Instance-Adaptive Online Multicalibration

Zhiming Huang[1], Jamie Morgenstern[1,3], Aaron Roth[2,4], and Jie Claire Zhang[1]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington

[2]Department of Computer and Information Sciences, University of Pennsylvania

[3]Amazon

[4]Resolution

Abstract

We study online multicalibration beyond the worst-case. We give a single, efficient algorithm that dynamically interpolates between benign and worst-case sequences by adaptively refining a dyadic grid of prediction values. Its error is controlled by the growth of the refinement tree. Our analysis recovers the known $\widetilde{O}(T^{2/3})$ worst-case-optimal rate for online multicalibration, while automatically adapting to easier instances: in the marginal stochastic setting it obtains a rate of $\widetilde{O}(\sqrt{T})$, and for piecewise-stationary means with J segments its rate is $\widetilde{O}(\sqrt{JT})$. More generally, the rate depends on how well the predictable means can be approximated by simple contextual scores whose thresholds can be represented by the group family. We also show that the threshold-complexity dependence is tight up to logarithmic factors.

1 Introduction

Calibration is a basic statistical property that we want for probabilistic predictions to be "trustworthy". At a surface level, calibration asks that predictions "mean what they say" in the sense that they should be unbiased predictors of the outcome, even conditional on the value of the prediction itself. For example, a calibrated weather forecaster should have the property that, amongst all days the forecaster predicts a 20% chance of rain, 20% of those days actually experience rain [Dawid, 1982]. Calibrated predictions are viewed as "trustworthy" because of how they mediate the interface between prediction and decision making: selecting an action that is a best response to calibrated forecasts is an optimal decision policy among all policies that map forecasts to actions, simultaneously for all sets of actions and utility functions [Foster and Vohra, 1997, Kleinberg et al., 2023, Noarov et al., 2025, Roth, 2022, Kiyani et al., 2026].

Remarkably, as first shown by Foster and Vohra [1997], it is possible to produce forecasts that are asymptotically calibrated even in sequential *adversarial* prediction settings, in which there is no underlying distribution, and outcomes can be chosen adaptively by an adversary. Throughout, a rate describes the growth of cumulative, unnormalized calibration error over T rounds. The achievable rate depends on the environment. In the standard batch/statistical learning setting, calibration can be obtained via marginal mean estimation: given T samples, simple algorithms can produce predictors whose expected calibration error, without the conventional $1/T$ normalization, scales as $\Theta(\sqrt{T})$ [Shabat et al., 2020, Gupta and Ramdas, 2021]. Calibration in the online adversarial setting, on the other hand, is a harder problem with the best known lower bounds of $\Omega(T^{0.543})$ separating it from the stochastic setting [Qiao and Valiant, 2021, Dagan et al., 2025].

Regardless of the rate at which one achieves it, calibration alone guarantees surprisingly little: it is a marginal property, demanding unbiased predictions only on average across the full sequence. We instead consider guarantees which ensure predictions are calibrated simultaneously on each of many subsequences that can be defined by the data [Dawid, 1985, Lehrer, 2003, Sandroni et al., 2003]. The modern formulation of this idea—multicalibration [Hébert-Johnson et al., 2018]—is achievable

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

at a rate of $\tilde{O}(T^{2/3})$ in online adversarial settings [Noarov et al., 2025, Ghuge et al., 2025], and this rate is tight in both the sequential adversarial and batch settings [Collina et al., 2026a,b]. The online lower bounds, however—$\Omega(T^{0.543})$ for marginal calibration [Dagan et al., 2025] and $\tilde{\Omega}(T^{2/3})$ for multicalibration [Collina et al., 2026b]—hold only in the worst case, leaving open whether typical instances admit substantially better rates. We ask:

> *Are there sequential multicalibration algorithms that go beyond the worst case, achieving instance-dependent bounds that interpolate between $O(\sqrt{T})$ rates on easy instances and the $O(T^{2/3})$ rates that are optimal in the worst case?*

We show that the answer is *yes*. A single, efficient algorithm attains instance-adaptive multicalibration rates. It recovers the minimax-optimal $\widetilde{O}(T^{2/3})$ worst-case rate for polynomial-sized group families, and more generally gives a continuum of bounds depending on how well the predictable means can be approximated by simple structured scores. For example, if the full predictable means μ_t are constant on each of J contiguous segments, with each segment's mean common to all realized contexts, the algorithm achieves multicalibration error $\widetilde{O}(\sqrt{JT})$ (Corollary 2). The $J = 1$ case gives $\widetilde{O}(\sqrt{T})$ and includes the marginal stochastic setting. More generally, if $C_{\mathrm{stat}}(\mu) = \inf_{c\in[0,1]} \sum_{t=1}^{T} |\mu_t - c|$, the same algorithm obtains $\widetilde{O}(\sqrt{T} + (TC_{\mathrm{stat}}(\mu))^{1/3})$ (Corollary 3). Both guarantees hold for any fixed finite group family, with logarithmic dependence on its size.

These guarantees are corollaries of our general multicalibration theorem, with marginal calibration as a special case. The relevant difficulty of an instance is governed by how well one can predict label means from the context in a way that the group family can represent. Concretely, we compare the predictable means to simple contextual scores. A score is useful when it takes only a few values and its thresholds can be represented cheaply using the group functions. The rate also pays for the residual error $\sum_{t\in S} |\mu_t - f_S(x_t)|$ left after approximating the predictable means by such scores on contiguous blocks. Thus label means may change at every round and still define an easy instance, provided those changes are predictable from context and their thresholds are represented by the group family.

Like many online calibration algorithms, our algorithm maintains a discrete grid of prediction values and computes the minimax optimal strategy in a game defined by a weighting of the worst-case next-round bias. Unlike many standard algorithms, ours does not maintain a fixed set of prediction values. Instead, it adaptively refines the grid over time, refining each interval once the algorithm has played it many times. A run of the algorithm produces an adaptively "grown" tree of prediction values. We first bound calibration error by accounting for the intervals that become active at each depth of this tree. The rest of the analysis upper bounds the number of active intervals at each depth in terms of the structure of the predictable means: intervals far from a good contextual approximation create detectable bias and therefore cannot be played often.

1.1 Related Work

Multicalibration was formalized by Hébert-Johnson et al. [2018], following earlier online algorithms [Lehrer, 2001, Sandroni et al., 2003, Foster and Kakade, 2006, Foster et al., 2011]. Gupta et al. [2022] gave the first efficient online multicalibration algorithm with logarithmic dependence on the number of groups and extensions to quantiles and moments. Following game-theoretic approaches to calibration [Hart, 2025, Fudenberg and Levine, 1999, Foster, 1999, Foster and Hart, 2018], Lee et al. [2022] study minimax multiobjective optimization, and Haghtalab et al. [2023] use online learning to solve zero-sum games. Farina and Perdomo [2026] formulate the minimax step as an expected variational inequality and apply fast solvers. Oracle-efficient methods handle rich benchmark classes [Garg et al., 2024, Ghuge et al., 2025, Hu et al., 2025, Farina and Perdomo, 2026], whereas our algorithm is efficient only for explicitly represented finite group families and gives instance-adaptive rates. Hu et al. [2025] also study swap multicalibration for elicitable properties beyond means (cf. Noarov and Roth [2023]). Recent work settles minimax rates in batch and online settings [Collina et al., 2026a,b], showing that the online upper bound of Noarov et al. [2025] is tight.

"Best of both worlds" algorithms adapt to favorable instances while retaining adversarial guarantees, in both bandits [Bubeck and Slivkins, 2012] and prediction with expert advice [De Rooij et al., 2014].

In online convex optimization, bounds adapting to the realized loss functions can yield $o(\sqrt{T})$ regret on favorable instances [Chiang et al., 2012, Chen et al., 2021].

Measuring instance difficulty through switches or drift also connects our work to "tracking the best expert" [Herbster and Warmuth, 1998, Bousquet and Warmuth, 2002] and adaptive regret [Hazan and Seshadhri, 2007].

Our adaptive grid refinement resembles "zooming" for Lipschitz bandits [Kleinberg et al., 2008, Slivkins, 2011], especially work giving instance-dependent bounds [Podimata and Slivkins, 2021, Krishnamurthy et al., 2020].

Hu et al. [2026] simultaneously use multiple discretization scales to bound swap regret for Lipschitz convex losses on continuous action spaces. Both approaches address limitations of fixed discretizations, but theirs gives worst-case regret bounds, while our adaptive tree yields calibration bounds depending on the realized mean trajectory.

Concurrently, Liu et al. [2026] study instance-adaptive marginal calibration. The initial bounds were incomparable: ours was $\min(T^{2/3}, \sqrt{JT})$ for means constant on J segments (the marginal specialization of Corollary 2), and theirs was $\sqrt{T} + (TC_{\mathrm{stat}}(q))^{1/3}$, suppressing logarithmic factors. Here $C_{\mathrm{stat}}(q)$ is the cumulative absolute deviation of the label mean from its median across time. After comparing the papers, we found that both methods yield both bounds. The marginal specialization of Corollary 3 recovers the bound first given by Liu et al. [2026]. Our marginal guarantees follow from adaptive multicalibration for ECE, while their algorithm covers marginal calibration under a broader collection of metrics.

2 Online Multicalibration and Dynamic Bins

2.1 Protocol and Error Notion

We consider the online multicalibration problem where in each round $t \in [T]$, the learner observes a context $x_t \in \mathcal{X}$, and chooses a mixed forecast π_t (a distribution over finitely many candidate prediction values in $[0, 1]$), and then privately samples a realized prediction p_t from π_t, and then observes an outcome $y_t \in [0, 1]$ revealed by nature.

We assume that y_t may depend on the past, on the current context x_t, and on the mixed forecast π_t, but not on the fresh randomization used to sample p_t. We describe this interaction protocol with filtration notation in Appendix A. Use $\mu_t \in [0, 1]$ to denote the conditional mean of realized outcome y_t, conditioned on all the transcript before p_t is drawn. Let $\mathcal{G}$ be a finite nonempty family of binary groups $g : \mathcal{X} \to \{0, 1\}$. If the all-ones group does not belong to the class we add it and write $\bar{\mathcal{G}} := \mathcal{G} \cup \{\mathbf{1}\}$.

Private randomization matters for the exact calibration error below. If Nature saw p_t first, choosing $y_t = \mathbf{1}\{p_t < 1/2\}$ would give residuals of one sign and magnitude at least $1/2$ within each prediction value, forcing marginal error at least $T/2$. Smooth calibration can admit deterministic forecasts [Foster and Hart, 2018]. We do not analyze smooth multicalibration here.

Definition 1 (Calibration error)**.** *For a group collection $\mathcal{G}$ and realized transcript, the multicalibration error is*

$$\mathrm{MCerr}_{\mathcal{G}}(T) := \max_{g \in \mathcal{G}} \sum_{v \in \{p_1, \ldots, p_T\}} \left| \sum_{t : p_t = v} g(x_t)(y_t - v) \right|.$$

The marginal calibration error is:

$$\mathrm{CalErr}(T) := \sum_{v \in \{p_1, \ldots, p_T\}} \left| \sum_{t : p_t = v} (y_t - v) \right|.$$

2.2 Dynamic bins

Now we introduce the core structure of our algorithm: a dynamic, multiscale partition of the prediction space $[0, 1]$. Let $d_{\max} := \left\lfloor \frac{1}{2} \log_2 T \right\rfloor + 1$. For each depth $d = 0, \ldots, d_{\max}$, define the dyadic grid

$$\mathcal{D}_d := \left\{ [k2^{-d}, (k+1)2^{-d}) : k = 0, \ldots, 2^d - 2 \right\} \cup \left\{ [1 - 2^{-d}, 1] \right\}.$$

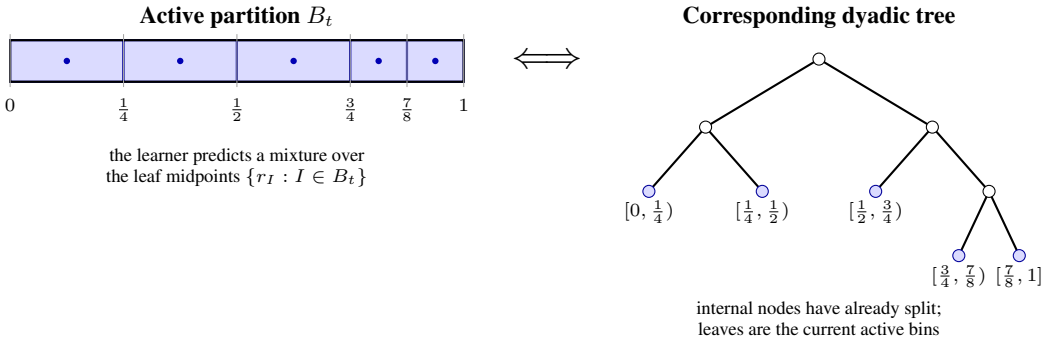

Figure 1: A typical state of the dynamic-bin data structure. The learner predicts using the midpoints of the current leaves, and an interval is refined only after the total mass assigned to it reaches the threshold L/w_I^2.

Use $\mathcal{D}$ to denote the union of intervals[1] at all depths, $\mathcal{D} := \bigcup_{d=0}^{d_{\max}} \mathcal{D}_d$. For each interval $I \in \mathcal{D}$, let r_I denote its midpoint and let $w_I = \sup_{x,y \in I} |x - y|$ denote its width. For intervals at depth d, we sometimes denote their width as w_d. The learner maintains an active dyadic partition $B_t \subseteq \mathcal{D}$ of $[0, 1]$, starting from $B_1 = \{[0, 1]\}$. We call an interval I "active" for rounds t where $I \in B_t$. We refer to $\mathcal{V}_T := \bigcup_{t=1}^{T} B_t$ as the set of intervals that are ever active. In round t, we denote the mixed forecast as $\pi_t \in \Delta(B_t)$, where $\pi_{t,I}$ denotes the probability assigned to the midpoint r_I of interval I.

For a set of rounds $S \subseteq [T]$, define the (expected) total play of interval I on S by $N_S(I) := \sum_{t \in S} \pi_{t,I}$, with the convention $\pi_{t,I} = 0$ when $I \notin B_t$ and $[0] := \varnothing$. We denote the full-horizon total play as $N(I) := N_{[T]}(I)$. Let $L := \log(eT|\bar{\mathcal{G}}|)$. Each interval I of depth strictly smaller than $d_{\max}$ is split at the end of the first round t where its running total play satisfies $N_{[t]}(I) \geq \frac{L}{w_I^2}$. When I splits, it is permanently removed from active partition B_t, meaning its final total play $N(I)$ remains fixed at $N_{[t]}(I)$. Its two dyadic children are added to the active partition in the next rounds. Each child interval is of width $w_I/2$ and initialized with total play value 0. Our algorithm has contiguous active lifetimes for every interval, which we denote as $A(I) := \{t \in [T] : I \in B_t\}$, and $\pi_{t,I} = 0$ for all $t \notin A(I)$. We write Λ_T for the number of leaves after the end-of-round T splits in the dyadic tree.

Why this split threshold? The threshold $N(I) \geq L/w_I^2$ is the scale at which further refinement becomes worthwhile. An interval of width w_I contributes deterministic discretization error on the order of $N(I) \cdot w_I$, while the corresponding online-learning term scales like $\sqrt{N(I)L}$. The crossover point $N(I) \asymp \frac{L}{w_I^2}$ is where these two terms are equal. The algorithm therefore splits an interval when the discretization error has become the dominant source of bias and not before. For comparison, a fixed grid of K values gives the global tradeoff $\widetilde{O}(\sqrt{KT} + T/K)$. Its optimized bound is $\widetilde{O}(T^{2/3})$. Dynamic refinement makes this tradeoff separately in each interval, allocating finer values where enough probability has accumulated to justify them.

2.3 The general algorithm

The algorithm has four modules: maintain the active prediction bins, weight their calibration constraints, choose a mixed forecast, and update the weights and refine the bins after feedback. The tree partitions prediction values. Context enters through the group values $g(x_t)$, which influence the forecast distribution and hence the mass assigned to each bin.

For every group $g \in \bar{\mathcal{G}}$, interval I, and sign $\sigma \in \{+, -\}$, define the group-weighted signed bias terms. The two signs monitor predictions that are too low and too high, respectively:

$$\phi_{g,I}^{+}(\pi_t, y_t) := g(x_t)\pi_{t,I}(y_t - r_I - w_I), \qquad \phi_{g,I}^{-}(\pi_t, y_t) := g(x_t)\pi_{t,I}(r_I - y_t - w_I).$$

[1]We use intervals and bins interchangeably throughout the paper.

The extra slack term w_I is a discretization buffer: deviations of size at most w_I are below the resolution of interval I, so they should not count as evidence that r_I is too large or too small.

We combine these weighted bias terms using a sleeping-experts algorithm, instantiated with confidence-rated AdaNormalHedge [Luo and Schapire, 2015]. Each expert $(s, g, (I, \sigma))$ monitors one such constraint starting at time s. It is awake in round t if and only if $s \le t$ and $I \in B_t$. We include start times s so that we can control weighted bias on every contiguous sub-interval of rounds, not just marginally.

The sleeping-experts interface. At round t, the wrapper maintains a nonnegative weight $\omega_{t,e}$ for each expert e that is awake in that round. These weights are normalized so that $\sum_{e:\text{ awake at } t} \omega_{t,e} = 1$. For each group g and interval $I \in B_t$, define the aggregate sign weights assigned to positive and negative copies of (g, I):

$$\lambda^+_{t,g,I} := \sum_{s \le t} \omega_{t,(s,g,(I,+))}, \qquad \lambda^-_{t,g,I} := \sum_{s \le t} \omega_{t,(s,g,(I,-))}.$$

We then track interval-level aggregations of these weights

$$a_{t,I} := \sum_{g \in \bar{\mathcal{G}}} g(x_t)\big(\lambda^+_{t,g,I} - \lambda^-_{t,g,I}\big), \qquad b_{t,I} := \sum_{g \in \bar{\mathcal{G}}} g(x_t)\big(\lambda^+_{t,g,I} + \lambda^-_{t,g,I}\big).$$

In words, $a_{t,I}$ records the net signed weight favoring outcomes above rather than below r_I, while $b_{t,I}$ records the total weight assigned to both of the corresponding one-sided bias terms for I. By construction, $|a_{t,I}| \le b_{t,I}$.

The wrapper's aggregate one-step bias is the weighted average of these expert-level biases:

$$\widehat{\phi}_t := \sum_{I \in B_t} \sum_{g \in \bar{\mathcal{G}}} \Big(\lambda^+_{t,g,I}\phi^+_{g,I}(\pi_t, y_t) \;+\; \lambda^-_{t,g,I}\phi^-_{g,I}(\pi_t, y_t)\Big),$$

and regrouping by interval gives the identity $\widehat{\phi}_t = \sum_{I \in B_t} \pi_{t,I}\big(a_{t,I}(y_t - r_I) - b_{t,I}w_I\big)$.

A forecast π_t is then chosen so that no convex combination of the currently active one-sided weighted bias terms has positive aggregate one-step bias, regardless of which outcome $y \in [0, 1]$ Nature reveals. In particular, the learner chooses $\pi_t \in \Delta(B_t)$ satisfying

$$\sum_{I \in B_t} \pi_{t,I}\big(a_{t,I}(y - r_I) - b_{t,I}w_I\big) \le 0 \qquad \text{for every } y \in [0, 1]. \tag{1}$$

The feasibility of this set of constraints, and the existence of such a mixed forecast, follow from a weighted feasibility argument based on Sion's minimax theorem (see Lemma 4 in Appendix E.1). Because the expression $\sum_{I \in B_t} \pi_{t,I}\left(a_{t,I}\left(y - r_I\right) - b_{t,I}w_I\right)$ is affine in y, Equation (1) holds for all $y \in [0, 1]$ if and only if it holds at the endpoints $y = 0$ and $y = 1$. Consequently, finding the mixed forecast π_t computationally reduces to solving a standard linear program over the simplex with exactly two constraints.

The learner then samples $p_t \sim \pi_t$, observes y_t, and supplies the normalized loss $(3/2 - \phi^\sigma_{g,I}(\pi_t, y_t))/3$ to each awake expert. AdaNormalHedge updates its cumulative regret and absolute-regret statistics, which determine the next weights. The learner then splits intervals satisfying the threshold L/w_I^2. Algorithm 1 summarizes the procedure. Appendix B gives marginal-calibration intuition, and Appendix E specifies the expert prior and updates used in the analysis.

Explicit implementation cost. At round t, there are $A_t = 2t|\bar{\mathcal{G}}||B_t|$ awake experts. Each stores a constant number of scalar statistics. Summing their weights and updating their states takes $O(A_t)$ arithmetic operations, in addition to evaluating the groups and solving the linear program over $|B_t|$ probabilities. Experts can be created when they first become awake and discarded when their interval splits, giving $O(A_t)$ stored expert state. The split-threshold resource bound gives $|B_t| = \widetilde{O}(T^{1/3})$ (Appendix F.3), so the worst-case per-round update time and memory are $\widetilde{O}(|\bar{\mathcal{G}}|T^{4/3})$, and total update time is $\widetilde{O}(|\bar{\mathcal{G}}|T^{7/3})$, plus the linear-program costs. Smaller trees reduce these costs.

Algorithm 1: Dynamic bins for online multicalibration

Input: horizon T, group family $\mathcal{G}$
Initialize $B_1 = \{[0,1]\}$ and $N_{[0]}([0,1]) = 0$;
Instantiate a sleeping-experts wrapper using confidence-rated AdaNormalHedge on the tuples $(s, g, (I, \sigma))$, $\sigma \in \{+, -\}$;
for $t = 1, \ldots, T$ **do**
  Observe the context x_t;
  Aggregate the weights of the awake experts into coefficients $a_{t,I}, b_{t,I}$;
  Find $\pi_t \in \Delta(B_t)$ satisfying (1);
  Sample a realized prediction $p_t = r_I$ with probability $\pi_{t,I}$;
  Observe the outcome $y_t \in [0,1]$;
  Update the losses of the awake experts using the group-weighted bias terms $\phi^{\pm}_{g,I}(\pi_t, y_t)$;
  Update $N_{[t]}(I) \leftarrow N_{[t-1]}(I) + \pi_{t,I}$ for every $I \in B_t$;
  Split every $I \in B_t$ of depth $d < d_{\max}$ with $N_{[t]}(I) \geq L/w_I^2$ into its two children, each child I' initialized with $N_{[t]}(I') = 0$;
end

3 A Hierarchy of Complexity Measures

We now define the complexity measures appearing in our bounds. Write $q_t := \mu_t$ for the predictable mean process. We begin with piecewise constancy, then introduce threshold representations and their residual approximation costs.

Definition 2 (Piecewise-constant predictable means)**.** *We say the scalar predictable mean sequence changes $J-1$ times if there exist $1 = \tau_0 < \tau_1 < \cdots < \tau_J = T+1$ and values $q^{(1)}, \ldots, q^{(J)} \in [0,1]$ such that $q_t = q^{(j)}$ for all $t \in \{\tau_{j-1}, \ldots, \tau_j - 1\}$.*

This benchmark is simple but it is overly conservative. In what follows we generalize it in two ways. First, we should consider a sequence to be simple even if it changes at every round, so long as those changes are only small deviations from some (piecewise) common center. Second, in the multicalibration setting in which there are contexts, there should be "simple" sequences that change dramatically at every round so long as those changes are predictable from the contexts in a way that is "visible" to the groups with respect to which the algorithm is multicalibrating.

We first introduce the appropriate generalization for multicalibration, which lets the mean change dramatically at every round so long as the changes are predictable from the group functions. Multicalibration algorithms control bias within every group. Thus what matters is whether the group family can represent, for each bin midpoint r_I, which side of r_I the relevant conditional means lie on. A sequence where μ_t changes every round can still be easy if those changes are predictable from context and their thresholds can be represented using $\bar{\mathcal{G}}$.

Just as Definition 2 required only that the mean be piecewise constant, and allowed it to change with time, we will similarly allow thresholds of label means to be *differently* predictable on different contiguous blocks of time S. At scale $w_d = 2^{-d}$, the useful object on a block S is a score f that approximately predicts μ_t, takes only K distinct values on the realized contexts, and has thresholds $\{f(x_t) \geq r\}$ that admit $\bar{\mathcal{G}}$-representations with ℓ_1 cost at most B. The product BK controls the number of depth-d splits that can be forced within block S: Lemma 8 shows that this number is $O(BK)$ when the approximation error is below scale w_d. A score is threshold representable if each of its level sets has a bounded-cost linear representation over the group family:

Definition 3 (Threshold-representable score)**.** *Fix $S \subseteq [T]$. A score $f : \mathcal{X} \to [0,1]$ is (K, B, η)-threshold representable on S if $|\{f(x_t) : t \in S\}| \leq K$ and for every $r \in [0,1]$ there exist coefficients $(\alpha_{r,g})_{g \in \bar{\mathcal{G}}}$ with $\sum_{g \in \bar{\mathcal{G}}} |\alpha_{r,g}| \leq B$ such that the function $h_r(x) := \sum_{g \in \bar{\mathcal{G}}} \alpha_{r,g} g(x)$ satisfies, for all $t \in S$,*

$$h_r(x_t) \geq 1 - \eta \qquad \text{whenever } f(x_t) \geq r,$$
$$h_r(x_t) \leq -(1-\eta) \quad \text{whenever } f(x_t) < r.$$

We fix the margin parameter $\eta = 1/4$ throughout to ensure the representation h_r separates by a margin strictly bounded away from zero. Any constant $\eta \in (0,1)$ would suffice for the subsequent

analysis as the specific choice only affects the absolute constants hidden by the Big-O notation. This condition is reminiscent of margin-based complexity notions such as fat-shattering, but it is instance-specific: rather than requiring a hypothesis class to realize arbitrary dichotomies on S, we only require low-cost representations of the nested threshold cuts induced by a chosen low-cardinality score f. As a sanity check, in the singleton-context case with $\bar{\mathcal{G}} = \{\mathbf{1}\}$, a constant score $f \equiv c$ is $(1, 1, 1/4)$-threshold representable. For an ordinal score $f(x) = a_{c(x)}$ with $a_1 < \cdots < a_K$, suppose the groups include $g_j(x) = \mathbf{1}\{c(x) \geq j\}$ for $j = 2, \ldots, K$. Each threshold sign is then 1, -1, or $2g_j - 1$, so its coefficient cost is at most 3. Thus K risk strata give a concrete $(K, 3, 1/4)$-threshold-representable score. Appendix C spells out the boundary cases.

To allow aggregate approximation error, we charge only deviations larger than the current prediction scale.

Definition 4 (Residual at scale d and threshold-simple block)**.** *For a contiguous block $S \subseteq [T]$, a local score $f : \mathcal{X} \to [0, 1]$, and depth d, the truncated residual is defined as*

$$\mathsf{R}_{S,d}(f) := \sum_{t \in S} \big(|\mu_t - f(x_t)| - w_d\big)_+, \qquad (a)_+ := \max\{a, 0\}.$$

Furthermore, we call S (K, B, d)-threshold simple if there exists a score f_S that is $(K, B, \frac{1}{4})$-threshold representable on S and satisfies $\mathsf{R}_{S,d}(f_S) = 0$. Zero residual is equivalent to $|\mu_t - f_S(x_t)| \leq w_d$ at every round in S.

Definition 5 (Depth-wise residual threshold cost)**.** *For each depth d, define*

$$\Psi_d^{\mathrm{res}}(\mu_{1:T}; \mathcal{G}) := \inf_{\mathcal{P}_d, \{(f_S, K_S, B_S)\}_{S \in \mathcal{P}_d}} \sum_{S \in \mathcal{P}_d} \left[B_S K_S + \sqrt{\frac{B_S K_S \, \mathsf{R}_{S,d}(f_S)\, w_d}{L}} \right],$$

where the infimum ranges over partitions $\mathcal{P}_d$ of $[T]$ into nonempty contiguous blocks and, for each block $S \in \mathcal{P}_d$, triples (f_S, K_S, B_S) such that f_S is $(K_S, B_S, \frac{1}{4})$-threshold representable on S with respect to $\bar{\mathcal{G}}$. If no admissible representation exists for a block, its cost is $+\infty$.

Zero scale-truncated residual defines the exact special case:

Definition 6 (Exact multiscale threshold complexity)**.** *For a fixed depth d, define*

$$M_d^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}) := \inf_{\mathcal{P}_d, \{(f_S, K_S, B_S)\}_{S \in \mathcal{P}_d}} \sum_{S \in \mathcal{P}_d} B_S K_S,$$

where the infimum ranges over partitions $\mathcal{P}_d$ of $[T]$ into nonempty contiguous blocks and triples (f_S, K_S, B_S) such that f_S is $(K_S, B_S, \frac{1}{4})$-threshold representable on S and satisfies

$$\mathsf{R}_{S,d}(f_S) = 0 \qquad \textit{for every } S \in \mathcal{P}_d.$$

The exact multiscale threshold complexity is

$$\Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}) := \sum_{d=0}^{d_{\max}} M_d^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}).$$

For one fixed partition and score collection, the following costs give an interpretable bound without a separate choice at each depth.

Definition 7 (Segmented threshold approximation cost)**.** *For a contiguous partition $\mathcal{P}$ of $[T]$ and triples $(f, K, B) = \{(f_S, K_S, B_S) : S \in \mathcal{P}\}$ such that each f_S is $(K_S, B_S, \frac{1}{4})$-threshold representable on S, define*

$$\mathsf{R}_S(f_S) := \sum_{t \in S} |\mu_t - f_S(x_t)|,$$

and

$$\mathsf{M}(\mathcal{P}, f) := 1 + \sum_{S \in \mathcal{P}} B_S K_S, \qquad \mathsf{A}(\mathcal{P}, f) := \sum_{S \in \mathcal{P}} \sqrt{B_S K_S \, \mathsf{R}_S(f_S)}.$$

We suppress the dependence of M and A on the chosen representation parameters K, B.

Setting	Complexity control	Calibration error
Worst-case adversarial	No structure assumed	$\widetilde{O}(T^{2/3})$
Exact threshold score (K levels, cost B)	$\mathsf{M}(\mathcal{P}, f) \lesssim 1 + BK$, $\mathsf{A}(\mathcal{P}, f) = 0$	$\widetilde{O}(\sqrt{BKT})$
Approximate threshold score (K levels, cost B)	$\mathsf{M}(\mathcal{P}, f) \lesssim 1 + BK$, $\mathsf{A}(\mathcal{P}, f) \lesssim \sqrt{BKR}$	$\widetilde{O}(\sqrt{BKT} + (TBKR)^{1/3})$
Segmentwise approximate threshold score	$\mathsf{M}(\mathcal{P}, f) = 1 + \sum_S B_S K_S$, $\mathsf{A}(\mathcal{P}, f) = \sum_S \sqrt{B_S K_S \mathsf{R}_S(f_S)}$	$\widetilde{O}(\sqrt{T\mathsf{M}(\mathcal{P}, f)} + T^{1/3}\mathsf{A}(\mathcal{P}, f)^{2/3})$
J-piecewise stationary means	$\mathsf{M}(\mathcal{P}, f) = 1 + J$, $\mathsf{A}(\mathcal{P}, f) = 0$	$\widetilde{O}(\sqrt{JT})$
Approximately stationary means	$\mathsf{M}(\mathcal{P}, f) = 2$, $\mathsf{A}(\mathcal{P}, f) = \sqrt{C_{\mathrm{stat}}(\mu)}$	$\widetilde{O}(\sqrt{T} + (TC_{\mathrm{stat}}(\mu))^{1/3})$

Table 1: Representative regimes covered by the dynamic-bin algorithm. The worst-case row recovers the known $\widetilde{O}(T^{2/3})$ online multicalibration rate achieved by prior algorithms [Noarov et al., 2025, Ghuge et al., 2025]. Stationarity refers to the full predictable means μ_t. Here $R := \sum_{t=1}^{T} |\mu_t - f(x_t)|$. In the last row, $C_{\mathrm{stat}}(\mu) := \inf_{c\in[0,1]} \sum_t |\mu_t - c|$. The algorithm does not need to know which regime it is running in. All of these bounds are obtained by the same algorithm.

4 Main Theorem and Corollaries

The main theorem controls multicalibration error through threshold-representable scores and their residual approximation error. Table 1 summarizes several representative consequences of the main theorem.

Theorem 1 (Instance-adaptive online multicalibration)**.** *For $d = 0, \dots, d_{\max}$, define*

$$\Theta_0 := 1, \qquad \Theta_d := \min\left\{2^d,\ \frac{Tw_d^2}{L},\ \Psi^{\mathrm{res}}_{d-1}(\mu_{1:T}; \mathcal{G})\right\} \quad (d \geq 1).$$

With probability at least $1 - O(1/T)$, the dynamic-bin algorithm satisfies

$$\mathrm{MCerr}_{\mathcal{G}}(T) \leq C \min\left\{T^{2/3}L^{1/3}, \sum_{d=0}^{d_{\max}} \min\left\{\sqrt{TL\Theta_d}, \frac{L\Theta_d}{w_d}\right\}\right\}.$$

Moreover, on the same high-probability event, for every contiguous partition $\mathcal{P}$ of $[T]$ and every choice of triples $(f, K, B) = \{(f_S, K_S, B_S) : S \in \mathcal{P}\}$ such that each f_S is $(K_S, B_S, \frac{1}{4})$-threshold representable on S,

$$\mathrm{MCerr}_{\mathcal{G}}(T) \leq C \min\left\{T^{2/3}L^{1/3}, \sqrt{TL\,\mathsf{M}(\mathcal{P}, f)} + T^{1/3}L^{1/3}\mathsf{A}(\mathcal{P}, f)^{2/3}\right\}.$$

The following three corollaries give multicalibration bounds for exact threshold structure, piecewise-constant predictable means, and predictable means close to a common constant. Their proofs appear in Appendix G.

Corollary 1 (Exact threshold-simple multicalibration)**.** *If the predictable mean process has exact multiscale threshold complexity $\Gamma^{\mathrm{thr}}_T(\mu_{1:T}; \mathcal{G})$ as in Definition 6, then, with probability at least $1 - O(1/T)$,*

$$\mathrm{MCerr}_{\mathcal{G}}(T) \leq \widetilde{O}\left(\min\left\{T^{2/3}, \sqrt{T\left(1 + \Gamma^{\mathrm{thr}}_T(\mu_{1:T}; \mathcal{G})\right)}\right\}\right).$$

In particular, if for every depth d the whole horizon $[T]$ is (K_d, B_d, d)-threshold simple with $B_d \leq B$ and $K_d \leq K$, then

$$\mathrm{MCerr}_{\mathcal{G}}(T) \leq \widetilde{O}\left(\min\left\{T^{2/3}, \sqrt{BKT}\right\}\right).$$

For any group family, means that are constant on J segments have $\Gamma_T^{\mathrm{thr}} \le J(d_{\max}+1)$, giving the next specialization.

Corollary 2 (Multicalibration with piecewise-constant predictable means)**.** *For any fixed finite nonempty family $\mathcal{G}$ of binary groups, suppose the predictable mean $q_t := \mu_t$ is constant on J contiguous segments. Then, with probability at least $1 - O(1/T)$, the dynamic-bin algorithm satisfies*

$$\mathrm{MCerr}_{\mathcal{G}}(T) \le \widetilde{O}\Big(\min\{T^{2/3}, \sqrt{JT}\}\Big).$$

The logarithmic factors depend on T and $|\bar{\mathcal{G}}|$.

With a singleton context space and $\mathcal{G} = \{\mathbf{1}\}$, this is the piecewise-stationary marginal guarantee.

For any fixed group family, predictable means that stay close to a common constant admit the following bound.

Corollary 3 (Multicalibration near a constant mean)**.** *For any fixed finite nonempty family $\mathcal{G}$ of binary groups, let $C_{\mathrm{stat}}(\mu) := \inf_{c\in[0,1]} \sum_{t=1}^{T} |\mu_t - c|$ be the best L_1 deviation of the predictable means from a common constant. Then, with probability at least $1 - O(1/T)$, the dynamic-bin algorithm satisfies*

$$\mathrm{MCerr}_{\mathcal{G}}(T) \le \widetilde{O}\Big(\sqrt{T} + (TC_{\mathrm{stat}}(\mu))^{1/3}\Big).$$

The logarithmic factors depend on T and $|\bar{\mathcal{G}}|$.

In the context-free marginal setting ($\mathcal{X}$ is a singleton and $\mathcal{G} = \{\mathbf{1}\}$), we have $\mu_t = q_t$ and $\mathrm{MCerr}_{\mathcal{G}}(T) = \mathrm{CalErr}(T)$. This marginal special case matches the guarantee first given by Liu et al. [2026].

Unknown horizon via doubling. Restarting with epoch horizons $1, 2, 4, \ldots$ preserves the $\widetilde{O}(T^{2/3})$ worst-case and the marginal $\widetilde{O}(\sqrt{T})$ stationary and $\widetilde{O}(\sqrt{JT})$ piecewise-stationary guarantees in expectation. For every fixed observation time, the general bound uses the sum of the observed epoch-wise threshold complexities, including the final partial epoch. Appendix H gives the bound and proof.

5 Proof of the Main Theorem

We give the proof chain here and full statements and derivations in Appendices E–G. Let $\mathcal{E}_{\mathrm{global}} := \mathcal{E}_{\mathrm{samp}} \cap \mathcal{E}_{\mathrm{conc}}$. The first event controls sampling error for every fixed group and dyadic bin, while the second controls mean noise on every contiguous time block (Appendix D.2). Their joint probability is at least $1 - 1/(2T)$.

Step 1: Regret controls local bias. Write $N_S^g(I) := \sum_{t\in S} g(x_t)\pi_{t,I}$. The two sign experts and weighted feasibility give, for every group g, bin I, and contiguous $S \subseteq A(I)$,

$$\left|\sum_{t\in S} g(x_t)\pi_{t,I}(y_t - r_I)\right| \le N_S^g(I) w_I + C_{\mathrm{loc}}\big(\sqrt{N_S^g(I)L} + L\big).$$

The first-order regret term scales with assigned probability mass, including when the bin's lifetime is long (Lemma 5).

Step 2: Account for error by depth. Write $\mathcal{V}_{T,d} = \mathcal{V}_T \cap \mathcal{D}_d$ and $m_d = |\mathcal{V}_{T,d}|$. For $L \le T$, the split rule gives $N(I)w_I^2 \le 2L$, including maximum-depth bins. Combining Step 1 with $\mathcal{E}_{\mathrm{samp}}$ and summing gives

$$\sum_{I\in\mathcal{V}_{T,d}} \left|\sum_{t:p_t=r_I} g(x_t)(y_t - r_I)\right| \le C \min\left\{\sqrt{TLm_d}, \frac{Lm_d}{w_d}\right\}$$

for every group and depth (Lemma 6).

Step 3: Residual mass pays for distant-bin play. For a threshold-representable score f on a contiguous block S, let $\delta_I = \operatorname{dist}(r_I, \{f(x_t) : t \in S\})$ and $\mathsf{D}_{S,I} = \sum_{t \in S \cap A(I)} \pi_{t,I}(|\mu_t - f(x_t)| - w_d)_+$. Expanding the threshold sign in groups and using $\mathcal{E}_{\mathrm{conc}}$ gives, when $\delta_I > 4Bw_d$,

$$N_S(I) \le C\left[\frac{B^2 L}{(\delta_I - 4Bw_d)^2} + \frac{B\mathsf{D}_{S,I}}{\delta_I - 4Bw_d}\right].$$

A distance-shell sum, using $\sum_I \mathsf{D}_{S,I} \le \mathsf{R}_{S,d}(f)$, therefore bounds the block's split charge by

$$\Xi_d(S) := \sum_{I \in \mathcal{D}_d} \min\{1, N_S(I) w_d^2/L\} \le C\left[BK + \sqrt{BK\mathsf{R}_{S,d}(f) w_d/L}\right]$$

(Lemmas 7–8).

Step 4: Count active bins. Every active nonroot bin has one split parent. Summing Step 3 over any temporal partition, then taking its infimum, gives $m_d \le C\Psi^{\mathrm{res}}_{d-1}$. The grid and total probability mass also give $m_d \le \min\{2^d, 8Tw_d^2/L\}$, while $m_0 = 1$. Thus $m_d \le C\Theta_d$ (Lemma 9).

Step 5: Sum the dyadic scales. For fixed partition/score costs $M = \mathsf{M}(\mathcal{P}, f)$ and $A = \mathsf{A}(\mathcal{P}, f)$, Step 4 gives $m_d \le C\min\{2^d, Tw_d^2/L, M + A\sqrt{w_d/L}\}$. Lemma 10 sums Step 2 to at most $C(\sqrt{TLM} + T^{1/3}L^{1/3}A^{2/3})$. The two crossovers are $w = \sqrt{LM/T}$ and $w = (A\sqrt{L}/T)^{2/3}$.

Step 6: Assemble the theorem. Dyadic midpoints are distinct, so exact empirical error is the sum of the depth contributions before maximizing over groups. Steps 2 and 4 prove the depth-wise bound, and Step 5 proves its segmented consequence. The deterministic caps give $\sum_d \min\{\sqrt{TL\Theta_d}, L\Theta_d/w_d\} \le CT^{2/3}L^{1/3}$ by splitting at $w = (L/T)^{1/3}$. If $L > T$, error at most T proves all branches directly, since the worst-case cap, depth-zero term and $M \ge 1$ term are each at least T. Appendix G.1 supplies the full assembly and uniformity argument.

6 Tightness of the $\sqrt{T\Gamma^{\mathrm{thr}}_T}$ Bound

The dependence on threshold complexity is tight throughout the intermediate range. For every power of two $2 \le m \le T^{1/3}$, the ordered-grid Walsh construction has $\Gamma^{\mathrm{thr}}_T = \widetilde{\Theta}(m)$ and forces expected error $\widetilde{\Omega}(\sqrt{mT})$ for every online forecaster. This matches our upper bound along the curve from $\sqrt{T}$ to $T^{2/3}$. Appendix I states the parameterized result and tracks the dependence on m in the lower-bound proof of Collina et al. [2026b].

Limitations and Broader Impacts

Our results are theoretical and come with several limitations. The explicit implementation maintains states for every group and start time, so large group families and long horizons can be costly. Threshold complexity may be difficult to estimate in practice, but the algorithm does not need to know or estimate it. This limits our ability to predict the improvement in advance. Although the bounds adapt to favorable instances, they still revert to the worst-case rate on sequences with high threshold complexity. Multicalibration can improve reliability across subpopulations, but calibration alone does not imply fairness, robustness, causal validity, or correct downstream decisions. These properties depend on the choice of groups and the decision problem, which we do not analyze.

Acknowledgments and Disclosures

We gratefully acknowledge support from the Simons Collaboration on Algorithmic Fairness and the NSF EnCoRE Tripods institute. Zhiming Huang acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Olivier Bousquet and Manfred K Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3(Nov):363–396, 2002.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proc. Conference on Learning Theory (COLT)*, pages 42.1–42.23. JMLR Workshop and Conference Proceedings, 2012.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Proc. Conference on Learning Theory (COLT)*, pages 1216–1259. PMLR, 2021.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Proc. Conference on Learning Theory (COLT)*, pages 6.1–6.20. JMLR Workshop and Conference Proceedings, 2012.

Natalie Collina, Jiuyao Lu, Georgy Noarov, and Aaron Roth. The sample complexity of multicalibration. *arXiv preprint arXiv:2604.21923*, 2026a.

Natalie Collina, Jiuyao Lu, Georgy Noarov, and Aaron Roth. Optimal lower bounds for online multicalibration. *arXiv preprint arXiv:2601.05245*, 2026b.

Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. Breaking the $T^{2/3}$ barrier for sequential calibration. In *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, pages 2007–2018, 2025.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

A Philip Dawid. Calibration-based empirical probability. *The Annals of Statistics*, 13(4):1251–1274, 1985.

Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

Gabriele Farina and Juan Carlos Perdomo. An efficient black-box reduction from online learning to multicalibration, and a new route to ϕ-regret minimization. *arXiv preprint arXiv:2604.19592*, 2026.

Dean P Foster. A proof of calibration via blackwell's approachability theorem. *Games and Economic Behavior*, 29(1-2):73–78, 1999.

Dean P Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games and Economic Behavior*, 109:271–293, 2018.

Dean P Foster and Sham M Kakade. Calibration via regression. In *Proc. IEEE Information Theory Workshop (ITW), Punta del Este*, pages 82–86. IEEE, 2006.

Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, 1997.

Dean P Foster, Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Complexity-based approach to calibration with checking rules. In *Proc. Conference on Learning Theory (COLT)*, pages 293–314. JMLR Workshop and Conference Proceedings, 2011.

Drew Fudenberg and David K Levine. An easier way to calibrate. *Games and economic behavior*, 29(1-2):131–137, 1999.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proc. Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2725–2792. SIAM, 2024.

Rohan Ghuge, Vidya Muthukumar, and Sahil Singla. Improved and oracle-efficient online ℓ_1-multicalibration. In *Proc. International Conference on Machine Learning (ICML)*, pages 19437–19457. PMLR, 2025.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *Proc. International Conference on Machine Learning (ICML)*, pages 3942–3952. PMLR, 2021.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *Proc. Innovations in Theoretical Computer Science Conference (ITCS)*, pages 82:1–82:24. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022.

Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 72464–72506, 2023.

Sergiu Hart. Calibrated forecasts: The minimax proof. In *Matching, Dynamics and Games for the Allocation of Resources: Essays in Celebration of David Gales 100th Birthday*, pages 153–159. Springer, 2025.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. Technical Report TR07-088, Electronic Colloquium on Computational Complexity (ECCC), 2007.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proc. International Conference on Machine Learning (ICML)*, pages 1939–1948. PMLR, 2018.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

Lunjia Hu, Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Efficient swap multicalibration of elicitable properties. *arXiv preprint arXiv:2511.04907*, 2025.

Lunjia Hu, Jon Schneider, and Yifan Wu. Near-optimal swap regret minimization for convex losses. *arXiv preprint arXiv:2602.08862*, 2026.

Shayan Kiyani, Hamed Hassani, George J. Pappas, and Aaron Roth. Robust decision making with partially calibrated forecasts. In *Proc. International Conference on Learning Representations (ICLR)*, 2026.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *Proc. Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 5143–5145. PMLR, 2023.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, pages 681–690, 2008.

Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *Journal of Machine Learning Research*, 21(137):1–45, 2020.

Daniel Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 29051–29063, 2022.

Ehud Lehrer. Any inspection is manipulable. *Econometrica*, 69(5):1333–1347, 2001.

Ehud Lehrer. Approachability in infinite dimensional spaces. *International Journal of Game Theory*, 31(2):253–268, 2003.

Junyan Liu, Haipeng Luo, and Lillian J Ratliff. Adaptive calibration in non-stationary environments. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: Adanormalhedge. In *Proc. Conference on Learning Theory (COLT)*, pages 1286 – 1304. PMLR, 2015.

Georgy Noarov and Aaron Roth. The statistical scope of multicalibration. In *Proc. International Conference on Machine Learning (ICML)*, pages 26283–26310. PMLR, 2023.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. In *Proc. International Conference on Machine Learning (ICML)*, 2025.

Chara Podimata and Alex Slivkins. Adaptive discretization for adversarial lipschitz bandits. In *Proc. Conference on Learning Theory (COLT)*, pages 3788–3805. PMLR, 2021.

Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Proc. Annual ACM Symposium on Theory of Computing (STOC)*, pages 456–466, 2021.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation. Lecture notes, 2022. URL `https://www.cis.upenn.edu/~aaroth/uncertainty-notes.pdf`.

Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28(1):141–153, 2003.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 13331–13340, 2020.

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proc. Conference on Learning Theory (COLT)*, pages 679–702. JMLR Workshop and Conference Proceedings, 2011.

Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262 – 270, 2011.

A Interaction Protocol Defined with Filtration

In this section, we describe the interaction protocol between learner and Nature with filtration notation.

Formally, at round $t \in [T]$:

1. the learner observes a context $x_t \in \mathcal{X}$;
2. the learner outputs a mixed forecast π_t, i.e. a distribution over finitely many prediction values in $[0, 1]$;
3. a realized prediction $p_t \in [0, 1]$ is sampled from π_t using fresh private randomization that is not revealed to Nature before the outcome is generated;
4. Nature reveals an outcome $y_t \in [0, 1]$.

We assume that y_t may depend on the past, on the current context x_t, and on the mixed forecast π_t, but not on the fresh randomization used to sample p_t. That is, Nature may react to π_t, but not to the private coin flip that turns π_t into the realized prediction p_t.

We work with the following standard two-stage filtration. The sigma-field $\mathcal{F}_{t-1}$ contains the realized transcript through the end of round $t - 1$. Define the pre-outcome sigma-field

$$\mathcal{H}_{t-1} := \sigma(\mathcal{F}_{t-1}, x_t, \pi_t),$$

and the end-of-round history

$$\mathcal{F}_t := \sigma(\mathcal{H}_{t-1}, p_t, y_t).$$

Thus $\mathcal{H}_{t-1}$ contains the information available before the outcome is generated, while $\mathcal{F}_t$ contains the full transcript through round t. Conditional on $\mathcal{H}_{t-1}$, the fresh randomization used to sample p_t has law π_t and is independent of y_t. Define the next pre-outcome sigma-field, for $t < T$, by

$$\mathcal{H}_t := \sigma(\mathcal{F}_t, x_{t+1}, \pi_{t+1}),$$

so that

$$\mathcal{H}_{t-1} \subseteq \mathcal{F}_t \subseteq \mathcal{H}_t$$

for every $t < T$. For notational convenience set $\mathcal{H}_T := \mathcal{F}_T$, so that the same interleaving convention holds for all $t \in [T]$. Define the predictable conditional mean $\mu_t := \mathbb{E}[y_t \mid \mathcal{H}_{t-1}] \in [0,1]$. This is just notation for the conditional mean of the realized outcome sequence. In particular, if Nature chooses y_t deterministically from $\mathcal{H}_{t-1}$, then $\mu_t = y_t$.

B Marginal Calibration Intuition

We explain the proof idea in the simplest special case: the context space is a singleton and the only group is the all-ones group. In this case multicalibration is just marginal calibration. The general multicalibration proof follows the same template (Section 5 and Appendix G.1).

Step 1: Regret controls per-interval bias. Fix an active interval I with midpoint r_I and width w_I. The sleeping-experts wrapper tracks whether predictions from this interval are accumulating positive or negative bias. The weighted feasibility step chooses π_t so that the aggregate one-step bias is nonpositive for every possible next outcome. As a result, over any contiguous block S inside the lifetime of I,

$$\left|\sum_{t\in S} \pi_{t,I}(y_t - r_I)\right| \lesssim N_S(I)w_I + \sqrt{N_S(I)L} + L.$$

The first term is deterministic discretization error: an interval of width w_I cannot distinguish outcomes that differ by less than its own resolution. The remaining terms are the regret cost of controlling one-sided bias on that interval. The bound is Lemma 5. This also explains the split rule. The algorithm refines I when $N(I) \asymp L/w_I^2$, the point where the discretization term $N(I)w_I$ and the regret term $\sqrt{N(I)L}$ are of the same order.

Step 2: Far away intervals accumulate bias quickly Write $q_t := \mu_t$ for the scalar predictable mean at time t. Now suppose that on a time block S, the mean sequence q_t is well approximated by a constant $c_S \in [0,1]$. If a depth-d interval has midpoint r_I far above c_S, then putting probability on that interval tends to produce negative bias. If r_I is far below c_S, it tends to produce positive bias. Step 1 says such one-sided bias cannot persist, so far-away intervals cannot receive too much mass.

If q_t is only approximately constant, this argument weakens exactly on rounds where q_t is not close to c_S. Errors smaller than the bin width w_d are below the resolution of a depth-d interval, so they are free in our accounting. The relevant residual on S at depth d is therefore

$$\mathsf{R}_{S,d}(c_S) := \sum_{t\in S}(|q_t - c_S| - w_d)_+.$$

How frequently a far interval can be played can be bounded by two terms: a square-root deviation term from regret and concentration, and a residual term measuring how much of $\mathsf{R}_{S,d}(c_S)$ is assigned to that interval (Lemma 7).

Step 3: Residual error controls how many intervals split. Only depths $d < d_{\max}$ split. A depth-d interval splits only after accumulating total probability mass on the order of L/w_d^2. Around the local constant c_S, there may be a band of nearby intervals for which the bias argument is too weak. Farther away, Step 2 gives a bound on how much probability mass an interval can receive.

The analysis balances these two effects. If we declare a wider band around c_S to be "near," we pay for more nearby intervals. If we declare a narrower band to be near, we must charge more of the far-away play to residual error. Optimizing this tradeoff bounds the block's capped split charge at depth d on the order of

$$1 + \sqrt{\mathsf{R}_{S,d}(c_S)w_d/L}$$

for a single block S. This is $\Xi_d(S)$ from Lemma 8. Summing charges over blocks bounds split parents even when their lifetimes cross block boundaries. If the block is exactly stationary at scale w_d, then the residual is zero and the block contributes only constant capped split charge at that depth. Summing these scale-by-scale split bounds is what ultimately produces the cube-root residual term in the main theorem.

Step 4: Depth-wise active intervals control calibration error. The tree contains intervals at many resolutions. A coarse interval has large discretization error but can only appear in small numbers. A fine interval has small width but can be expensive to control if many such intervals are active. The proof therefore accounts for calibration separately at each depth. If m_d depth-d intervals ever become active, their total contribution is bounded by

$$C \min\{\sqrt{T\, m_d L},\, m_d L / w_d\}.$$

The first term is the cost of summing square-root deviation terms over the active intervals at that depth. The second uses the fact that an interval is refined once its counter reaches the split threshold. Combining this depth-wise accounting with the split bound from Step 3 yields the per-segment rates in the main theorem (Lemmas 6 and 10).

Extension to multicalibration. For multicalibration, the local constant c_S is replaced by a contextual score $f_S(x)$ mapping context to label means. The role of "r_I is above or below c_S" is played by the threshold comparison $f_S(x_t) \geq r_I$. If these threshold comparisons can be represented as low-weight linear combinations of the groups, then the groupwise bias control from Step 1 applies to this more general case as well, where the bias control is worse by a factor of the weight of the linear representation.

C Ordinal Risk Strata as Threshold-Representable Scores

Example 1 (Ordinal risk strata)**.** *Suppose that on a block S, the predictable mean is well approximated by an ordinal K-level score*

$$f(x) = a_{c(x)}, \qquad a_1 < \cdots < a_K, \quad c(x) \in \{1, \ldots, K\}.$$

Think of $c(x)$ as a coarse risk stratum such as routine, elevated, high, or critical. Assume the group family contains the all-ones group and the cumulative tier indicators

$$g_j(x) = \mathbf{1}\{c(x) \geq j\}, \qquad j = 2, \ldots, K.$$

Then every threshold cut of f has constant representation cost. Fix any $r \in [0, 1]$. If $r \leq a_1$, then $f(x) \geq r$ for all realized contexts, and we take $h_r(x) = 1$. If $r > a_K$, then $f(x) < r$ for all realized contexts, and we take $h_r(x) = -1$. Otherwise, there is an index $j \in \{2, \ldots, K\}$ such that $a_{j-1} < r \leq a_j$. In this case,

$$f(x) \geq r \quad \Longleftrightarrow \quad c(x) \geq j,$$

so we take

$$h_r(x) = 2g_j(x) - 1.$$

Thus, on the realized contexts,

$$\begin{cases} h_r(x_t) = 1, & f(x_t) \geq r, \\ h_r(x_t) = -1, & f(x_t) < r. \end{cases}$$

Its coefficient cost is at most 3, since h_r is either 1, -1, or $2g_j - 1$. Therefore the ordinal score is $(K, 3, 0)$-threshold representable on S. In particular, it is also $(K, 3, \eta)$-threshold representable for any $\eta \in [0, 1]$.

D Concentration Inequality

D.1 Two-stage Freedman Inequality (Lemma 1)

Lemma 1 (Two-stage Freedman inequality)**.** *Let $(\mathcal{H}_{t-1}, \mathcal{F}_t)_{t=1}^n$ be a two-stage filtration with $\mathcal{H}_{t-1} \subseteq \mathcal{F}_t \subseteq \mathcal{H}_t$ for every t, and let Δ_t be $\mathcal{F}_t$-measurable random variables satisfying*

$$\mathbb{E}[\Delta_t \mid \mathcal{H}_{t-1}] = 0, \qquad |\Delta_t| \leq 1 \quad \textit{a.s.}$$

for every t. Then for every $x, v > 0$,

$$\Pr\left(\left|\sum_{t=1}^n \Delta_t\right| \geq x \text{ and } \sum_{t=1}^n \mathbb{E}[\Delta_t^2 \mid \mathcal{H}_{t-1}] \leq v\right) \leq 2\exp\left(-\frac{x^2}{2(v + x/3)}\right).$$

Proof. We reduce the two-stage case to Freedman's inequality by constructing an interleaved filtration. Let $\widetilde{\mathcal{F}}_0 := \mathcal{H}_0$ and define $(\widetilde{\mathcal{F}}_k)_{k=1}^{2n}$ as follows:

$$\widetilde{\mathcal{F}}_{2t-1} := \mathcal{H}_{t-1}, \qquad \widetilde{\mathcal{F}}_{2t} := \mathcal{F}_t, \qquad t = 1, \dots, n.$$

Correspondingly, define the sequence of increments $(D_k)_{k=1}^{2n}$ by $D_{2t-1} := 0$ and $D_{2t} := \Delta_t$. Let $M_k := \sum_{s=1}^{k} D_s$ be the associated partial sums. By construction, D_k is $\widetilde{\mathcal{F}}_k$-measurable and satisfies:

$$\mathbb{E}[D_k \mid \widetilde{\mathcal{F}}_{k-1}] = \begin{cases} \mathbb{E}[0 \mid \widetilde{\mathcal{F}}_{k-1}] = 0 & \text{if } k = 2t-1, \\ \mathbb{E}[\Delta_t \mid \mathcal{H}_{t-1}] = 0 & \text{if } k = 2t. \end{cases}$$

Then $(M_k, \widetilde{\mathcal{F}}_k)$ is a martingale, its increments are bounded by 1 in absolute value, and its predictable quadratic variation at time $2n$ is

$$\langle M \rangle_{2n} = \sum_{k=1}^{2n} \mathbb{E}[D_k^2 \mid \widetilde{\mathcal{F}}_{k-1}] = \sum_{t=1}^{n} \mathbb{E}[\Delta_t^2 \mid \mathcal{H}_{t-1}].$$

Apply the standard one-sided Freedman inequality to M_k and to $-M_k$, then union bound the two tails; see Tropp [2011, Theorem 1.1]. □

D.2 The Global Concentration Event

D.2.1 Sampling Error (Lemma 2)

Lemma 2 (Sampling error)**.** *There exists an event $\mathcal{E}_{\mathrm{samp}}$ of probability at least $1 - \frac{1}{4T}$ such that, on $\mathcal{E}_{\mathrm{samp}}$, for every $g \in \bar{\mathcal{G}}$ and every interval $I \in \mathcal{D}$,*

$$\left| \sum_{t=1}^{T} \big(\mathbf{1}[p_t = r_I] - \pi_{t,I}\big) g(x_t)(y_t - r_I) \right| \le C_{\mathrm{samp}} \big(\sqrt{N_{[T]}^g(I) L} + L \big).$$

Proof. Fix a group $g \in \bar{\mathcal{G}}$ and an interval $I \in \mathcal{D}$. To bound the sampling error, we analyze the concentration of the sequence

$$Z_t^g(I) := \big(\mathbf{1}[p_t = r_I] - \pi_{t,I}\big) g(x_t)(y_t - r_I).$$

The proof is based on an application of Freedman's inequality with the peeling technique to handle the random conditional variance.

Step 1: Martingale properties and variance bounds. Let $M_T^g(I) := \sum_{t=1}^{T} Z_t^g(I)$. By the protocol, $Z_t^g(I)$ is $\mathcal{F}_t$-measurable and satisfies the martingale property $\mathbb{E}[Z_t^g(I) \mid \mathcal{H}_{t-1}] = 0$. The increments are bounded by $|Z_t^g(I)| \le 1$. The conditional variance is bounded as follows:

$$\mathbb{E}[(Z_t^g(I))^2 \mid \mathcal{H}_{t-1}] \le g(x_t)\pi_{t,I},$$

which holds because $g(x_t) \in \{0, 1\}$, $(y_t - r_I)^2 \le 1$, and the conditional variance of the indicator $\mathbf{1}[p_t = r_I]$ is $\pi_{t,I}(1 - \pi_{t,I}) \le \pi_{t,I}$. Consequently, the predictable quadratic variation satisfies:

$$\begin{aligned} V_T^g(I) &:= \sum_{t=1}^{T} \mathbb{E}[(Z_t^g(I))^2 \mid \mathcal{H}_{t-1}] \\ &= \sum_{t=1}^{T} \mathbb{E}\left[\big(\mathbf{1}[p_t = r_I] - \pi_{t,I}\big)^2 g(x_t)^2 (y_t - r_I)^2 \mid \mathcal{H}_{t-1} \right] \\ &\le \sum_{t=1}^{T} \mathbb{E}\left[\pi_{t,I}(1 - \pi_{t,I}) g(x_t) \mid \mathcal{H}_{t-1} \right] \\ &\le \sum_{t=1}^{T} g(x_t)\pi_{t,I} = N_{[T]}^g(I), \end{aligned}$$

where the first inequality is because $g(x_t)^2 = g(x_t)$ and $(y_t - r_I)^2 \le 1$.

Step 2: Peeling over variance buckets. To handle the random nature of $N^g_{[T]}(I)$, we employ a peeling technique. Let $m_{\max} := \lceil \log_2 T \rceil$. For $m = 0, 1, \ldots, m_{\max}$, define the buckets:

$$\mathcal{U}_m := \{2^{m-1} < N^g_{[T]}(I) \le 2^m\},$$

with the convention $\mathcal{U}_0 = \{N^g_{[T]}(I) \le 1\}$. On $\mathcal{U}_m$ we have $V^g_T(I) \le 2^m$, so Lemma 1 gives, for every $x \ge 0$,

$$\Pr(|M^g_T(I)| \ge x \text{ and } \mathcal{U}_m) \le 2\exp\left(-\frac{x^2}{2(2^m + x/3)}\right).$$

Step 3: Union bound over buckets and (g, I) pairs. Fix $\delta \in (0, 1)$. For each bucket m, setting the threshold $x_m := \sqrt{2^{m+1}\Lambda} + \frac{2}{3}\Lambda$ where $\Lambda := \log(2(m_{\max} + 1)/\delta)$ ensures the tail probability is at most $\delta/(m_{\max} + 1)$. Since $\sqrt{2^{m+1}\Lambda} \le \sqrt{2 \cdot 2^m \Lambda} \le 2\sqrt{(N^g_{[T]}(I) \vee 1)\Lambda}$. Summing over $m = 0, \ldots, m_{\max}$, we obtain that with probability at least $1 - \delta$:

$$|M^g_T(I)| \le 2\sqrt{(N^g_{[T]}(I) \vee 1)\Lambda} + \frac{2}{3}\Lambda.$$

Using the inequality $\sqrt{N^g_{[T]}(I) \vee 1} \le \sqrt{N^g_{[T]}(I)} + 1$, and simplifying constants, we have:

$$|M^g_T(I)| \le 4\left(\sqrt{N^g_{[T]}(I)\Lambda} + \Lambda\right).$$

Finally, we perform a union bound over all $|\bar{\mathcal{G}}| \cdot |\mathcal{D}| \le 4|\bar{\mathcal{G}}|\sqrt{T}$ pairs of (g, I). Setting $\delta := (16|\bar{\mathcal{G}}|T^{3/2})^{-1}$ yields an event $\mathcal{E}_{\text{samp}}$ with $\Pr(\mathcal{E}_{\text{samp}}) \ge 1 - \frac{1}{4T}$. Since $\Lambda = O(\log(T|\bar{\mathcal{G}}|)) = O(L)$, we conclude that on $\mathcal{E}_{\text{samp}}$:

$$|M^g_T(I)| \le C\left(\sqrt{N^g_{[T]}(I)L} + L\right)$$

simultaneously for all $g \in \bar{\mathcal{G}}$ and $I \in \mathcal{D}$, for a suitable absolute constant C.

□

D.2.2 Grouped Martingale Concentration

Lemma 3 (Grouped martingale concentration)**.** *There exists an event $\mathcal{E}_{\text{conc}}$ of probability at least $1 - \frac{1}{4T}$ such that, on $\mathcal{E}_{\text{conc}}$, for every $g \in \bar{\mathcal{G}}$, every interval $I \in \mathcal{D}$, and every contiguous block $S \subseteq [T]$,*

$$\left|\sum_{t \in S} g(x_t)\pi_{t,I}(y_t - \mu_t)\right| \le C_{\text{conc}}\big(\sqrt{N^g_S(I)L} + L\big).$$

Proof. Fix a triple (g, I, S) where $g \in \bar{\mathcal{G}}$, $I \in \mathcal{D}$, and $S \subseteq [T]$ is a contiguous block. The analysis proceeds in three steps, mirroring the structure of Lemma 2.

Step 1: Martingale properties and variance bounds. Consider the process

$$X^g_t(I, S) := g(x_t)\pi_{t,I}(y_t - \mu_t)\mathbf{1}[t \in S].$$

This is a two-stage martingale difference sequence with respect to $(\mathcal{H}_{t-1}, \mathcal{F}_t)_t$. Since $\mu_t = \mathbb{E}[y_t \mid \mathcal{H}_{t-1}]$ and the weight $g(x_t)\pi_{t,I}\mathbf{1}[t \in S]$ is $\mathcal{H}_{t-1}$-measurable, we have:

$$\mathbb{E}[X^g_t(I, S) \mid \mathcal{H}_{t-1}] = g(x_t)\pi_{t,I}\mathbf{1}[t \in S] \cdot \mathbb{E}[y_t - \mu_t \mid \mathcal{H}_{t-1}] = 0.$$

The increments are bounded by $|X^g_t(I, S)| \le 1$. The predictable quadratic variation of the sum $M^g_S(I) := \sum_{t=1}^T X^g_t(I, S)$ satisfies:

$$V^g_S(I) := \sum_{t=1}^T \mathbb{E}[(X^g_t(I, S))^2 \mid \mathcal{H}_{t-1}] \le \sum_{t \in S} g(x_t)\pi_{t,I} = N^g_S(I),$$

where we used the facts that $(y_t - \mu_t)^2 \le 1$, $g(x_t) \in \{0, 1\}$, and $\pi_{t,I} \in [0, 1]$.

Step 2: Peeling over variance buckets. To handle the random nature of $N_S^g(I)$, we employ a peeling argument. Let $m_{\max} := \lceil \log_2 T \rceil$. For each $m = 0, 1, \ldots, m_{\max}$, define the variance bucket:

$$\mathcal{U}_m := \{2^{m-1} < N_S^g(I) \leq 2^m\}, \quad \text{with } \mathcal{U}_0 := \{N_S^g(I) \leq 1\}.$$

On the event $\mathcal{U}_m$, the quadratic variation $V_S^g(I)$ is bounded by 2^m. Applying the two-stage Freedman inequality (Lemma 1), for any $x \geq 0$:

$$\Pr\left(|M_S^g(I)| \geq x \text{ and } \mathcal{U}_m\right) \leq 2\exp\left(-\frac{x^2}{2(2^m + x/3)}\right).$$

Step 3: Union bound over buckets and (g, I, S) triples. Fix $\delta \in (0, 1)$ and let $\Lambda := \log(2(m_{\max} + 1)/\delta)$. Setting the threshold $x_m := \sqrt{2^{m+1}\Lambda} + \frac{2}{3}\Lambda$ and summing over all $m_{\max} + 1$ buckets, we find that with probability at least $1 - \delta$:

$$|M_S^g(I)| \leq 4\left(\sqrt{N_S^g(I)\Lambda} + \Lambda\right).$$

There are at most $|\bar{\mathcal{G}}| \cdot |\mathcal{D}| \cdot T^2 \leq 4|\bar{\mathcal{G}}|T^{5/2}$ such triples (g, I, S). Setting $\delta := (16|\bar{\mathcal{G}}|T^{7/2})^{-1}$ and performing a union bound over all triples, we obtain the event $\mathcal{E}_{\text{conc}}$ with probability $\Pr(\mathcal{E}_{\text{conc}}) \geq 1 - 1/(4T)$.

Finally, as $m_{\max} = O(\log T)$ and $\log(1/\delta) = O(\log(T|\bar{\mathcal{G}}|))$, we have $\Lambda = O(L)$. Absorbing the absolute constants into C_{conc} yields the claimed bound.

□

E Algorithmic Properties

E.1 Weighted Feasibility (Lemma 4)

Lemma 4. *Let B be any finite partition of $[0, 1]$ into intervals I with midpoints r_I and widths w_I. Let $\{a_I, b_I\}_{I \in B}$ satisfy $|a_I| \leq b_I$ for every $I \in B$. Then there exists $\pi \in \Delta(B)$ such that*

$$\sum_{I \in B} \pi_I(a_I(y - r_I) - b_I w_I) \leq 0 \qquad \textit{for every } y \in [0, 1].$$

Proof. We reformulate the problem as a zero-sum game between a learner (choosing $\pi \in \Delta(B)$) and an adversary (choosing $y \in [0, 1]$). Let the payoff function be defined as

$$F(\pi, y) := \sum_{I \in B} \pi_I \left(a_I(y - r_I) - b_I w_I\right).$$

For any fixed $y \in [0, 1]$, let $I_y \in B$ denote the unique interval containing y. By the definition of the grid, we have $|y - r_{I_y}| \leq w_{I_y}/2$. Using the assumption $|a_{I_y}| \leq b_{I_y}$, we observe

$$\begin{aligned}
\min_{I \in B}\left(a_I(y - r_I) - b_I w_I\right) &\leq a_{I_y}(y - r_{I_y}) - b_{I_y} w_{I_y} \\
&\leq |a_{I_y}| \cdot \frac{w_{I_y}}{2} - b_{I_y} w_{I_y} \\
&\leq b_{I_y}\left(\frac{w_{I_y}}{2} - w_{I_y}\right) \\
&= -\frac{b_{I_y} w_{I_y}}{2} \leq 0.
\end{aligned}$$

The function F is thus bilinear, and the strategy spaces $\Delta(B)$ and $[0, 1]$ are compact convex sets. Thus, applying Sion's minimax theorem, we can swap the order of optimization:

$$\min_{\pi \in \Delta(B)} \max_{y \in [0,1]} F(\pi, y) = \max_{y \in [0,1]} \min_{\pi \in \Delta(B)} F(\pi, y) = \max_{y \in [0,1]} \min_{I \in B}\left(a_I(y - r_I) - b_I w_I\right) \leq 0$$

Consequently, there exists a fixed $\pi \in \Delta(B)$ such that $F(\pi, y) \leq 0$ for all $y \in [0, 1]$. □

E.2 AdaNormalHedge Guarantee

We now analyze the sleeping-experts wrapper from Section 2. Concretely, we instantiate it with the confidence-rated AdaNormalHedge algorithm of Luo and Schapire [2015].

Expert construction. Let $\mathcal{J} := \{(I,+),(I,-) : I \in \mathcal{D}\}$ be the set of coordinates representing the positive and negative parts of the multicalibration constraints, with $M := |\mathcal{J}|$ as the total number of coordinates. Since the interval partition satisfies $|\mathcal{D}| = \sum_{d=0}^{d_{\max}} 2^d \le 4\sqrt{T}$, the number of coordinates is bounded by $M \le 8\sqrt{T}$.

For every start time $s \in [T]$, group $g \in \bar{\mathcal{G}}$, and coordinate $j \in \mathcal{J}$, create a sleeping expert $e = (s, g, j)$ with prior

$$q_e := \frac{1}{|\bar{\mathcal{G}}| M H_{T,2}} \frac{1}{s^2}, \qquad H_{T,2} := \sum_{u=1}^{T} \frac{1}{u^2}.$$

Then $\sum_e q_e = 1$.

Loss for experts. To apply AdaNormalHedge, we map our calibration constraints into the range $[0,1]$. Let $G = 3/2$ be a uniform bound on $|\phi^{\pm}_{g,I}(\pi_t, y_t)|$. For an expert $e = (s, g, (I,+))$, the normalized loss on an awake round is defined as:

$$\ell_{t,e} := \frac{G - \phi^{+}_{g,I}(\pi_t, y_t)}{2G}.$$

For $e = (s, g, (I,-))$ we define the loss analogously with $\phi^{-}_{g,I}$ in place of $\phi^{+}_{g,I}$. Let

$$I_{t,e} := \mathbf{1}\{s \le t,\ I \in B_t\}$$

denote the sleeping-specialist confidence for expert e; in this paper these confidences are indicators, while the factors $g(x_t)\pi_{t,I}$ enter through the expert loss itself.

Loss for the wrapper. The wrapper instantiated with AdaNormalHedge maintains a set of weights $\omega_{t,e}$ over the experts e that are awake at round t (i.e., $I_{t,e} = 1$). The wrapper loss is the weighted average $\widehat{\ell}_t := \sum_{e:I_{t,e}=1} \omega_{t,e}\ell_{t,e}$. By our normalization, this corresponds to the predicted violation $\widehat{\phi}_t = \sum_{I \in B_t} \pi_{t,I}\,(a_{t,I}\,(y_t - r_I) - b_{t,I} w_I)$ from the potential-based update:

$$\widehat{\ell}_t = \frac{G - \widehat{\phi}_t}{2G}.$$

To quantify the performance relative to a specific expert e, we define the positive relative-loss term up to time b as:

$$\widetilde{L}_{b,e} := \sum_{t=1}^{b} I_{t,e}[\ell_{t,e} - \widehat{\ell}_t]_+.$$

This quantity measures, up to time b, how much worse expert e can be than the wrapper on the rounds when e is awake.

Regret guarantee. We now state the first-order performance guarantee of the AdaNormalHedge-based wrapper. The following bound is the confidence-rated/sleeping-specialist version of the first-order AdaNormalHedge guarantee.

Proposition 1 (First-order confidence-rated AdaNormalHedge bound)**.** *There exists an absolute constant $C_{\mathrm{anh}} > 0$ such that, simultaneously for every expert e and every horizon $b \le T$,*

$$\sum_{t=1}^{b} I_{t,e}(\widehat{\ell}_t - \ell_{t,e}) \le C_{\mathrm{anh}} \left(\sqrt{\widetilde{L}_{b,e} L} + L \right),$$

where $L = \log(eT|\bar{\mathcal{G}}|)$.

Proof. We combine the confidence-rated AdaNormalHedge bound of Luo and Schapire [2015, Theorem 3] with their first-order conversion in Luo and Schapire [2015, Theorem 2]. The total number of grouped sleeping experts is

$$N = |\bar{\mathcal{G}}| M T \le 8|\bar{\mathcal{G}}| T^{3/2}.$$

For expert $e = (s, g, j)$, define the instantaneous regret

$$r_{t,e} := I_{t,e}(\widehat{\ell}_t - \ell_{t,e}), \qquad R_{b,e} := \sum_{t=1}^{b} r_{t,e}, \qquad C_{b,e} := \sum_{t=1}^{b} |r_{t,e}|.$$

The confidence-rated AdaNormalHedge guarantee gives

$$R_{b,e} \le \sqrt{C_{b,e} A_{b,e}},$$

where $A_{b,e}$ captures the complexity cost of expert e. Moreover,

$$C_{b,e} = R_{b,e} + 2\widetilde{L}_{b,e}.$$

If $R_{b,e} < 0$, the claimed upper bound is trivial. Otherwise, solving $R_{b,e} \le \sqrt{(R_{b,e} + 2\widetilde{L}_{b,e}) A_{b,e}}$ gives

$$R_{b,e} \le \sqrt{2\widetilde{L}_{b,e} A_{b,e}} + A_{b,e},$$

which is the first-order form used below. According to Luo and Schapire [2015], this cost is:

$$A_{b,e} := 3\left(\log \frac{1}{q_e} + \log \mathsf{B}_b + \log(1 + \log N)\right),$$

where $\mathsf{B}_b \le \frac{5}{2} + \frac{3}{2}\ln(1+T) = O(\log T)$. Thus, our prior yields

$$\log \frac{1}{q_e} = \log|\bar{\mathcal{G}}| + \log M + 2\log s + \log H_{T,2} = O(\log(T|\bar{\mathcal{G}}|)) = O(L),$$

uniformly over all experts. Since also $\log \mathsf{B}_b = O(\log\log T)$ and $\log(1 + \log N) = O(\log\log(T|\bar{\mathcal{G}}|))$, we obtain $A_{b,e} \le C'L$ for a universal constant C'. Substituting into the regret inequality yields the claim. □

E.2.1 Bias Control via Regret (Lemma 5)

Lemma 5 (Bias control via regret)**.** *For every group $g \in \bar{\mathcal{G}}$, every dyadic interval I, and every contiguous block $S \subseteq A(I)$,*

$$\left|\sum_{t \in S} g(x_t)\pi_{t,I}(y_t - r_I)\right| \le N_S^g(I) w_I + C_{\mathrm{loc}}\left(\sqrt{N_S^g(I)L} + L\right).$$

Proof of Lemma 5. We prove the positive direction; the negative one is identical. Let

$$S = [a, b] \subseteq A(I),$$

and consider the expert

$$e = (a, g, (I, +)).$$

Up to horizon b, this expert is awake exactly on the rounds in S. The wrapper loss satisfies

$$\widehat{\ell}_t = \frac{G - \widehat{\phi}_t}{2G}.$$

Using the wrapper identity from Section 2,

$$\widehat{\phi}_t = \sum_{I' \in B_t} \pi_{t,I'}\big(a_{t,I'}(y_t - r_{I'}) - b_{t,I'} w_{I'}\big) \le 0$$

by the weighted-feasibility condition (1). Therefore the regret of the wrapper against e upper bounds the cumulative positive one-sided weighted bias:

$$\sum_{t=1}^{b} I_{t,e}(\widehat{\ell}_t - \ell_{t,e}) = \sum_{t \in S} \frac{\phi^+_{g,I}(\pi_t, y_t) - \widehat{\phi}_t}{2G} \ge \frac{1}{2G} \sum_{t \in S} \phi^+_{g,I}(\pi_t, y_t).$$

Proposition 1 yields

$$\sum_{t\in S}\phi^+_{g,I}(\pi_t,y_t)\le C\left(\sqrt{\widetilde{L}_{b,e}L}+L\right).$$

Finally, recall $N^g_S(I)=\sum_{t\in S}g(x_t)\pi_{t,I}$, and we have

$$\widetilde{L}_{b,e}\;=\;\sum_{t\in S}\left[\frac{\widehat{\phi}_t-\phi^+_{g,I}(\pi_t,y_t)}{2G}\right]_+\le\sum_{t\in S}\frac{(-\phi^+_{g,I}(\pi_t,y_t))_+}{2G}\le\frac{1}{2}\sum_{t\in S}g(x_t)\pi_{t,I}=\frac{1}{2}N^g_S(I),$$

because $\widehat{\phi}_t\le 0$ and $|\phi^+_{g,I}(\pi_t,y_t)|\le Gg(x_t)\pi_{t,I}$. Rearranging gives

$$\sum_{t\in S}g(x_t)\pi_{t,I}(y_t-r_I-w_I)\le C\left(\sqrt{N^g_S(I)L}+L\right).$$

This is the point where the first-order form of Proposition 1 is essential: although the expert is awake throughout S, rounds with $g(x_t)\pi_{t,I}=0$ have $\phi^+_{g,I}(\pi_t,y_t)=0$ and do not increase $\widetilde{L}_{b,e}$. Thus the regret term scales with $N^g_S(I)=\sum_{t\in S}g(x_t)\pi_{t,I}$ rather than with $|S|$ or the active lifetime of I. Adding back the width term yields

$$\sum_{t\in S}g(x_t)\pi_{t,I}(y_t-r_I)\le N^g_S(I)w_I+C\left(\sqrt{N^g_S(I)L}+L\right),$$

which is the desired positive one-sided bound. The negative direction follows from the expert $(a,g,(I,-))$. □

E.2.2 Depth-Wise Calibration Accounting

Lemma 6 (Depth-wise calibration accounting)**.** *Let $\mathcal{V}_{T,d}:=\mathcal{V}_T\cap\mathcal{D}_d$ be the set of depth-d intervals that are active in at least one prediction round, and let $m_d:=|\mathcal{V}_{T,d}|$. Assume $L\le T$. On $\mathcal{E}_{\mathrm{samp}}$, for every group $g\in\mathcal{G}$ and every depth d,*

$$\sum_{I\in\mathcal{V}_{T,d}}\left|\sum_{t:p_t=r_I}g(x_t)(y_t-r_I)\right|\le C\min\left\{\sqrt{TLm_d},\frac{Lm_d}{w_d}\right\}.$$

Proof of Lemma 6. Fix a group g and a depth d. We first control each bin and then sum over its depth.

Step 1: Bound each bin's error. Below maximum depth, the split threshold can be overshot by less than one round of mass, and the bin is then permanently retired. At maximum depth, $N(I)w_d^2\le Tw_{d_{\max}}^2\le 1$. Thus $N(I)w_d^2\le 2L$ for every ever-active bin. Local bias on the contiguous lifetime $A(I)$ and the sampling event give

$$\left|\sum_{t:p_t=r_I}g(x_t)(y_t-r_I)\right|\le C_{\mathrm{bin}}\big(\sqrt{N^g_{[T]}(I)L}+L\big),\quad C_{\mathrm{bin}}:=\sqrt{2}+C_{\mathrm{loc}}+C_{\mathrm{samp}},$$

because $N^g_{[T]}(I)\le N(I)$ implies $N^g_{[T]}(I)w_d\le\sqrt{2N^g_{[T]}(I)L}$.

Step 2: Prove the square-root branch. At a fixed depth, $\sum_{I\in\mathcal{V}_{T,d}}N^g_{[T]}(I)\le T$, so Cauchy–Schwarz bounds the square-root terms by $\sqrt{TLm_d}$. For $d\ge 1$, every ever-active bin has a split parent of width $2w_d$. Distinct parents have total assigned mass at most T, and each paid $L/(2w_d)^2$ before splitting. Consequently $m_d\le 8Tw_d^2/L$, and hence $Lm_d\le\sqrt{8}\sqrt{TLm_d}$. At the root, $m_0=1$ and $L\le T$ give $L\le\sqrt{TL}$. Thus the depth error is at most $(1+\sqrt{8})C_{\mathrm{bin}}\sqrt{TLm_d}$ in both cases.

Step 3: Prove the linear branch. The bound $N^g_{[T]}(I)\le 2L/w_d^2$ gives $\sqrt{N^g_{[T]}(I)L}\le\sqrt{2}L/w_d$ for each bin. Since $w_d\le 1$, the additive term also satisfies $L\le L/w_d$. Summing gives depth error at most $(1+\sqrt{2})C_{\mathrm{bin}}Lm_d/w_d$. Taking the smaller branch proves the lemma with a constant uniform in g,d,T. □

F Bounding Splits From Approximation Structure

F.1 Intervals Far From the Score

Lemma 7 (Intervals far from the score are rarely played)**.** *Condition on $\mathcal{E}_{conc}$. Fix a depth d, write $w = w_d$, let S be a nonempty contiguous block, and let f be $(K, B, \frac{1}{4})$-threshold representable on S. Let $C_f(S) := \{f(x_t) : t \in S\}$. For a depth-$d$ interval I with midpoint r_I, define*

$$\delta_I := \operatorname{dist}(r_I, C_f(S)), \qquad \mathsf{D}_{S,I} := \sum_{t \in S \cap A(I)} \pi_{t,I}\big(|\mu_t - f(x_t)| - w\big)_+.$$

Let $C_0 := 4$. There is an absolute $C > 0$ such that, whenever $\delta_I > C_0 B w$,

$$N_S(I) \le C\left[\frac{B^2 L}{(\delta_I - C_0 B w)^2} + \frac{B\mathsf{D}_{S,I}}{\delta_I - C_0 B w}\right].$$

The trivial bound $N_S(I) \le |S|$ always holds.

Proof of Lemma 7. Assume $\delta_I > 4Bw$ and write $S_I := S \cap A(I)$. This intersection is contiguous, and $N_S(I) = N_{S_I}(I)$ because the bin has no mass outside its lifetime. If S_I is empty, the claim is immediate. A nonempty represented block has $B \ge 3/4$, since $|h_r(x_t)| \ge 3/4$ and $|h_r(x_t)| \le \sum_g |\alpha_g| \le B$.

Step 1: Lower-bound the signed mean bias. Choose the threshold representation $h_{r_I} = \sum_g \alpha_g g$ of f. Its signs agree with $f(x_t) - r_I$, its magnitude is at least $3/4$, and its magnitude is at most B. For $\rho_t := (|\mu_t - f(x_t)| - w)_+$, this gives

$$h_{r_I}(x_t)(\mu_t - r_I) \ge \tfrac{3}{4}\delta_I - Bw - B\rho_t.$$

Therefore

$$A_\mu := \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(\mu_t - r_I) \ge (\tfrac{3}{4}\delta_I - Bw) N_S(I) - B\mathsf{D}_{S,I}.$$

Step 2: Upper-bound both bias terms in absolute value. Decompose

$$A_\mu = \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - r_I) - \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - \mu_t).$$

Expanding in groups, using $\sum_g |\alpha_g| \le B$ and $N^g_{S_I}(I) \le N_S(I)$, local bias and $\mathcal{E}_{\mathrm{conc}}$ give

$$\left|\sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - r_I)\right| \le B N_S(I) w + C_{\mathrm{loc}} B\big(\sqrt{N_S(I)L} + L\big),$$

$$\left|\sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - \mu_t)\right| \le C_{\mathrm{conc}} B\big(\sqrt{N_S(I)L} + L\big).$$

These inequalities hold for every admissible representation after fixing the common event. They give $|A_\mu| \le B N_S(I) w + (C_{\mathrm{loc}} + C_{\mathrm{conc}}) B(\sqrt{N_S(I)L} + L)$.

Step 3: Solve the play inequality. Let $C_* := \frac{4}{3}\max\{1, C_{\mathrm{loc}} + C_{\mathrm{conc}}\}$ and $\Delta_I := \delta_I - 4Bw > 0$. Comparing Step 1 with Step 2 and using $4 \ge 8/3$ gives

$$\Delta_I N_S(I) \le C_* B\big(\mathsf{D}_{S,I} + \sqrt{N_S(I)L} + L\big).$$

Young's inequality bounds the square-root term by $C_* B\sqrt{N_S(I)L} \le \Delta_I N_S(I)/2 + C_*^2 B^2 L/(2\Delta_I)$. Rearranging yields

$$N_S(I) \le \frac{C_*^2 B^2 L}{\Delta_I^2} + \frac{2C_* B L}{\Delta_I} + \frac{2C_* B \mathsf{D}_{S,I}}{\Delta_I}.$$

Since $\Delta_I \le 1$ and $B \ge 3/4$, the first two terms are at most $(C_*^2 + \frac{8}{3}C_*)B^2 L/\Delta_I^2$. Taking $C = \max\{C_*^2 + \frac{8}{3}C_*, 2C_*\}$ proves the bound. The independent inequality $N_S(I) \le |S|$ always holds. □

F.2 Split Count on One Block

Lemma 8 (Split count on one block)**.** *Condition on $\mathcal{E}_{conc}$. Fix depth d, write $w = w_d$, and let S be a nonempty contiguous block. If f is $(K, B, \frac{1}{4})$-threshold representable on S, then*

$$\Xi_d(S) := \sum_{I \in \mathcal{D}_d} \min\left\{1, \frac{N_S(I) w_d^2}{L}\right\} \le C\left(BK + \sqrt{\frac{BK\,\mathsf{R}_{S,d}(f)\,w}{L}}\right).$$

Proof of Lemma 8. Write $C_f(S) = \{c_1, \dots, c_{K'}\}$, where $K' \le K$, and let $R := \mathsf{R}_{S,d}(f)$. Assign every interval to one nearest score, breaking ties by its index, and write $\delta_I = |r_I - c_{k(I)}|$.

Step 1: Count near intervals. Choose

$$Q := 16B + \sqrt{\frac{BRw}{KL}}.$$

Since $B \ge 3/4$ and $K \ge 1$, we have $Q \ge 12$ and $Q \ge 2C_0B$. At most $2Q + 2 \le 3Q$ depth-d midpoints lie within distance Qw of any one score. Every capped summand is at most one, so intervals with $\delta_I \le Qw$ contribute at most $3KQ$.

Step 2: Sum all far distance shells. For integers $\ell \ge \lfloor Q \rfloor$, define

$$\mathcal{S}_{k,\ell} := \{I : k(I) = k,\ \delta_I > Qw,\ \ell w < \delta_I \le (\ell+1)w\}.$$

These shells include the possibly partial first shell when Q is not an integer, and partition the far region. Each shell contains at most four midpoints. For every member, $\Delta_I = \delta_I - C_0 B w \ge \delta_I/2 \ge \ell w/2$ and $\ell \ge Q/2$. The preceding lemma therefore gives

$$\min\{1, N_S(I) w^2/L\} \le C\left[\frac{B^2}{\ell^2} + \frac{B\mathsf{D}_{S,I} w}{\ell L}\right].$$

The first terms sum to at most $CB^2K/Q \le CBK$, because $\sum_{\ell \ge \lfloor Q \rfloor} \ell^{-2} \le 1/(\lfloor Q \rfloor - 1) \le 2/Q$. For the residual terms, at each round the total probability assigned to this depth is at most one, so

$$\sum_{I \in \mathcal{D}_d} \mathsf{D}_{S,I} = \sum_{t \in S} (|\mu_t - f(x_t)| - w)_+ \sum_{I \in \mathcal{D}_d} \pi_{t,I} \le R.$$

They therefore sum to at most $CBRw/(QL)$.

Step 3: Substitute the cutoff. Together the near and far bounds give $\Xi_d(S) \le C(KQ + BRw/(QL))$, with a constant depending only on the preceding lemma. For $u := \sqrt{BRw/(KL)}$, the two terms satisfy

$$KQ = 16BK + Ku, \qquad \frac{BRw}{QL} = \frac{Ku^2}{16B + u} \le Ku.$$

This also covers $R = 0$, when $u = 0$. Since $Ku = \sqrt{BKRw/L}$, the claim follows. □

F.3 Active Intervals From Split Bounds

Let $\mathcal{A}_d$ be the set of depth-d bins that actually split by the end of round T.

Lemma 9 (Active intervals from split bounds)**.** *Condition on $\mathcal{E}_{\mathrm{conc}}$ and let $m_d = |\mathcal{V}_{T,d}|$. Then*

$$m_0 = 1, \qquad m_d \le C \min\left\{2^d, \frac{T w_d^2}{L}, \Psi^{\mathrm{res}}_{d-1}(\mu_{1:T}; \mathcal{G})\right\} \quad (d \ge 1).$$

Proof of Lemma 9. The root is the only active depth-0 interval, so $m_0 = 1$.

Step 1: Prove the deterministic caps. For $d \geq 1$, the grid has only 2^d depth-d intervals, hence $m_d \leq 2^d$. For the resource bound, each active depth-d interval is created by a split of a depth-$(d-1)$ parent. Such a parent must have accumulated at least L/w_{d-1}^2 total mass before splitting. The total mass accumulated by all depth-$(d-1)$ intervals over the horizon is at most T, so the number of such parents is at most Tw_{d-1}^2/L, and therefore

$$m_d \leq 2Tw_{d-1}^2/L = 8Tw_d^2/L.$$

Step 2: Charge split parents to temporal blocks. Every active depth-d interval is a child of a split depth-$(d-1)$ interval, so $m_d \leq 2|\mathcal{A}_{d-1}|$. Fix any admissible partition $\mathcal{P}_{d-1}$ and triples $\{(f_S, K_S, B_S) : S \in \mathcal{P}_{d-1}\}$ for $\Psi_{d-1}^{\mathrm{res}}$. If $I \in \mathcal{A}_{d-1}$, then $N(I) \geq L/w_{d-1}^2$. Since $\mathcal{P}_{d-1}$ partitions $[T]$,

$$N(I) = \sum_{S \in \mathcal{P}_{d-1}} N_S(I),$$

and therefore

$$1 \leq \sum_{S \in \mathcal{P}_{d-1}} \min\left\{1, \frac{N_S(I) w_{d-1}^2}{L}\right\}.$$

Summing over $I \in \mathcal{A}_{d-1}$ and exchanging sums,

$$|\mathcal{A}_{d-1}| \leq \sum_{S \in \mathcal{P}_{d-1}} \sum_{I \in \mathcal{D}_{d-1}} \min\left\{1, \frac{N_S(I) w_{d-1}^2}{L}\right\}.$$

Applying Lemma 8 to each block at depth $d-1$ gives

$$|\mathcal{A}_{d-1}| \leq C \sum_{S \in \mathcal{P}_{d-1}} \left[B_S K_S + \sqrt{\frac{B_S K_S \, \mathsf{R}_{S,d-1}(f_S) w_{d-1}}{L}}\right].$$

Step 3: Take the infimum and record the leaf cap. Taking the infimum over $\mathcal{P}_{d-1}$ and the scores gives $|\mathcal{A}_{d-1}| \leq C\Psi_{d-1}^{\mathrm{res}}(\mu_{1:T}; \mathcal{G})$, and hence $m_d \leq C\Psi_{d-1}^{\mathrm{res}}(\mu_{1:T}; \mathcal{G})$. Combining the three bounds proves the lemma. For the implementation bound, an actual split at depth d costs L/w_d^2, so $|\mathcal{A}_d| \leq \min\{2^d, T/(L4^d)\}$. The final tree, including end-of-round T splits, therefore satisfies

$$\Lambda_T = 1 + \sum_{d < d_{\max}} |\mathcal{A}_d| \leq 1 + \sum_{d \geq 0} \min\{2^d, T/(L4^d)\} \leq 1 + 4(T/L)^{1/3}.$$

For $T/L \geq 1$, split the sum at $\lfloor \log_2(T/L)/3 \rfloor$ and sum the ratios 2 and $1/4$. For $T/L < 1$, its sum is at most $4T/(3L) \leq 4(T/L)^{1/3}$. Since $|B_t| \leq \Lambda_T$, this is the resource bound used in the implementation paragraph. □

F.4 Dyadic Summation

Lemma 10 (Dyadic summation)**.** *Let $w_d = 2^{-d}$, and let $M \geq 1$ and $A \geq 0$. Suppose $m_0 \leq C$ and, for every $d \geq 1$,*

$$m_d \leq C \min\left\{2^d, \frac{Tw_d^2}{L}, M + A\sqrt{\frac{w_d}{L}}\right\}.$$

Then

$$\sum_{d=0}^{d_{\max}} \min\left\{\sqrt{TLm_d}, \frac{Lm_d}{w_d}\right\} \leq C\left(\sqrt{TLM} + T^{1/3}L^{1/3}A^{2/3}\right).$$

Proof of Lemma 10. Let $C_* \geq 1$ be the constant in the hypothesis. The root contributes at most $\sqrt{C_* TL} \leq C_* \sqrt{TLM}$ because $M \geq 1$.

Step 1: Separate the two depth contributions. For $d \geq 1$, use the mass cap in the square-root branch and the profile bound in the linear branch to obtain

$$\min\{\sqrt{TLm_d}, Lm_d/w_d\} \leq C_* \min\{Tw_d, LM/w_d + A\sqrt{L/w_d}\}.$$

The inequality $\min\{a, b+c\} \leq \min\{a,b\} + \min\{a,c\}$ bounds this by

$$C_*\big[\min\{Tw_d, LM/w_d\} + \min\{Tw_d, A\sqrt{L/w_d}\}\big].$$

Step 2: Sum the threshold-cost contribution. The terms Tw and LM/w cross at $w_M = \sqrt{LM/T}$, with common value $\sqrt{TLM}$. The terms on each side form geometric tails of ratios $1/2$ away from the crossing. Extending the available depths if necessary gives

$$\sum_{d\geq 1} \min\{Tw_d, LM/w_d\} \leq 4\sqrt{TLM}.$$

This also covers a crossover outside the actual depth range, since discarding tail terms decreases the sum.

Step 3: Sum the residual contribution. If $A = 0$, this contribution is zero. Otherwise the crossover is $w_A = (A\sqrt{L}/T)^{2/3}$, with common value $T^{1/3}L^{1/3}A^{2/3}$. The fine-depth tail has ratio $1/2$ and the coarse-depth tail has ratio $2^{-1/2}$ away from the crossing, so

$$\sum_{d\geq 1} \min\{Tw_d, A\sqrt{L/w_d}\} \leq \left(2 + \frac{1}{1-2^{-1/2}}\right) T^{1/3}L^{1/3}A^{2/3}.$$

The same extension argument covers either endpoint case.

Step 4: Assemble the bound. Add the root, Step 2 and Step 3. Their constants depend only on C_*, not on T, L, M, A or the number of depths. This proves the stated sum. □

G Proofs of the Main Theorem and Corollaries

G.1 Proof of the Main Theorem

Proof of Theorem 1. We treat the boundary case, combine the depth bounds, and then substitute the segmented costs.

Step 1: Handle large logarithmic parameters. If $L > T$, then the trivial bound $\mathrm{MCerr}_{\mathcal{G}}(T) \leq T$ is covered by the worst-case term $T^{2/3}L^{1/3}$; it is also covered by the depth-wise branch because the depth-0 term is at least T. The segmented branch is also at least $\sqrt{TL} \geq T$ because $\mathsf{M} \geq 1$. Hence assume $L \leq T$ and condition on $\mathcal{E}_{\mathrm{global}}$.

Step 2: Assemble the depth-wise bound. Each realized prediction value is the midpoint of a unique interval that is active at some time. Group these ever-active intervals by their depth. For any fixed group $g \in \mathcal{G}$,

$$\sum_{v} \left| \sum_{t:p_t=v} g(x_t)(y_t - v) \right| = \sum_{d=0}^{d_{\max}} \sum_{I\in\mathcal{V}_{T,d}} \left| \sum_{t:p_t=r_I} g(x_t)(y_t - r_I) \right|.$$

Applying Lemma 6 at each depth and then maximizing over g gives

$$\mathrm{MCerr}_{\mathcal{G}}(T) \leq C \sum_{d=0}^{d_{\max}} \min\left\{ \sqrt{TLm_d}, \frac{Lm_d}{w_d} \right\}.$$

Lemma 9 gives $m_d \leq C\Theta_d$ for every depth d. Since the count constant is uniform in depth, increasing C by that fixed factor proves the depth-wise bound.

Step 3: Recover the worst-case cap. The deterministic parts of Θ_d also give the worst-case branch. Indeed, for $d \geq 1$ we use only

$$\Theta_d \leq 2^d = \frac{1}{w_d}, \qquad \Theta_d \leq \frac{Tw_d^2}{L}.$$

Thus the dth summand is at most

$$\min\left\{ \sqrt{\frac{TL}{w_d}}, Tw_d \right\}.$$

Let $w_* := (L/T)^{1/3}$, where the two terms are equal. For depths with $w_d \geq w_*$, the terms $\sqrt{TL/w_d}$ form a geometrically increasing sequence as d increases, so their total contribution is at most a constant times their value at w_*. For depths with $w_d < w_*$, the terms Tw_d form a geometrically decreasing tail, so their total contribution is at most a constant times Tw_*. Since

$$\sqrt{\frac{TL}{w_*}} = Tw_* = T^{2/3}L^{1/3},$$

and the depth-0 term is $O(\sqrt{TL}) \leq O(T^{2/3}L^{1/3})$ when $L \leq T$, we obtain

$$\sum_{d=0}^{d_{\max}} \min\left\{\sqrt{TL\Theta_d}, \frac{L\Theta_d}{w_d}\right\} \leq CT^{2/3}L^{1/3}$$

as claimed.

Step 4: Substitute the segmented costs on the same event. For the segmented consequence, fix an admissible partition $\mathcal{P}$ and triples $(f, K, B) = \{(f_S, K_S, B_S) : S \in \mathcal{P}\}$. For every $d \geq 1$, since $\mathsf{R}_{S,d-1}(f_S) \leq \mathsf{R}_S(f_S)$,

$$\Psi^{\mathrm{res}}_{d-1}(\mu_{1:T}; \mathcal{G}) \leq \sum_{S\in\mathcal{P}} \left[B_S K_S + \sqrt{\frac{B_S K_S \,\mathsf{R}_S(f_S)\, w_{d-1}}{L}}\right] \leq \mathsf{M}(\mathcal{P}, f) + \sqrt{2}\,\mathsf{A}(\mathcal{P}, f)\sqrt{\frac{w_d}{L}},$$

where the last step uses $w_{d-1} = 2w_d$. Combining this with Lemma 9, the active interval counts satisfy the hypothesis of Lemma 10 with $M = \mathsf{M}(\mathcal{P}, f)$ and A equal to an absolute-constant multiple of $\mathsf{A}(\mathcal{P}, f)$. Applying that lemma gives the segmented bound. The worst-case cap is the first branch proved above. The group/bin concentration event and the local regret inequalities hold for all contiguous blocks before any partition, score or representation coefficients are chosen. All later steps are pathwise, so the result holds simultaneously for every admissible choice on that same event. □

G.2 Proof of Corollary 1

Proof of Corollary 1. Use the depth-wise part of Theorem 1. For each depth d, every admissible exact threshold-simple partition has zero scale-truncated residual, hence $\Psi^{\mathrm{res}}_d(\mu_{1:T}; \mathcal{G}) \leq M^{\mathrm{thr}}_d(\mu_{1:T}; \mathcal{G})$. Therefore, for $d \geq 1$, $\Theta_d \leq M^{\mathrm{thr}}_{d-1}$, while $\Theta_0 = 1$. The depth-wise sum is bounded by

$$\sqrt{TL} + \sum_{d=1}^{d_{\max}} \sqrt{TL\, M^{\mathrm{thr}}_{d-1}} \leq \widetilde{O}\left(\sqrt{TL\left(1 + \Gamma^{\mathrm{thr}}_T(\mu_{1:T}; \mathcal{G})\right)}\right),$$

where the last step uses Cauchy–Schwarz over the $O(\log(eT))$ depths. Taking the minimum with the worst-case branch proves the first claim.

For the special case, the one-block partition at each depth gives $M^{\mathrm{thr}}_d \leq B_d K_d \leq BK$. Hence $\Gamma^{\mathrm{thr}}_T \leq O(BK \log(eT))$. Since $BK \geq 3/4$ on a nonempty represented horizon, $1 + \Gamma^{\mathrm{thr}}_T \leq O(BK \log(eT))$ as well. Substitution into the first claim proves the specialization. □

G.3 Proof of Corollary 2

Proof of Corollary 2. Work on the event of Theorem 1 that is simultaneous over all admissible score/partition choices. Let $S_1, \ldots, S_J$ be the nonempty contiguous segments partitioning $[T]$ in the assumption, and let $c_j \in [0, 1]$ denote the constant value of $q_t = \mu_t$ on S_j. Let $\mathcal{P} = \{S_1, \ldots, S_J\}$ and use the constant score $f_{S_j} \equiv c_j$ on each segment. Because $\mathbf{1} \in \bar{\mathcal{G}}$, each threshold sign is represented by $\mathbf{1}$ when $r \leq c_j$ and $-\mathbf{1}$ when $r > c_j$. Each score is therefore $(1, 1, \frac{1}{4})$-threshold representable and has zero residual. Thus $\mathsf{M}(\mathcal{P}, f) = 1 + J$ and $\mathsf{A}(\mathcal{P}, f) = 0$. The segmented part of Theorem 1 gives

$$\mathrm{MCerr}_{\mathcal{G}}(T) \leq C \min\left\{T^{2/3}L^{1/3}, \sqrt{TL(1 + J)}\right\}.$$

Since $1 + J \leq 2J$ and $L = \log(eT|\bar{\mathcal{G}}|)$, this proves the claim. □

G.4 Proof of Corollary 3

Proof of Corollary 3. Work on the event of Theorem 1 that is simultaneous over all admissible score/partition choices. Choose $c^* \in \arg\min_{c\in[0,1]} \sum_{t=1}^{T} |\mu_t - c|$, which exists by continuity on $[0,1]$. Use the one-block partition $\mathcal{P} = \{[T]\}$ and the constant score $f \equiv c^*$. Because $\mathbf{1} \in \bar{\mathcal{G}}$, the threshold signs of f are represented by $\mathbf{1}$ when $r \le c^*$ and $-\mathbf{1}$ when $r > c^*$. Thus f is $(1, 1, \frac{1}{4})$-threshold representable on $[T]$, regardless of the contexts. The segmented costs are $\mathsf{M}(\mathcal{P}, f) = 2$ and $\mathsf{A}(\mathcal{P}, f) = \sqrt{C_{\mathrm{stat}}(\mu)}$. Substitution into Theorem 1 gives

$$\mathrm{MCerr}_{\mathcal{G}}(T) \le C\left(\sqrt{2TL} + (TLC_{\mathrm{stat}}(\mu))^{1/3}\right).$$

Here $L = \log(eT|\bar{\mathcal{G}}|)$, so this is the claimed bound. □

H Unknown Horizon by Restarting

We prove the expectation guarantee stated in Section 4 for a restarting wrapper around the finite-horizon algorithm. This changes only the wrapper: each epoch runs the original tree and experts with its own fixed horizon parameters. For $k = 0, 1, \ldots$, let $H_k := 2^k$ and run epoch k on rounds $H_k, \ldots, 2H_k - 1$. For any fixed deterministic observation time $T \ge 1$, let $K := \lfloor \log_2 T \rfloor$ and $E_k := [H_k, \min\{2H_k - 1, T\}]$ for $k \le K$. Define $L_k := \log(eH_k|\bar{\mathcal{G}}|)$. Let Γ_k be the threshold complexity of the observed block E_k, using depths $0, \ldots, d_{\max}(H_k)$ and local epoch time. Write

$$\Gamma_{\mathrm{ep}}(T) := \sum_{k=0}^{K}(1 + \Gamma_k), \qquad \bar{L} := \log(2eT|\bar{\mathcal{G}}|).$$

Then for a universal constant C,

$$\mathbb{E}[\mathrm{MCerr}_{\mathcal{G}}(T)] \le C\,\mathbb{E}\left[\min\left\{T^{2/3}\bar{L}^{1/3}, \bar{L}\sqrt{T\Gamma_{\mathrm{ep}}(T)}\right\}\right] + C\log(eT). \tag{2}$$

The expectation in (2) also covers the possibly random epoch-wise complexities.

Proof. We first control every observed epoch prefix and then aggregate its exact calibration error.

Step 1: Control the final partial epoch. Fix an epoch with planned horizon H and parameters $L_H = \log(eH|\bar{\mathcal{G}}|)$ and $d_{\max}(H)$. The sleeping-expert guarantee already holds for every endpoint $b \le H$, and the martingale event controls all contiguous blocks. For the sampling event, repeat the proof of Lemma 2 with a union bound over all $b \in [H]$ as well as groups, fixed dyadic bins, and variance buckets. There are at most $4|\bar{\mathcal{G}}|H^{3/2}(1 + \lceil \log_2 H \rceil)$ such choices. Assigning probability $1/(8H)$ in total to these sampling failures makes the Freedman logarithm at most

$$\log\Big(64|\bar{\mathcal{G}}|H^{5/2}(1 + \lceil \log_2 H \rceil)\Big) \le C_{\mathrm{pref}} L_H$$

for a universal C_{pref}. Together with the existing block event, this gives an event $\mathcal{E}_H$ of conditional failure probability at most C_{fail}/H, uniformly over the history before the epoch. The case $H = 1$ is covered directly by error at most one. For each local prefix $[b]$, $b \le H$, let $m_{b,d}$ count bins used at depth d. Let $M^{\mathrm{thr}}_{b,d}$ be its exact threshold cost and $\Gamma^{(H)}_b = \sum_{d=0}^{d_{\max}(H)} M^{\mathrm{thr}}_{b,d}$, using the planned depth range. Truncate each bin lifetime at b and retain the planned-H priors, depths and split thresholds. The scores realizing threshold simplicity have zero scale residual, so the active-profile proof gives

$$m_{b,0} = 1, \qquad m_{b,d} \le C\min\{2^d, Hw_d^2/L_H, M^{\mathrm{thr}}_{b,d-1}\} \quad (d \ge 1).$$

Only actual prefix blocks are used, whose concentration bounds are included in $\mathcal{E}_H$. Maximum-depth bins obey $N_{[b]}(I)w_I^2 \le Hw^2_{d_{\max}(H)} \le 1$. If $L_H \le H$, the depth-accounting proof with H as the total-mass upper bound therefore gives

$$\mathrm{MCerr}_{\mathcal{G}}(b) \le C \sum_{d=0}^{d_{\max}(H)} \min\{\sqrt{HL_H m_{b,d}}, L_H m_{b,d}/w_d\}.$$

The grid and mass caps yield $CH^{2/3}L_H^{1/3}$ by the same crossover as the main theorem. For the instance branch, let $D_H := d_{\max}(H) + 1 \le C' \log(eH) \le C' L_H$. Cauchy–Schwarz and $\sum_d m_{b,d} \le 1 + C\Gamma_b^{(H)}$ give

$$\mathrm{MCerr}_{\mathcal{G}}(b) \le C\sqrt{HL_H D_H(1+\Gamma_b^{(H)})} \le C'' L_H \sqrt{H(1+\Gamma_b^{(H)})}.$$

If $L_H > H$, both branches follow instead from $\mathrm{MCerr}_{\mathcal{G}}(b) \le b \le H$. Thus the observed part of epoch k has error at most

$$C_{\mathrm{epoch}} \min\{H_k^{2/3} L_k^{1/3}, L_k\sqrt{H_k(1+\Gamma_k)}\}$$

on $\mathcal{E}_{H_k}$, uniformly in the epoch and prefix length. The explicit logarithmic factor here includes the Cauchy–Schwarz sum over planned depths.

Step 2: Aggregate the actual prediction values. For any group g and realized value v, let $B_{k,g}(v) := \sum_{t \in E_k : p_t = v} g(x_t)(y_t - v)$. The triangle inequality gives

$$\max_g \sum_v \left| \sum_k B_{k,g}(v) \right| \le \sum_k \max_g \sum_v |B_{k,g}(v)|.$$

Thus reusing a prediction value in several epochs causes no problem. Outside $\mathcal{E}_{H_k}$, bound the epoch error by H_k. The total expected failure contribution is at most $C_{\mathrm{fail}}(K+1)$ because $\mathbb{E}[H_k \mathbf{1}_{\mathcal{E}_{H_k}^c}] \le C_{\mathrm{fail}}$. No independence between epoch failure events is required.

Step 3: Sum the epoch bounds. The planned horizons satisfy $\sum_{k=0}^{K} H_k = 2^{K+1} - 1 \le 2T$, including the unfinished final epoch, and $L_k \le \bar{L}$. Their worst-case terms form a geometric series, so

$$\sum_{k=0}^{K} H_k^{2/3} L_k^{1/3} \le \frac{T^{2/3} \bar{L}^{1/3}}{1 - 2^{-2/3}}.$$

Cauchy–Schwarz also gives

$$\sum_{k=0}^{K} L_k \sqrt{H_k(1+\Gamma_k)} \le \bar{L} \sqrt{\left(\sum_{k=0}^{K} H_k\right)\left(\sum_{k=0}^{K}(1+\Gamma_k)\right)} \le \bar{L}\sqrt{2T\Gamma_{\mathrm{ep}}(T)}.$$

The sum of the epoch-wise minima is at most the minimum of these two summed bounds. Taking expectations and adding the failure contribution proves (2).

Step 4: Recover the stated marginal rates. For stationary marginal means, each observed epoch has a constant score, so $\Gamma_k \le d_{\max}(H_k) + 1$. Consequently $\Gamma_{\mathrm{ep}}(T) = O(\log^2(eT))$, giving $\widetilde{O}(\sqrt{T})$ expected error. If the number of constant-mean segments is bounded by a deterministic J, let J_k count their nonempty intersections with E_k. The original segment boundaries and the K epoch boundaries form at most $J + K$ pieces, so $\sum_k J_k \le J + K$ and $\Gamma_k \le J_k(d_{\max}(H_k) + 1)$. Thus $\Gamma_{\mathrm{ep}}(T) = O((J + \log(eT))\log(eT))$. Substitution into (2), using $1 \le J \le T$, gives $\widetilde{O}(\sqrt{JT})$. The worst-case branch always gives $\widetilde{O}(T^{2/3})$. This proves the stated expectation guarantees. □

I Tightness of the Threshold-Complexity Dependence

We now show that the exact threshold-complexity corollary is tight, up to polylogarithmic factors, along the whole curve $\min\{\sqrt{T\Gamma_T^{\mathrm{thr}}}, T^{2/3}\}$. The construction is a parameterized version of the Walsh lower-bound instance of Collina et al. [2026b, Section 4]. In their instance, a grid discretization parameter m controls the threshold complexity: on an ordered grid of size m, every group family has $\Gamma_T^{\mathrm{thr}} \ge \widetilde{\Omega}(m)$, while the Walsh family has $\Gamma_T^{\mathrm{thr}} = \widetilde{\Theta}(m)$. The lower-bound proof of Collina et al. [2026b] then gives multicalibration error $\widetilde{\Omega}(\sqrt{mT})$ for every $m \le T^{1/3}$.

I.1 The Ordered-Grid Problem Instance.

To establish the lower bound, we construct an instance where the contexts are restricted to a uniformly spaced one-dimensional grid. Fix a power of two m satisfying $2 \le m \le T^{1/3}$. The contexts $x_1, \dots, x_m$ are uniformly spaced in the interval $[1/4, 3/4]$:

$$x_i := \frac{1}{4} + \frac{i-1}{2(m-1)}, \quad i \in \{1, \dots, m\}.$$

The learner faces these contexts in a periodic, round-robin fashion, i.e., $x_t := x_{1+((t-1) \pmod m)}$. The outcomes y_t are generated as:

$$y_t = x_t + \frac{\xi_t}{4}, \quad \xi_t \in \{\pm 1\} \text{ (independent Rademacher noise)}.$$

The signs ξ_t are also independent of the forecaster's internal randomness. The predictable mean is $\mu_t = x_t$, and we write $f^\star(x) = x$ for the identity score.

I.2 Fundamental Limits of Threshold Complexity

Lemma 11 (Generic fine-scale threshold lower bound)**.** *For all sufficiently large T, the ordered-grid instance above and any group family $\mathcal{G}$ satisfy*

$$\Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}) \ge c'\, m \log T = \widetilde{\Omega}(m),$$

for some constant $c' > 0$.

Lemma 12 (Threshold complexity of the Walsh family)**.** *There exists a binary group family $\mathcal{G}_{T,m}$ of size $O(\log^3(m+1))$ such that for the ordered-grid instance, the threshold complexity satisfies:*

$$\Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) = \widetilde{\Theta}(m).$$

I.3 Tightness of Multicalibration Error

The next theorem is a parameterized restatement of the Walsh lower bound of Collina et al. [2026b]. We do not reprove it from scratch; we only track the dependencies in its proof on the grid size m, whereas Collina et al. [2026b] state their theorem with m optimized for the worst-case bound.

Theorem 2 (Parameterized Walsh-family lower bound)**.** *There exist universal constants $c, C > 0$ such that, for all sufficiently large T and every power of two m with $2 \le m \le T^{1/3}$, every nonanticipating (possibly randomized) online forecaster on the ordered-grid instance above and the Walsh family $\mathcal{G}_{T,m}$ satisfies*

$$\mathbb{E}[\mathrm{MCerr}_{\mathcal{G}_{T,m}}(T)] \ge c\, \frac{\sqrt{mT}}{\log^C(T+1)}.$$

Moreover, by Lemma 12, this same instance has

$$\Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) = \widetilde{\Theta}(m),$$

so the lower bound can equivalently be written as

$$\mathbb{E}[\mathrm{MCerr}_{\mathcal{G}_{T,m}}(T)] \ge \widetilde{\Omega}\left(\sqrt{T\, \Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m})}\right).$$

Consequently, up to powers of two and logarithmic factors, the bound in Corollary 1 is tight throughout the regime $\Gamma_T^{\mathrm{thr}} \le T^{1/3}$. Taking $m \asymp T^{1/3}$ gives the worst-case lower bound $\widetilde{\Omega}(T^{2/3})$, which is the cap in Theorem 1. Equivalently, for every threshold-complexity budget $\Gamma_\star \ge 2$, there are instances with $\Gamma_T^{\mathrm{thr}} \le \widetilde{O}(\Gamma_\star)$ and expected error at least

$$\widetilde{\Omega}\left(\min\{\sqrt{T\Gamma_\star}, T^{2/3}\}\right).$$

Proofs of Lemmas 11, 12, and Theorem 2 appear in Appendix I.4. Conversely, plugging Lemma 12 into Corollary 1, and converting the high-probability bound to expectation using the failure-event argument from Theorem 1, gives the matching upper bound

$$\mathbb{E}[\mathrm{MCerr}_{\mathcal{G}_{T,m}}(T)] \le \widetilde{O}\left(\sqrt{mT}\right).$$

Therefore the upper and lower bounds match, up to polylogarithmic factors, along the full curve obtained by varying m up to the critical scale $T^{1/3}$, and the largest such m matches the worst-case bound.

I.4 Proofs for the Parameterized Tightness Bound

I.4.1 Proof of Lemma 11

Proof of Lemma 11. The proof demonstrates that for a sufficiently fine grid, every exact threshold-simple block must incur a complexity proportional to the number of distinct grid points it contains.

Step 1: Distinct grid points take distinct scores. Fix a depth d and write $w_d = 2^{-d}$. Let $c \in (0, \frac{1}{4})$ be a small absolute constant. We focus on fine-scale depths where the grid spacing dominates the discretization error:

$$w_d \le \frac{c}{m} \implies |x_{i+1} - x_i| = \frac{1}{2(m-1)} > 2w_d.$$

Consider any $(K_{S,d}, B_{S,d}, d)$-threshold simple block S and let $X(S)$ be the set of grid indices appearing in S. We claim that the score function $f_{S,d}$ must assign distinct values to every distinct grid point x_i, x_j for $i, j \in X(S)$.

Suppose for contradiction that $f_{S,d}(x_i) = f_{S,d}(x_j)$ for $i \neq j$. By the definition of threshold simplicity, the mean μ (which in this instance equals x) must satisfy $|\mu - f_{S,d}| \le w_d$. Thus,

$$|x_i - x_j| \le |x_i - f_{S,d}(x_i)| + |f_{S,d}(x_j) - x_j| \le 2w_d.$$

This contradicts our choice of w_d, which ensures $|x_i - x_j| > 2w_d$. Therefore, the score function must take at least $|X(S)|$ distinct values, implying $K_{S,d} \ge |X(S)|$.

Step 2: Lower bound threshold complexity at fine-scale depths. Since $B_{S,d} \ge 3/4$ for any non-empty block, the cost of any admissible partition $\mathcal{P}_d$ at this depth is:

$$\sum_{S \in \mathcal{P}_d} B_{S,d} K_{S,d} \ge \frac{3}{4} \sum_{S \in \mathcal{P}_d} K_{S,d} \ge \frac{3}{4} \sum_{S \in \mathcal{P}_d} |X(S)|.$$

Because the context sequence is round-robin and $m \le T^{1/3} \le T$, every grid point x_i appears at least once during the first T rounds. Since the blocks in $\mathcal{P}_d$ partition $[T]$, we have $\sum_{S \in \mathcal{P}_d} |X(S)| \ge m$. This yields the lower bound for a single fine-scale depth:

$$M_d^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}) = \inf_{\mathcal{P}_d} \sum_{S \in \mathcal{P}_d} B_{S,d} K_{S,d} \ge \frac{3m}{4}.$$

Step 3: Summation over depths. Take $c = 1/8$. Since $\log_2 m \le \frac{1}{3} \log_2 T$, the number of depths $\lceil \log_2(8m) \rceil \le d \le d_{\max}$ is at least $\frac{1}{6} \log_2 T - 3$, which is at least $\frac{1}{12} \log_2 T$ for $T \ge 2^{36}$. This threshold on T is uniform over the allowed values of m. Summing the above lower bound over those depths gives

$$\Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}) = \sum_{d=0}^{d_{\max}} M_d^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}) \ge c'\, m \log T = \widetilde{\Omega}(m).$$

□

I.4.2 Proof of Lemma 12

Proof of Lemma 12. We construct $\mathcal{G}_{T,m}$ using the subsampled Walsh system and bound its complexity in three steps.

Step 1: Construction of $\mathcal{G}_{T,m}$. We define a grid index map $\mathrm{idx} : [0,1] \to [m]$ such that $\mathrm{idx}(x_i) = i$ for each grid point and $\mathrm{idx}(x) = 1$ elsewhere. Let $\psi_\ell^{\mathrm{Wal}} : \{0, \dots, m-1\} \to \{\pm 1\}$ be the standard Walsh system. The signed Walsh features are then defined as:

$$w_\ell(x) := \psi_\ell^{\mathrm{Wal}}(\mathrm{idx}(x) - 1) \in \{\pm 1\}.$$

Each feature w_ℓ induces two binary groups $g_\ell^{\mathrm{Wal},+} := \frac{1+w_\ell}{2}$ and $g_\ell^{\mathrm{Wal},-} := \frac{1-w_\ell}{2}$, allowing us to write the feature as a difference $w_\ell = g_\ell^{\mathrm{Wal},+} - g_\ell^{\mathrm{Wal},-}$.

The subsampled Walsh family is defined as $\mathcal{G}_{T,m} := \{\mathbf{1}\} \cup \{g_\ell^{\mathrm{Wal},+}, g_\ell^{\mathrm{Wal},-} : \ell \in S\}$. Choose the fixed subset $S \subseteq \{1, \dots, m-1\}$ from Collina et al. [2026b, Lemma 11] with approximation error $\varepsilon = 1/4$. It has size $|S| = O(\log^3(m+1))$ and is independent of the forecaster.

Step 2: Logarithmic representation of discrete thresholds. In this proof, we use the convention $\mathrm{sgn}(u) = 1$ for $u \geq 0$ and $\mathrm{sgn}(u) = -1$ for $u < 0$. The key property of $\mathcal{G}_{T,m}$ is its sparse representation of thresholds. For any threshold r, let $k(r) := \big|\{i \in [m] : x_i < r\}\big|$ be the number of grid points below r. The target sign pattern on the grid, $i \mapsto \mathrm{sgn}(f^\star(x_i) - r)$, represents a simple step function: it is -1 for $i \leq k(r)$ and $+1$ for $i > k(r)$.

By the Walsh expansion and subsampling lemmas of Collina et al. [2026b, Section 4], there exists a signed linear combination of the Walsh features:

$$h_r(x) = \alpha_0(r)\mathbf{1}(x) + \sum_{\ell \in S} c_\ell(r) w_\ell(x),$$

which satisfies the following two properties.

First, for every grid point x_i, the approximation is within the margin required for stable decision-making:

$$\begin{cases} h_r(x_i) \geq \frac{3}{4} & f^\star(x_i) \geq r, \\ h_r(x_i) \leq -\frac{3}{4} & f^\star(x_i) < r. \end{cases}$$

Second, the cost of this representation, measured by the sum of absolute coefficients, is logarithmic:

$$\max_r |\alpha_0(r)| \leq 1 \quad \text{and} \quad \sum_{\ell \in S} \max_r |c_\ell(r)| \leq C \log m,$$

for some constant $C > 0$. This implies that every threshold of $f^\star$ has representation cost $B \leq 1 + 2\sum_{\ell \in S} \max_r |c_\ell(r)| \leq C \log m$ under the family $\mathcal{G}_{T,m}$.

Step 3: Threshold complexity on the ordered grid. We next upper bound the threshold complexity of this Walsh instance. Fix a depth d and write $\alpha_d := 2w_d$. We define the score function f_d via a monotone quantizer:

$$f_d(x) := \min\left\{1,\ \alpha_d\left(\left\lfloor \frac{x}{\alpha_d} \right\rfloor + \frac{1}{2}\right)\right\}.$$

This map is nondecreasing and satisfies $|f_d(x) - x| \leq \alpha_d/2 = w_d$ for every $x \in [0,1]$, including the capped endpoint $x = 1$. Thus the single-block score satisfies $\mathsf{R}_{[T],d}(f_d) = 0$. The threshold complexity M_d^{thr} is then bounded by two factors:

- Score count: the number of values taken by f_d on the realized contexts is at most $K_d \leq \min\left\{\left\lceil \frac{1}{\alpha_d} \right\rceil + 1,\ m\right\} \leq \min\left\{2^d + 1,\ m\right\} \leq C \min\{2^d, m\}$ values for some constant C.
- Representation cost: because f_d is monotone in the ordered grid index i, for every threshold r the set $\{i \in [m] : f_d(x_i) \geq r\}$ is a suffix of the ordered grid and its complement is a prefix. Thus, every threshold of f_d is again a discrete prefix/suffix sign pattern of exactly the type handled above, so the same Walsh representation gives cost $B_d \leq C \log m$.

Taking the single-block partition $\mathcal{P}_d = \{[T]\}$ in the definition of M_d^{thr} yields

$$M_d^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) \leq C \log m \cdot \min\{2^d, m\}.$$

Aggregating over all depths $d = 0, \ldots, d_{\max}$:

$$\begin{aligned} \Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) &= \sum_{d=0}^{d_{\max}} M_d^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) \\ &\leq C \log m \sum_{d=0}^{d_{\max}} \min\{2^d, m\} \\ &\leq Cm \log m \log T \\ &= \widetilde{O}(m). \end{aligned}$$

Applying Lemma 11 with $\mathcal{G} = \mathcal{G}_{T,m}$ gives

$$\Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) = \widetilde{\Theta}(m).$$

□

I.4.3 Proof of Theorem 2

Proof of Theorem 2. The proof is the parameterized algebra underlying the Walsh lower bound of Collina et al. [2026b, Section 4]. We recall the imported ingredients, translated into our notation. Let

$$M := \mathbb{E}[\mathrm{MCerr}_{\mathcal{G}_{T,m}}(T)], \quad A := \sum_{t=1}^{T} |p_t - x_t|,$$
$$\mathsf{Pred}_T := \{p_1, \dots, p_T\}, \qquad N := \sum_{v \in \mathsf{Pred}_T} \sqrt{|\{t : p_t = v\}|}.$$

For the Walsh family on a grid of size m, the ℓ_1-truthfulness lemma of Collina et al. [2026b, Lemma 12], using their subsampled-Walsh approximation lemma [Collina et al., 2026b, Lemma 11], gives the following bound for the same fixed family. Its approximation error $1/4$ is smaller than the $1/2$ required there, and its coefficient bound is uniform over cuts:

$$\mathbb{E}[A] \le C_1 \log(m+1)\, M.$$

Their diverse-predictions lemma [Collina et al., 2026b, Lemma 13] gives

$$\mathbb{E}[N] \ge \frac{T}{4\sqrt{\mathbb{E}[A] + T/m + 1}}.$$

Finally, their adaptive-noise-bucketing theorem and its constant-group corollary, combined with the bias–noise decomposition in Collina et al. [2026b, Section 4.7], yield, for $L_0 := \log(T+1)$,

$$M \ge \frac{c_2}{L_0}\, \mathbb{E}[N] - \mathbb{E}[A].$$

For a randomized forecaster, include its initial random seed in the filtration. Its predictions are then predictable for the independent Rademacher labels, as required by the noise theorem. These imported inequalities apply for every power-of-two $m \le T^{1/3}$; the statement in Collina et al. [2026b] specializes them to the optimized choice $m \asymp T^{1/3}$.

Since $\log(m+1) \le L_0$, the first and third inequalities imply

$$M \ge \frac{c_3}{L_0^2}\, \mathbb{E}[N]$$

for a universal constant $c_3 > 0$. Combining this with the diversity lower bound and the truthfulness bound yields

$$M \ge \frac{c_3 T}{4L_0^2 \sqrt{C_1 L_0\, M + T/m + 1}}.$$

If $C_1 L_0\, M \le T/m + 1$, then, since $m \le T$,

$$M \ge c_4\, \frac{T}{L_0^2 \sqrt{T/m + 1}} \ge c_5\, \frac{\sqrt{mT}}{L_0^2}.$$

If instead $C_1 L_0\, M > T/m + 1$, then

$$M \ge \frac{c_6 T}{L_0^{5/2} \sqrt{M}},$$

and hence $M \ge c_7 T^{2/3}/L_0^{5/3}$. Because $m \le T^{1/3}$, this is at least $c_7 \sqrt{mT}/L_0^{5/3}$. Thus in all cases

$$M \ge c\, \frac{\sqrt{mT}}{\log^C(T+1)}$$

for universal constants $c, C > 0$.

The equivalence with $\sqrt{T\Gamma_T^{\mathrm{thr}}}$ follows from Lemma 12. Given any target complexity budget $\Gamma_\star \ge 2$, choose m to be the largest power of two such that $m \le \min\{\Gamma_\star, T^{1/3}\}$. Then

$$\Gamma_T^{\mathrm{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) \le \widetilde{O}(\Gamma_\star),$$

and the lower bound above gives

$$M \ge \widetilde{\Omega}\Big(\min\{\sqrt{T\Gamma_\star}, T^{2/3}\}\Big).$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the adaptive multicalibration theorem, its piecewise-constant and near-constant mean specializations, and the parameterized tightness statement. These match Sections 4, 5, and I.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A short limitations discussion appears in the unnumbered "Limitations and Broader Impacts" section near the end of the paper, including both modeling and computational limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions accompany the formal statements. Section 5 outlines the proof. Formal proofs appear in the appendices, including Appendices G, H, and I for the main theorem and its three corollaries, restarting construction, and parameterized lower bound.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not contain empirical experiments; its contributions are theoretical.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not report experiments or release empirical assets. Reproducibility is addressed through full theorem statements and proofs rather than experimental code.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not contain training-based experiments.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [N/A]

Justification: The paper does not contain experiments, so there are no experimental compute resources to report.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is a theoretical study and does not involve human subjects, private data, or deceptive evaluation practices. We are not aware of any aspect of the work that conflicts with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The "Limitations and Broader Impacts" section discusses both the reliability motivations for calibration and the fact that calibration alone does not ensure fairness or good downstream decisions.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data, pretrained models, or other high-risk assets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external code, datasets, or models as assets; it cites prior theoretical work in the standard way.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, code, or models as submission assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subject research.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM tools assisted with writing, editing, and proof derivations.