of varying volatility. The full ε-HQ-DCOCI scheme further improves overall coverage, confirming that both the ε-HQ gradient module and the momentum calibration module contribute meaningfully to performance enhancement.

D.9 Empirical Comparison of Scaled and Composite Losses

We compare the proposed algorithms, built on the scaled loss, with their composite-loss counterparts. As analyzed in Appendix B.1, replacing the scaled loss by the composite loss yields the variants ε-HQ-OGD-CL and ε-HQ-DCOGD-CL: the scaled-loss algorithms update via $\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta g_t)$, whereas the composite-loss variants use $\hat{q}_{t+1} = \Pi_{\mathcal{Q}}(\hat{q}_t - \eta a_t^s g_t)$, where a_t^s is the slope of the adaptive residual mapping. Tables 16 and 17 report the coverage performance of the four OCI algorithms on financial time series and medical image datasets, respectively.

Financial time series. Table 16 contains 12 trials, each with a best and a second-best result, yielding 24 top-two entries in total. The scaled-loss algorithms attain all 24 positions: ε-HQ-OGD achieves 8 best and 4 second-best, while ε-HQ-DCOGD achieves 4 best and 8 second-best. In contrast, the composite-loss variants perform substantially worse across coverage accuracy (Cov), coverage efficiency (PI-Width), and the overall MIS metric. This degradation is consistent with Appendix B.1: the mapping factor a_t^s amplifies the update during calm periods and shrinks it under volatility, opposite to the direction required for adaptive tracking.

Medical image classification. Table 17 contains 6 trials (2 datasets $\times$ 3 base predictors), there are 12 top-two placements in total. The scaled-loss algorithms again occupy all 12 placements, splitting them evenly: ε-HQ-OGD and ε-HQ-DCOGD each achieve 3 best and 3 second-best. The composite-loss variants are markedly inferior, with pronounced coverage deviations for ε-HQ-OGD-CL on MRI and for ε-HQ-DCOGD-CL on DERM12345. This directly reflects the amplified Lipschitz and GWPC constants in the composite-loss formulation, which worsen the regret bound and, in turn, the coverage error bound.

Table 16: Performance comparison of scaled-loss and composite-loss variants on four financial time series datasets. For each dataset–predictor pair, the **best** and second-best MIS results are bolded and underlined in red.

Algorithm	Model	Bitcoin			DX			IXIC			XAU		
		Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
ε-HQ-OGD	LightGBM	89.078	2368.859	**3510.987**	89.280	1.352	1.824	90.021	444.509	**624.004**	89.490	52.211	**73.633**
	Theta	88.297	2320.906	**3553.846**	89.716	1.400	**1.882**	89.534	454.960	644.797	88.751	51.116	75.526
	Prophet	89.315	2393.302	**3486.548**	89.618	1.399	**1.850**	89.534	444.661	631.639	89.255	52.334	**73.656**
ε-HQ-DCOGD	LightGBM	89.173	2377.241	3512.139	88.739	1.341	**1.822**	90.230	447.119	624.794	88.919	50.076	73.937
	Theta	89.600	2439.278	3558.182	88.367	1.357	1.883	89.673	460.297	**644.337**	89.154	51.526	**75.523**
	Prophet	89.315	2388.768	3486.916	88.705	1.352	1.851	89.360	442.671	**631.028**	88.818	50.119	73.977
ε-HQ-OGD-CL	LightGBM	92.158	2612.728	3734.922	88.570	1.628	2.224	87.378	478.387	670.106	89.456	56.838	78.628
	Theta	88.415	2127.047	3840.315	89.381	1.706	2.304	86.509	451.678	687.304	88.382	55.463	79.465
	Prophet	94.219	2184.625	3728.457	87.115	1.601	2.357	90.473	561.827	713.680	86.098	51.207	80.862
ε-HQ-DCOGD-CL	LightGBM	92.727	2781.250	3748.478	85.830	1.976	2.640	85.327	404.975	640.632	91.437	108.829	125.996
	Theta	92.040	2499.395	3722.788	85.932	1.441	2.109	88.178	437.703	681.380	87.139	52.606	80.748
	Prophet	91.637	2477.331	3749.777	89.888	2.473	2.918	86.474	407.337	664.525	91.605	102.051	117.493

Table 17: Performance comparison of scaled-loss and composite-loss variants on two medical image datasets. For each dataset–predictor pair, the **best** and second-best MDSC results are bolded and underlined in red.

Algorithm	Model	DERM12345			MRI		
		Cov(%)	PI-Size ↓	MDSC ↑	Cov(%)	PI-Size ↓	MDSC ↑
ε-HQ-OGD	ResNet-50	89.613 ± 0.471	2.272 ± 0.030	0.663 ± 0.002	92.489 ± 0.183	1.083 ± 0.002	$\mathbf{0.879 \pm 0.002}$
	DenseNet-121	87.471 ± 0.585	2.075 ± 0.016	$\mathbf{0.663 \pm 0.004}$	92.584 ± 0.239	1.067 ± 0.004	$\mathbf{0.884 \pm 0.002}$
	ResNext-50	89.527 ± 0.533	2.004 ± 0.024	0.720 ± 0.003	89.753 ± 0.341	1.024 ± 0.003	0.862 ± 0.003
ε-HQ-DCOGD	ResNet-50	87.271 ± 0.328	2.003 ± 0.027	$\mathbf{0.673 \pm 0.002}$	91.977 ± 0.233	1.074 ± 0.003	0.876 ± 0.002
	DenseNet-121	87.346 ± 0.560	2.059 ± 0.018	0.663 ± 0.004	92.190 ± 0.276	1.075 ± 0.006	0.880 ± 0.002
	ResNext-50	89.644 ± 0.540	1.968 ± 0.015	$\mathbf{0.723 \pm 0.003}$	92.079 ± 0.234	1.068 ± 0.003	$\mathbf{0.880 \pm 0.002}$
ε-HQ-OGD-CL	ResNet-50	83.509 ± 0.765	3.856 ± 0.720	0.552 ± 0.004	80.053 ± 0.442	2.940 ± 0.279	0.610 ± 0.017
	DenseNet-121	83.966 ± 1.150	3.543 ± 0.942	0.548 ± 0.025	80.638 ± 0.489	3.005 ± 0.277	0.611 ± 0.016
	ResNext-50	85.723 ± 0.702	3.819 ± 0.879	0.562 ± 0.040	80.520 ± 0.372	3.190 ± 0.181	0.597 ± 0.010
ε-HQ-DCOGD-CL	ResNet-50	80.661 ± 0.778	1.651 ± 0.036	0.661 ± 0.003	91.659 ± 1.607	3.410 ± 3.916	0.791 ± 0.099
	DenseNet-121	80.912 ± 0.945	1.710 ± 0.047	0.651 ± 0.005	89.383 ± 0.929	1.447 ± 0.166	0.823 ± 0.008
	ResNext-50	85.483 ± 1.795	1.923 ± 0.797	0.712 ± 0.037	90.092 ± 0.753	1.265 ± 0.249	0.846 ± 0.013

In summary, the empirical results for the composite-loss variants align with the theoretical analysis in Appendix B.1. The fixed-learning-rate scaled-loss formulation preserves the full adaptivity of the residual mapping mechanism, yielding a better accuracy–efficiency trade-off in coverage.

E Ablation Studies and Sensitivity Analysis

This appendix supplements Section 6.1 with an ablation study on the loss structure and adaptive mechanisms (Appendix E.1), an evaluation of alternative insensitive-zone tuning mechanisms (Appendix E.2), and a sensitivity analysis of the key parameters τ, Γ, W, and k (Appendix E.3).

E.1 Ablation Study on Loss Structure and Adaptive Mechanisms

We conduct an ablation study to isolate the contribution of the hybrid structure, adaptive residual mapping, and insensitive-zone fine-tuning by evaluating the following nine configurations on the four financial datasets with ε-HQ-DCOGD (Table 18):

- $(\tau = 1, \varepsilon = 0, \Gamma = 0)$: standard pinball loss L_{Q} (baseline).
- $(\tau = 1, \varepsilon \neq 0, \Gamma = 0)$: insensitive pinball loss $L_{\varepsilon\text{-Q}}$, evaluating the effect of the insensitive zone on L_{Q}.
- $(\tau = 0, \varepsilon = 0, \Gamma = 0)$: bounded asymmetric least squares loss L_{BALS}, representing fully smooth bounded losses.
- $(\tau = 0, \varepsilon = 0, \Gamma \neq 0)$: L_{BALS} with a default residual mapping range $[-10, 10]$.
- $(\tau = 0, \varepsilon \neq 0, \Gamma = 0)$: insensitive bounded loss $L_{\varepsilon\text{-BALS}}$, evaluating the effect of the insensitive zone on L_{BALS}.
- $(\tau = 0, \varepsilon \neq 0, \Gamma \neq 0)$: $L_{\varepsilon\text{-BALS}}$ with residual mapping to $[-10 - \varepsilon_{\min}, 10 + \varepsilon_{\min}]$.

Table 18: MIS of ε-HQ-DCOGD under 10 combinations of τ, ε, and Γ on four financial time series datasets. For each dataset-predictor pair, the **best** and second-best results are bolded and underlined in red.

Loss	Configuration	Bitcoin			DX			IXIC			XAU		
		LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet
L_{Q}	$\tau=1,\varepsilon=0,\Gamma=0$	3513.560	3587.314	3557.778	1.830	1.854	1.883	624.854	630.776	645.775	74.207	75.760	75.826
$L_{\varepsilon\text{-Q}}$	$\tau=1,\varepsilon\neq0,\Gamma=0$	3513.560	**3556.941**	3487.314	1.829	1.883	1.852	624.854	645.775	**630.776**	74.107	75.828	**73.760**
L_{BALS}	$\tau=0,\varepsilon=0,\Gamma=0$	5831.402	5924.148	5887.755	2.138	2.185	2.134	756.661	781.147	762.342	83.172	84.825	82.653
L_{BALS}	$\tau=0,\varepsilon=0,\Gamma\neq0$	4789.748	4889.860	4772.604	2.716	2.802	2.707	845.726	867.553	849.915	105.620	106.803	104.029
$L_{\varepsilon\text{-BALS}}$	$\tau=0,\varepsilon\neq0,\Gamma=0$	10831.366	5913.776	5902.731	2.138	2.185	2.134	756.661	781.147	762.342	83.172	84.825	82.653
$L_{\varepsilon\text{-BALS}}$	$\tau=0,\varepsilon\neq0,\Gamma\neq0$	4789.904	4889.915	4773.392	2.715	2.805	2.708	844.489	868.559	850.167	105.337	106.831	104.187
L_{HQ}	$\tau\in(0,1),\varepsilon=0,\Gamma=0$	3531.487	3575.866	3515.173	1.830	1.892	1.855	625.192	646.077	634.449	74.225	75.984	73.836
L_{HQ}	$\tau\in(0,1),\varepsilon=0,\Gamma\neq0$	3514.212	3566.592	**3482.932**	1.824	**1.883**	1.851	**624.403**	644.661	630.791	**73.724**	**75.329**	73.977
$L_{\varepsilon\text{-HQ}}$	$\tau\in(0,1),\varepsilon\neq0,\Gamma=0$	3541.060	3580.511	3522.180	1.830	1.886	1.850	630.886	652.762	637.300	74.233	75.984	73.836
$L_{\varepsilon\text{-HQ}}$	$\tau\in(0,1),\varepsilon\neq0,\Gamma\neq0$	**3512.139**	3558.182	3486.916	**1.822**	1.883	1.851	624.794	**644.337**	631.028	73.937	75.523	73.977

- $(\tau \in (0,1), \varepsilon = 0, \Gamma = 0)$: plain hybrid quantile loss L_{HQ}, providing only the hybrid structure without any dynamic adjustment.
- $(\tau \in (0,1), \varepsilon = 0, \Gamma \neq 0)$: L_{HQ} with adaptive residual mapping, isolating the effect of residual rescaling.
- $(\tau \in (0,1), \varepsilon \neq 0, \Gamma = 0)$: $L_{\varepsilon\text{-HQ}}$ without adaptive residual mapping, isolating the effect of insensitive-zone tuning.

Configurations with $\tau = 1$ and $\Gamma \neq 0$ reduce to the baseline and are excluded. $L_{\varepsilon\text{-Q}}$ and $L_{\varepsilon\text{-BALS}}$ are orthogonal special cases of $L_{\varepsilon\text{-HQ}}$, while L_{Q} and L_{BALS} are degenerate hybrid forms.

Across the 12 dataset–predictor combinations, the full $L_{\varepsilon\text{-HQ}}$ achieves the highest total of top-two placements (10), comprising 3 best and 7 second-best results. The mapping-only variant (L_{HQ} with $\varepsilon = 0, \Gamma \neq 0$) attains 8 top-two placements (5 best, 3 second-best), whereas the tuning-only variant ($L_{\varepsilon\text{-HQ}}$ with $\varepsilon \neq 0, \Gamma = 0$) obtains only 2 second-best placements, and the plain L_{HQ} without dynamic adjustment yields none. Notably, the mapping-only variant attains more best-case results (5 vs. 3), consistent with Corollary 5.4: eliminating the insensitive-zone bias reduces coverage error, thereby favoring best-case outcomes. Nevertheless, the full $L_{\varepsilon\text{-HQ}}$ secures more top-two placements overall (10 vs. 8), indicating that the insensitive zone, despite introducing a small bias, provides a stability benefit that prevents worst-case degradation. These results demonstrate that both dynamic mechanisms contribute independently, and their combination is essential for the most stable performance.

The standard quantile loss L_{Q} obtains only 1 second-best placement, while its insensitive variant $L_{\varepsilon\text{-Q}}$ obtains 3 best results, confirming that an insensitive zone alone offers limited stabilization. The fully smooth L_{BALS} and $L_{\varepsilon\text{-BALS}}$ perform poorly across all configurations. Two limitations explain this: (1) the full smoothness delays threshold adaptation in the early stage, producing overly wide intervals; (2) the fully bounded nature causes gradient saturation for moderate-to-large residuals, suppressing informative gradients precisely when large coverage corrections are needed. The ε-HQ loss avoids both issues through its partially bounded hybrid design: the linear component prevents gradient saturation, while the quadratic component controls outliers.

Overall, the ablation results demonstrate that continuous gradient feedback from the hybrid structure is necessary but not sufficient for superior coverage performance. The decisive factors are the two dynamic adjustment mechanisms: adaptive residual mapping ensures stable and informative gradients, and insensitive-zone tuning adapts the loss to the prevailing noise level. Their close coupling jointly produces the marked performance gains observed throughout the experiments.

E.2 Evaluation of Insensitive-Zone Tuning Mechanisms

The insensitive-zone tuning mechanism is a key component of ε-HQ-DCOCI, as it adaptively adjusts the algorithm's perception of the current noise level. It combines a perturbation magnitude from the Tukey lower fence with the gradient-direction consistency norm. To validate the choice of the scale estimator, we compare three strategies: MAD-based, Sigma-based, and Tukey-based tuning. Results on the four financial time series datasets are reported in Table 19.

As presented in Table 19, the Tukey-based variant achieves 7 best and 3 second-best MIS results, while the MAD-based variant obtains 2 best and 5 second-best, and the Sigma-based variant obtains

Table 19: Performance comparison of ε-HQ-DCOGD under three scale estimators on four financial time series datasets with three base predictors. For each dataset-predictor pair, the **best** and second-best MIS results are bolded and underlined in red.

Estimator	Model	Bitcoin			DX			IXIC			XAU		
		Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓	Cov(%)	PI-Width ↓	MIS ↓
MAD-based	LightGBM	88.913	2355.191	3531.540	87.690	1.321	1.829	89.917	441.491	624.819	89.087	51.102	**73.815**
	Theta	89.055	2404.932	3575.009	89.111	1.401	1.892	89.360	449.159	645.002	90.094	53.746	75.473
	Prophet	88.794	2332.951	3490.211	87.183	1.296	1.862	89.569	451.442	632.269	89.322	51.700	**73.680**
Sigma-based	LightGBM	88.534	2311.817	3535.814	87.521	1.299	1.829	89.673	439.303	**624.021**	88.852	50.490	73.978
	Theta	88.676	2364.221	3595.092	89.043	1.400	1.889	89.256	447.350	644.932	90.094	53.547	**75.448**
	Prophet	88.368	2288.956	3497.591	87.758	1.319	1.859	89.604	448.262	**630.623**	89.255	51.829	73.812
Turkey-based	LightGBM	89.173	2377.241	**3512.139**	88.739	1.341	**1.822**	90.230	447.119	624.794	88.919	50.076	73.937
	Theta	89.600	2439.278	**3558.182**	88.367	1.357	**1.883**	89.673	460.297	**644.337**	89.154	51.526	75.523
	Prophet	89.315	2388.768	**3486.916**	88.705	1.352	**1.851**	89.360	442.671	631.028	88.818	50.119	73.977

Table 20: Performance comparison and average ranks of ε-HQ-DCOGD under different τ values.

Metric	τ	Bitcoin			DX			IXIC			XAU			Average rank
		LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	
ΔCov	0.1	0.614	0.993	0.779	1.498	1.126	1.194	0.918	0.257	1.648	0.880	1.182	0.981	2.250
	0.2	0.590	1.514	0.756	2.242	0.686	1.261	0.257	0.230	1.370	0.880	0.846	1.316	2.167
	0.3	1.087	0.590	0.519	1.566	0.822	1.769	0.160	0.257	0.779	0.712	0.376	0.443	1.667
	0.4	3.765	4.049	4.333	5.590	6.706	5.928	3.665	3.873	3.595	5.379	4.574	4.842	5.500
	0.5	12.104	12.175	12.578	15.736	15.127	15.702	12.983	12.705	12.670	14.480	14.446	13.573	8.417
	0.6	14.899	15.042	14.189	13.977	14.552	14.281	13.296	14.548	14.270	14.211	14.614	13.405	8.583
	0.7	8.811	8.290	8.147	8.769	9.615	9.040	9.019	8.533	8.950	8.838	8.805	9.174	7.000
	0.8	3.409	3.717	3.575	5.894	5.489	5.421	4.638	4.256	4.117	4.909	5.111	5.178	5.500
	0.9	1.467	1.182	1.395	3.865	4.339	3.257	1.891	1.822	1.891	2.995	3.130	2.525	3.917
MIS	0.1	3516.447	3562.380	3486.362	1.828	1.882	1.849	624.460	643.994	629.847	73.789	75.584	73.686	1.500
	0.2	3519.313	3567.975	3488.424	1.825	1.884	1.849	623.879	645.735	630.675	73.881	75.458	74.027	2.083
	0.3	3525.899	3576.080	3493.737	1.828	1.889	1.849	624.682	644.881	631.476	73.735	75.352	73.664	2.500
	0.4	3548.707	3619.085	3592.948	1.859	1.963	1.890	635.951	653.146	645.880	76.660	77.722	76.882	5.417
	0.5	3913.551	4001.559	3883.863	2.168	2.250	2.141	713.768	725.673	724.581	86.634	87.332	83.675	8.417
	0.6	3952.174	4031.094	4069.879	2.106	2.161	2.138	696.154	738.000	716.548	87.073	87.964	84.874	8.583
	0.7	3724.567	3802.230	3715.736	1.967	2.025	1.990	659.427	684.610	668.344	81.056	81.941	79.927	7.000
	0.8	3559.865	3601.267	3541.514	1.875	1.937	1.907	643.845	665.200	645.355	76.435	77.962	76.540	5.417
	0.9	3522.502	3593.120	3553.484	1.853	1.918	1.871	632.080	654.061	642.523	75.122	76.430	74.616	4.083

3 best and 4 second-best. In terms of top-two placements, the Tukey-based variant attains 10, compared with 7 for each of the other two. These results support the use of the Tukey lower fence as the scale estimator within the consistency-gated tuning framework.

E.3 Sensitivity Analysis for Key Hyperparameters

We evaluate the sensitivity of ε-HQ-DCOGD to four key hyperparameters: the mixing coefficient τ, the gradient decay factor Γ, the sliding window width W, and the Tukey multiplier k. For each parameter, we report the average rank across 12 tasks (4 financial datasets $\times$ 3 base predictors) under two metrics: the absolute coverage deviation $\Delta\text{Cov} = |\text{Cov} - (1-\alpha)|$ and the mean interval score (MIS). Lower ranks indicate better performance. Statistical significance is assessed using the Friedman test (Friedman, 1937), followed by the Nemenyi post-hoc test (Demšar, 2006) at the 5% significance level. The Friedman statistic is given by

$$\chi_F^2 = \frac{12N}{k(k+1)} \sum_{j=1}^{k} \left(\bar{R}_j - \frac{k+1}{2} \right)^2, \tag{89}$$

where k is the number of settings, N the number of tasks, and $\bar{R}_j$ the average rank of the j-th setting. The Nemenyi critical difference is

$$\text{CD} = q_{0.05} \sqrt{\frac{k(k+1)}{6N}}, \tag{90}$$

where $q_{0.05}$ is the critical value of the Studentized range statistic.

Mixing coefficient τ. The mixing coefficient τ balances the linear and bounded quadratic components of the ε-HQ loss, directly affecting coverage accuracy and interval efficiency. We evaluate $\tau \in \{0.1, 0.2, \ldots, 0.9\}$ on the 12 tasks, with average ranks reported in Table 20. The Friedman test yields $\chi_F^2 \approx 91.02$ for ΔCov and $\chi_F^2 \approx 90.74$ for MIS ($k = 9$, $N = 12$). With $k - 1 = 8$ degrees of freedom, $\chi^2_{0.05}(8) = 15.51$, both statistics are highly significant ($p < 10^{-10}$), rejecting the null hypothesis of equal performance.

Table 21: Performance comparison and average ranks of ε-HQ-DCOGD under different Γ values.

Metric	Γ	Bitcoin			DX			IXIC			XAU			Average rank
		LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	
ΔCov	0.1	1.206	0.614	0.140	2.242	2.344	1.464	0.160	0.848	1.648	1.048	0.544	1.081	3.083
	0.2	0.637	0.353	1.087	2.310	2.513	2.174	0.090	0.535	0.501	1.619	1.652	1.014	3.167
	0.3	1.585	1.585	1.632	2.851	1.938	3.054	1.613	2.065	1.231	2.055	2.122	1.014	6.583
	0.4	1.324	0.732	1.585	2.851	4.001	1.870	1.544	1.613	1.857	1.417	1.753	1.014	5.667
	0.5	1.111	0.377	0.495	2.749	1.938	1.836	3.908	1.092	0.535	1.585	1.753	1.753	4.583
	0.6	0.945	1.064	0.472	2.851	2.445	5.894	1.092	1.266	4.465	1.417	1.753	1.014	5.500
	0.7	0.329	0.898	0.495	2.783	1.904	5.962	0.848	4.186	1.579	1.417	1.753	1.014	4.833
	0.8	1.230	0.685	5.802	2.513	2.479	1.566	0.848	0.883	6.307	1.417	1.753	1.014	5.583
	0.9	1.159	0.756	1.988	2.851	1.938	1.938	1.509	0.814	1.648	1.417	1.753	1.014	6.000
MIS	0.1	3526.721	3567.426	3491.502	1.825	1.887	1.849	624.657	646.334	630.991	73.869	75.367	74.026	3.083
	0.2	3517.763	3567.760	3489.256	1.831	1.884	1.851	624.992	644.706	631.080	73.920	75.721	73.694	2.250
	0.3	3520.187	3564.635	3486.241	1.831	1.886	1.856	625.002	644.916	631.120	74.278	76.185	73.833	3.833
	0.4	3524.568	3580.760	3494.043	1.829	1.906	1.850	625.922	645.191	632.265	74.280	76.028	73.833	4.667
	0.5	3521.169	3575.077	3511.248	1.833	1.886	1.856	629.706	653.670	634.381	74.388	76.028	74.299	6.167
	0.6	3529.930	3570.945	3525.054	1.831	1.890	1.889	623.974	648.167	636.184	74.509	76.028	73.707	5.750
	0.7	3511.563	3551.213	3512.402	1.831	1.890	1.894	628.588	654.337	635.904	74.509	76.028	73.707	5.750
	0.8	3528.288	3574.372	3576.476	1.830	1.886	1.856	626.415	650.280	647.767	74.509	76.028	73.707	6.500
	0.9	3533.073	3582.619	3522.584	1.831	1.886	1.852	627.911	648.820	637.242	74.509	76.028	73.707	7.000

Table 22: Performance comparison and average ranks of ε-HQ-DCOGD under different W values.

Metric	W	Bitcoin			DX			IXIC			XAU			Average rank
		LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	
ΔCov	10	0.637	0.827	0.756	1.938	1.160	1.667	0.056	0.466	1.300	0.712	0.678	0.007	2.917
	50	0.661	0.898	0.732	2.310	1.025	0.788	0.334	0.327	0.640	0.712	0.094	1.216	2.917
	100	0.827	0.400	0.685	1.261	1.633	1.295	0.230	0.327	0.640	1.081	0.846	1.182	2.917
	150	0.708	0.472	0.708	1.464	0.720	1.667	0.118	0.674	1.266	0.846	0.510	0.343	2.833
	200	0.614	0.874	0.708	2.208	1.701	0.619	0.779	0.396	1.440	0.678	0.779	1.316	3.417
MIS	10	3518.854	3569.673	3485.272	1.827	1.883	1.848	624.820	644.432	630.849	73.809	75.446	73.631	3.417
	50	3516.304	3561.258	3488.140	1.828	1.884	1.849	624.356	644.006	631.008	73.783	75.324	74.022	3.167
	100	3512.139	3558.182	3486.916	1.822	1.883	1.851	624.794	644.337	631.028	73.937	75.523	73.977	3.333
	150	3515.374	3563.022	3484.181	1.826	1.884	1.848	624.386	644.174	630.927	73.718	75.338	73.609	2.417
	200	3513.362	3568.359	3488.494	1.825	1.885	1.850	624.294	643.824	630.880	73.779	75.389	73.606	2.667

The Nemenyi critical difference is CD ≈ 3.468. For ΔCov, the best setting is $\tau = 0.3$ (average rank 1.667). Its rank differences from $\tau = 0.2, 0.1, 0.9$ are 0.500, 0.583, and 2.250, all below the critical difference; in contrast, $\tau = 0.4$ and $\tau = 0.8$ differ by 3.833, $\tau = 0.7$ by 5.333, $\tau = 0.5$ by 6.750, and $\tau = 0.6$ by 6.916, all exceeding the critical difference and thus significantly worse. For MIS, the best setting is $\tau = 0.1$ (average rank 1.500). Its rank differences from $\tau = 0.2, 0.3, 0.9$ are 0.583, 1.000, and 2.583, all below the critical difference; whereas $\tau = 0.4$ and $\tau = 0.8$ differ by 3.917, $\tau = 0.7$ by 5.500, $\tau = 0.5$ by 6.917, and $\tau = 0.6$ by 7.083, all significantly worse. In summary, $\tau \in [0.1, 0.3]$ achieves the best performance with no significant difference among these values, while $\tau \in [0.4, 0.8]$ degrades significantly, with $\tau = 0.5$ and $\tau = 0.6$ performing worst. Therefore, we recommend $\tau = 0.1$ or $\tau = 0.3$ as the default.

Gradient decay factor Γ. The gradient decay factor Γ controls the rescaling range of the adaptive residual mapping, influencing gradient sensitivity and stability. We evaluate $\Gamma \in \{0.1, 0.2, \ldots, 0.9\}$ on the 12 tasks, with average ranks reported in Table 21. The Friedman test yields $\chi_F^2 \approx 18.85$ for ΔCov and $\chi_F^2 \approx 34.32$ for MIS ($k = 9$, $N = 12$). With $\chi^2_{0.05}(8) = 15.51$, both statistics are significant ($p < 0.05$ and $p < 0.001$, respectively). The Nemenyi critical difference is CD ≈ 3.468.

For ΔCov, the best setting is $\Gamma = 0.1$ (average rank 3.083). Its rank differences from all other settings are below the critical difference except for $\Gamma = 0.3$ (difference 3.500), which is significantly worse. For MIS, the best setting is $\Gamma = 0.2$ (average rank 2.250). Its rank differences from $\Gamma = 0.1, 0.3, 0.4$ are 0.833, 1.583, and 2.417, all below the critical difference; whereas $\Gamma = 0.5$ differs by 3.917, $\Gamma = 0.6$ by 3.500, $\Gamma = 0.7$ by 3.500, $\Gamma = 0.8$ by 4.250, and $\Gamma = 0.9$ by 4.750, all significantly worse. Hence, smaller Γ (0.1 or 0.2) yields the best performance, with $\Gamma = 0.1$ best for ΔCov and $\Gamma = 0.2$ best for MIS; these two are not significantly different. We recommend $\Gamma = 0.1$ or $\Gamma = 0.2$ as the default.

Sliding window width W. The sliding window width W determines the smoothing range of historical gradient directions in the momentum calibration. Although fixed at $W = 100$ in all experiments, we evaluate its sensitivity for completeness. We consider $W \in \{10, 50, 100, 150, 200\}$ on the 12 tasks, with average ranks reported in Table 22. The Friedman test yields $\chi_F^2 \approx 1.07$ for ΔCov and $\chi_F^2 \approx 3.67$ for MIS ($k = 5$, $N = 12$). With $k - 1 = 4$ degrees of freedom, $\chi^2_{0.05}(4) = 9.49$, neither statistic is significant ($p > 0.05$), indicating that performance is insensitive

Table 23: Performance comparison and average ranks of ε-HQ-DCOGD under different k values.

Metric	k	Bitcoin			DX			IXIC			XAU			Average rank
		LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	LightGBM	Theta	Prophet	
ΔCov	0.5	0.922	1.467	0.519	1.701	0.788	1.160	0.195	0.535	0.744	0.712	0.913	0.343	2.417
	1.0	1.324	0.424	1.087	1.870	1.701	1.194	0.257	0.570	0.709	0.746	0.712	1.619	3.750
	1.5	0.827	0.400	0.685	1.261	1.633	1.295	0.230	0.327	0.640	1.081	0.846	1.182	2.417
	2.0	0.661	1.869	0.614	1.261	1.633	1.295	0.230	0.466	1.300	0.880	0.175	1.182	2.750
	2.5	0.661	1.064	0.661	1.261	1.633	1.295	0.230	0.640	1.300	0.880	0.175	1.182	3.667
MIS	0.5	3513.108	3563.507	3494.224	1.826	1.886	1.848	624.474	644.804	631.115	73.659	75.498	73.542	3.500
	1.0	3506.818	3573.495	3488.204	1.822	1.884	1.848	624.138	644.582	631.670	73.720	75.447	74.099	2.917
	1.5	3512.139	3558.182	3486.916	1.822	1.883	1.851	624.794	644.337	631.028	73.937	75.523	73.977	2.667
	2.0	3511.880	3560.852	3484.527	1.825	1.884	1.851	624.794	644.753	630.791	73.826	75.330	73.977	2.500
	2.5	3511.880	3569.327	3485.610	1.825	1.884	1.851	624.888	644.683	630.791	73.826	75.330	73.977	3.417

to W. For $\Delta\mathrm{Cov}$, the average ranks range from 2.833 to 3.417; for MIS, they range from 2.417 to 3.417, with $W = 150$ best (2.417) and $W = 10$ worst (3.417), but the variation is negligible. We therefore fix $W = 100$ throughout.

Tukey multiplier k. The Tukey multiplier k determines the truncation scale of the interquartile-range lower fence in the insensitive-zone tuning, affecting noise filtering strength. Although fixed at $k = 1.5$ in all experiments, we evaluate its sensitivity. We consider $k \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$ on the 12 tasks, with average ranks reported in Table 23. The Friedman test yields $\chi_F^2 \approx 8.41$ for $\Delta\mathrm{Cov}$ and $\chi_F^2 \approx 3.80$ for MIS ($k = 5$, $N = 12$). With $\chi^2_{0.05}(4) = 9.49$, neither statistic is significant ($p > 0.05$), indicating that performance is insensitive to k. For $\Delta\mathrm{Cov}$, the average ranks range from 2.417 to 3.750, with $k = 0.5$ and $k = 1.5$ jointly best (2.417) and $k = 1.0$ worst (3.750); for MIS, they range from 2.500 to 3.500, with $k = 2.0$ best (2.500) and $k = 0.5$ worst (3.500). The variation is small, and we fix $k = 1.5$ as the default.

In summary, the ε-HQ-DCOGD algorithm exhibits mild sensitivity to τ and Γ, for which we recommend $\tau \in [0.1, 0.3]$ and $\Gamma \in [0.1, 0.2]$, while it is largely insensitive to W and k, for which the default values $W = 100$ and $k = 1.5$ are robust.