

Instance-Adaptive Online Multicalibration

Zhiming Huang¹, Jamie Morgenstern^{1,3}, Aaron Roth², and Claire Jie Zhang¹

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

²Department of Computer and Information Sciences, University of Pennsylvania

³Amazon

May 22, 2026

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 $\tilde{O}(T^{2/3})$ worst-case-optimal rate for online multicalibration, while automatically adapting to easier instances: in the marginal stochastic setting it obtains a rate of $\tilde{O}(\sqrt{T})$, and for piecewise-stationary means with J segments its rate is $\tilde{O}(\sqrt{JT})$. 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 [1998], 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. The rate at which algorithms can achieve calibration 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 empirical 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, 2001, Sandroni et al., 2003]. The modern formulation of this idea—multicalibration [Hébert-Johnson et al., 2018]—is achievable 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*. There is a single, efficient algorithm that obtains (multi)calibration rates adaptive to the instance. It obtains $\tilde{O}(\sqrt{T})$ calibration rates on stochastic instances without groups (which is impossible for worst case adversaries), recovers the minimax-optimal $\tilde{O}(T^{2/3})$ 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, without groups, if the sequence of outcomes is produced by a piecewise-stationary environment in which the outcome y_t is drawn from a distribution with mean q_t which changes at most $J - 1$ times across the T rounds, we show that our algorithm obtains calibration error scaling as $\tilde{O}(\sqrt{JT})$ — recovering the claimed stochastic bound as the special case in which $J = 1$. We also obtain nontrivial rates when the means are not exactly stationary on any long segment: if $C_{\text{stat}}(q) = \inf_{c \in [0,1]} \sum_t |q_t - c|$, then the same algorithm obtains $\tilde{O}(\sqrt{T} + (TC_{\text{stat}}(q))^{1/3})$.

These marginal calibration guarantees are corollaries of our more general theorem for multicalibration over groups. The relevant difficulty of a multicalibration 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 to us when it takes only a few values and its thresholds can be represented cheaply using the group functions. The final 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

There is a large literature on online multicalibration and closely related guarantees. The modern formulation of multicalibration is due to Hébert-Johnson et al. [2018], but the first online multicalibration algorithms predate this formulation [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 rates scaling logarithmically with the number of groups, as well as generalizations to related guarantees (roughly speaking multicalibration for quantiles and moments). Lee et al. [2022] formulated online minimax multiobjective optimization and applied it to a variety of problems including multicalibration-style objectives, following the game theoretic tradition of deriving online calibration algorithms [Hart, 2025, Fudenberg and Levine, 1999, Foster, 1999, Foster and Hart, 2018]. Haghtalab et al. [2023] give an alternative unifying treatment of multicalibration algorithms through the lens of solving zero sum games with online learning algorithms. Farina and Perdomo [2026] interpret the kind of minimax step used in these papers as solving an expected variational inequality, and bring fast solvers to bear on it. Garg et al. [2024], Ghuge et al. [2025], Hu et al. [2025], Farina and Perdomo [2026] give oracle-efficient online multicalibration guarantees for rich benchmark classes. Their

emphasis is oracle efficiency for large/continuous classes, whereas our algorithm is efficient only for explicitly represented finite group families and gives instance-adaptive rates. More recently, Hu et al. [2025] studied efficient swap multicalibration for elicitable properties, giving online multicalibration algorithms for properties beyond means (cf. Noarov and Roth [2023]). The minimax sample complexity of online multicalibration was recently settled by Collina et al. [2026a,b] in both the batch and online settings, showing that the worst-case upper bound given by Noarov et al. [2025] was tight.

Bandit learning is another setting where there is a gap between stochastic and adversarial worst-case rates. There is a similarly motivated line of work that gives single algorithms obtaining the “best of both worlds” guarantees — i.e. optimal stochastic rates on stochastic instances without giving up on worst case rates on adversarial instances [Bubeck and Slivkins, 2012, De Rooij et al., 2014]. Similarly a branch of the online convex optimization literature studies more fine-grained regret bounds that adapt to properties of the realized loss functions, that can give better-than-worst-case $o(\sqrt{T})$ regret bounds on well behaved instances (see e.g. [Chiang et al., 2012, Chen et al., 2021]).

The idea of parameterizing the complexity of an adversary by how many times it changes strategies also appears in the literature on “tracking the best expert” [Herbster and Warmuth, 1998, Bousquet and Warmuth, 2002] and adaptive regret [Hazan and Seshadhri, 2007], which give regret bounds not just to the best expert in hindsight, but to the best sequence of experts that change a small number of times, or that have a small smoother measure of drift.

Our adaptive refinement of prediction values is reminiscent of “zooming algorithms” for the Lipschitz bandits problem in which algorithms iteratively refine attention to more promising regions of action space [Kleinberg et al., 2008, Slivkins, 2011, Podimata and Slivkins, 2021, Krishnamurthy et al., 2020]. From this literature, Podimata and Slivkins [2021], Krishnamurthy et al. [2020] are the most closely related in that they give instance-dependent bounds.

Recently, Hu et al. [2026] gave a swap regret algorithm for Lipschitz convex losses (where the action space is continuous) by using multiple scales of action discretizations. This is conceptually related to our approach; both approaches are designed to circumvent bottlenecks that result from algorithms tied to fixed discretizations of continuous action spaces. Despite the conceptual similarity, the techniques and results differ significantly; their algorithm simultaneously uses multiple discretization scales, and they prove worst-case regret bounds for different swap-regret objectives. Our algorithm adaptively refines predictions in a tree, and we prove beyond-worst-case-analysis style calibration bounds that take advantage of trajectory-specific structure.

Concurrently, Liu et al. [2026] give an alternative algorithm for obtaining instance-adaptive marginal calibration bounds. The first versions of our papers gave incomparable results in the marginal calibration setting. In the first version of our paper, our most interpretable corollary in the marginal calibration setting was the bound we state as Corollary 2 in the current version, which obtains calibration at the rate of $\min(T^{2/3}, \sqrt{JT})$ for stochastic instances whose means change J times, whereas the main result of Liu et al. [2026] was a bound of $\sqrt{T} + (TC_{\text{stat}}(q))^{1/3}$, where $C_{\text{stat}}(q)$ is a measure of the average distance of the label mean from its median across time. Neither bound implies the other. However after seeing each other’s papers, we realized that both methods could be used to give both bounds — we now also state the $\sqrt{T} + (TC_{\text{stat}}(q))^{1/3}$ bound as corollary 3, matching a theorem first given by Liu et al. [2026]. Beyond the overlapping results in the marginal calibration setting, the two papers generalize in different directions. The marginal calibration results we state are corollaries of the more general adaptive multicalibration bound we give for our algorithm, but we restrict our analysis to the ECE metric. Liu et al. [2026] give an algorithm only for marginal calibration, but give results for a broader collection of calibration 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.

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 $\mu_t = y_t$.

Let $\mathcal{G}$ be a finite family of binary groups $g : \mathcal{X} \rightarrow \{0, 1\}$. If the all-ones group does not belong to the class we add it and write $\bar{\mathcal{G}} := \mathcal{G} \cup \{\mathbf{1}\}$.

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

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

The marginal calibration error is:

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

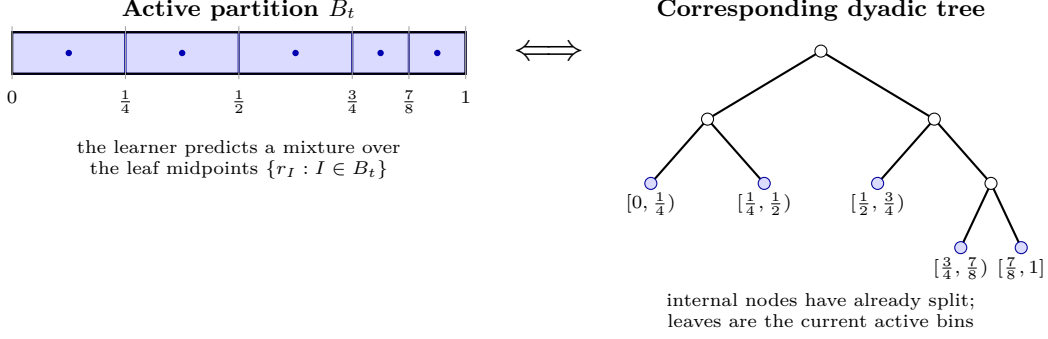


Figure 1: A typical state of the dynamic-bin data structure. The learner predicts using the midpoints of the current leaves, and an interval is refined only after the total mass assigned to it reaches the threshold L/w_I^2 .

2.2 Dynamic bins

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

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

Use $\mathcal{D}$ to denote the union of intervals¹ 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 $I \in B_t$. Let $\beta(I) := \min\{t \in [T] : I \in B_t\}$, $\delta(I) := \max\{t \in [T] : I \in B_t\}$. We refer to $\mathcal{V}_T := \bigcup_{t=1}^T B_t$ as the set of intervals that are ever active. We denote by $\mathcal{V}_d$ the set of depth- d intervals that are ever active, $\mathcal{V}_d = \mathcal{V}_T \cap \mathcal{D}_d$, and m_d is the cardinality of $\mathcal{V}_d$. 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 $I \notin B_t$. 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 $A(I) := [\beta(I), \delta(I)]$, and $\pi_{t,I} = 0$ for all $t \notin A(I)$.

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.

2.3 The general algorithm

For every group $g \in \bar{\mathcal{G}}$, interval I , and sign $\sigma \in \{+, -\}$, define the group-weighted signed bias terms

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

¹We 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 interval I has many experts corresponding to it, indexed by a start time s , group g , and sign σ . A particular expert (s, g, I, σ) is awake in round t if and only if $s \leq t$ (the expert's start time s has passed) and $I \in B_t$ (the corresponding interval is active). 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) :

$$\lambda_{t,g,I}^+ := \sum_{s \leq t} \omega_{t,(s,g,I,+)}, \quad \lambda_{t,g,I}^- := \sum_{s \leq t} \omega_{t,(s,g,I,-)}.$$

We then track interval-level aggregations of these weights

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

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, $|a_{t,I}| \leq b_{t,I}$.

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

$$\hat{\phi}_t := \sum_{I \in B_t} \sum_{g \in \bar{\mathcal{G}}} \left(\lambda_{t,g,I}^+ \phi_{g,I}^+(\pi_t, y_t) + \lambda_{t,g,I}^- \phi_{g,I}^-(\pi_t, y_t) \right),$$

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

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

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

Lemma 1 (Weighted feasibility). *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*

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

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 Appendix B.1). Because the expression $\sum_{I \in B_t} \pi_{t,I} (a_{t,I}(y - r_I) - b_{t,I}w_I)$ 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 , provides the normalized losses derived from $\phi_{g,I}^\pm(\pi_t, y_t)$ to the sleeping-experts wrapper, and splits any interval satisfying the threshold L/w_I^2 . The complete round-by-round procedure is summarized in Algorithm 1.

Algorithm 1: Dynamic bins for online multicalibration

Input: horizon T , group family $\mathcal{G}$
Initialize $B_1 = \{[0, 1]\}$ and $N(I) = 0$ for the root interval;
Instantiate a sleeping-experts wrapper using confidence-rated AdaNormalHedge on the tuples $(s, g, (I, \pm))$;
for $t = 1, \dots, T$ **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 $\phi_{g,I}^\pm(\pi_t, y_t)$;
 Update $N(I) \leftarrow N(I) + \pi_{t,I}$ 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 initialized with counter value 0;
end

3 Why the Algorithm Works: Marginal Calibration Intuition

Before defining the instance-complexity measures formally, 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 full proof in Section 6 follows the same template but in greater generality.

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 ,

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

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. This also explains the split rule. The algorithm refines I when $N(I) \asymp \log T/w_I^2$, the point where the discretization term $N(I)w_I$ and the regret term $\sqrt{N(I) \log T}$ 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 . 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

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

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 $R_{S,d}(c_S)$ is assigned to that interval.

Step 3: Residual error controls how many intervals split. A depth- d interval splits only after accumulating total probability mass on the order of $\log T/w_d^2$. 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 gives a bound on the number of splits at depth d on the order of:

$$1 + \sqrt{R_{S,d}(c_S)w_d/\log T}$$

for a single block S . If the block is exactly stationary at scale w_d , then the residual is zero and the block contributes only a constant number of splits 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

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

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.

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.

4 A Hierarchy of Complexity Measures

We now define the complexity measures appearing in our bounds, which track various ways in which the sequence of label means can be (un)predictable. In the marginal case, where the context space is a singleton, the predictable mean process reduces to a scalar sequence $q_t := \mu_t$. The simplest notion of simplicity is piecewise stationarity of the mean — that it changes only a small number of times:

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

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 5 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 $f : \mathcal{X} \rightarrow [0, 1]$ 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 $t \in S$,*

$$\begin{aligned} h_r(x_t) &\geq 1 - \eta && \text{whenever } f(x_t) \geq r, \\ h_r(x_t) &\leq -(1 - \eta) && \text{whenever } f(x_t) < r. \end{aligned}$$

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.

Remark 1. *With $\eta = 1/4$, every nonempty block satisfying the threshold-representable condition has $B \geq 3/4$. Indeed, choose $t_0 \in S$ and set $r = f(x_{t_0})$. Then $f(x_{t_0}) \geq r$, so the representation guarantee gives $h_r(x_{t_0}) \geq 3/4$. Since every group in $\bar{\mathcal{G}}$ is binary-valued,*

$$|h_r(x_{t_0})| \leq \sum_g |\alpha_g| \leq B.$$

The threshold-representable score 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 thresholds 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.

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

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

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

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

Then every threshold 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 $j \in \{2, \dots, K\}$ such that $a_{j-1} < r \leq a_j$. In this case,

$$f(x) \geq r \iff c(x) \geq j,$$

so we take

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

Thus, on the realized contexts,

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

Moreover, the sum of absolute coefficients 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 $\eta \in [0, 1]$.

Next, we introduce the generalization that allows the predictable mean sequence μ_t to vary from its score $f(x_t)$ (or, in the marginal case, to vary from a constant within a block) so long as those variations are small in aggregate. Approximation errors below the resolution w_d of a depth- d interval should be free, because the algorithm's analysis already pays for error at this scale. This motivates the following residual, which ignores such small errors.

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

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

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 $R_{S,d}(f_S) = 0$.

We now define the complexity measure that our main theorem depends on. For a partition of time into blocks S , the measure depends on the best threshold-representable score function f_S that can be defined on each block S , and scales with the parameters K_S and B_S of the score function along with its residuals $R_{S,d}(f_S)$. It then sums a complexity term depending on these parameters across all blocks S in the partition. Since the algorithm will obtain bounds scaling with the *best* such choice of partition and score functions, the complexity measure takes the infimum across all choices of partition, score functions, and valid values of the complexity parameters.

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

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

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 $+\infty$.

If we ask for perfect approximation within scale d (i.e. with zero truncated residual) then the measure simplifies:

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

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

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

$$R_{S,d}(f_S) = 0 \quad \text{for every } S \in \mathcal{P}_d.$$

The exact multiscale threshold complexity is

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

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

Table 1: Representative regimes covered by the dynamic-bin algorithm. The worst-case row recovers the known $\tilde{O}(T^{2/3})$ online multicalibration rate achieved by prior algorithms [Noarov et al., 2025, Ghuge et al., 2025]. In the last row, $C_{\text{stat}}(q) := \inf_{c \in [0,1]} \sum_t |q_t - c|$. The algorithm does not need to know which regime it is running in; all of these bounds are obtained by the same algorithm.

For a fixed partition and collection of score functions, it is useful to have notation for the relevant complexity terms:

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*

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

and

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

5 Main Theorem and Marginal Corollaries

We now state the main guarantee. The theorem has two forms. The first is a depth-by-depth bound in terms of the depth- d costs Ψ_d^{res} , which is our tightest bound and closest to the proof. The second is a cleaner consequence stated in terms of one contiguous partition and one threshold-representable score per segment. Table 1 summarizes representative consequences.

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

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

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

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

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 ,

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

Since $\text{MCerr}_{\mathcal{G}}(T) \leq T$ deterministically, the same bounds also hold in expectation up to an additive $O(1)$ term.

The main theorem gives a versatile bound, but can be confusing because of its generality. We now state a couple of more easily interpretable corollaries. For example, if the entire predictable mean process is exactly described by a single score function with low multiscale threshold complexity then we get the following corollary:

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

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

In particular, if for every depth d the whole horizon admits a (K_d, B_d, d) -threshold-simple score with $B_d \leq B$ and $K_d \leq K$, then

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

Similarly, suppose there is no group structure (so we are in the marginal calibration special case) and the sequence of label means is piecewise constant on J pieces. Then we get the following:

Corollary 2 (Piecewise-stationary marginal calibration). *In the context-free marginal calibration setting ($\mathcal{X}$ is a singleton and $\mathcal{G} = \{1\}$), 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*

$$\text{CalErr}(T) \leq \tilde{O} \left(\min \{ T^{2/3}, \sqrt{JT} \} \right).$$

Similarly, suppose the label mean may change at every round, but on average stays close to its median. Then we get the following corollary.

Corollary 3 (Slowly drifting marginal calibration). *In the context-free marginal calibration setting ($\mathcal{X}$ is a singleton and $\mathcal{G} = \{1\}$), let $C_{\text{stat}}(q) = \inf_{c \in [0,1]} \sum_{t=1}^T |q_t - c|$ be the best L_1 deviation from a constant mean. Then, with probability at least $1 - O(1/T)$,*

$$\text{CalErr}(T) \leq \tilde{O} \left(\sqrt{T} + (TC_{\text{stat}}(q))^{1/3} \right).$$

We note that it is possible to mix and match these bounds — for example, we can read off from the main theorem a bound for the marginal calibration case in which there are J blocks on each of which the label mean stays close to its median, but for which the median is very different on each of the J blocks. This emphasizes that the algorithm is the same for each of these corollaries — they are just different specializations of the main theorem — and so we always get the best of any of the stated bounds.

6 Proof of the Main Theorem

Proof Overview. The proof has six steps. First, we establish that the sleeping-experts wrapper controls one-sided weighted bias for every group, interval, and contiguous sub-block. Second, we show that if a depth- d interval midpoint is far from all values of a threshold-representable score on a block, then playing that interval grows signed bias at a linear rate; when μ_t is not exactly the score, the extra cost is the residual

approximation error on the rounds when that interval is played. Together these facts limit how frequently an interval far from a score value can be played. Third, we use this play bound to control how many depth- d intervals can be close to splitting during a block S . Formally, we bound

$$\Xi_d(S) := \sum_{I \in \mathcal{D}_d} \min \left\{ 1, \frac{N_S(I)w_d^2}{L} \right\},$$

where an interval contributes 1 if it receives enough mass on S to reach the depth- d split threshold, and contributes proportionally otherwise. Grouping intervals by distance from the score values gives

$$\Xi_d(S) \lesssim BK + \sqrt{BK \mathcal{R}_{S,d}(f)w_d/L}.$$

Fourth, we show these blockwise bounds control the number of intervals that ever become active at each depth. Fifth, calibration is accounted for by separately bounding the contribution of predictions made using intervals at each depth. Finally, a dyadic summation lemma converts the depth-wise bounds into the clean segmented approximation rate in Theorem 1.

To separate the probabilistic terms from the later split-count analysis, we define a global concentration event $\mathcal{E}_{\text{global}} := \mathcal{E}_{\text{samp}} \cap \mathcal{E}_{\text{conc}}$, and we further define $N_S^g(I) := \sum_{t \in S} g(x_t)\pi_{t,I}$. These events separate the stochasticity into two sources:

- **Sampling Error ($\mathcal{E}_{\text{samp}}$):** This event accounts for the variance introduced by the learner's randomized point predictions $p_t \sim \pi_t$. It ensures that for any group g and interval I , the realized weighted outcomes stay close to their expected values under the mixed forecast:

$$\mathcal{E}_{\text{samp}} : \max_{g \in \bar{\mathcal{G}}, I \in \mathcal{D}} \left| \sum_{t=1}^T (\mathbf{1}[p_t = r_I] - \pi_{t,I})g(x_t)(y_t - r_I) \right| \leq C_{\text{samp}} \left(\sqrt{N_T^g(I)L} + L \right),$$

where $N_T^g(I) = N_{[T]}^g(I)$.

- **Martingale Concentration ($\mathcal{E}_{\text{conc}}$):** This event bounds the outcome noise relative to the predictable means μ_t . It ensures that on any contiguous block S , the realized outcomes do not drift too far from their conditional means:

$$\mathcal{E}_{\text{conc}} : \max_{\substack{g \in \bar{\mathcal{G}}, I \in \mathcal{D} \\ S \subseteq [T] \text{ contiguous}}} \left| \sum_{t \in S} g(x_t)\pi_{t,I}(y_t - \mu_t) \right| \leq C_{\text{conc}} \left(\sqrt{N_S^g(I)L} + L \right)$$

By applying Freedman's inequality (see Appendix A.2), each event fails with probability at most $1/4T$, ensuring $\Pr(\mathcal{E}_{\text{global}}) \geq 1 - \frac{1}{2T}$. Our analysis will condition on this event.

6.1 From Bias Control via Regret to Depth-Wise Calibration

To bound the global multicalibration error, we first control cumulative bias within each activated interval. This follows from the regret-minimization properties of our wrapper algorithm.

Lemma 2 (Bias control via regret). *For every group $g \in \bar{\mathcal{G}}$, every dyadic interval I , and every contiguous block $S \subseteq A(I)$,*

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

Proof Sketch. We bound the cumulative one-sided weighted bias by relating it to the wrapper's regret against the sleeping expert associated with $\phi_{g,I}^+$ and start time $\min S$. Since the feasibility condition (1) ensures the aggregate one-step bias satisfies $\hat{\phi}_t \leq 0$, regret against this expert upper-bounds the cumulative positive bias

term $\sum_{t \in S} \phi_{g,I}^+(\pi_t, y_t)$. The first-order AdaNormalHedge guarantee scales with the expert's positive relative loss, and in this construction that quantity is at most a constant multiple of the total group-weighted play $N_S^g(I)$. Therefore the positive one-sided bias is at most $O(\sqrt{N_S^g(I)L} + L)$. Adding back the width slack in the definition of $\phi_{g,I}^+$ gives

$$\sum_{t \in S} g(x_t) \pi_{t,I} (y_t - r_I) \leq N_S^g(I) w_I + O(\sqrt{N_S^g(I)L} + L).$$

The negative direction is identical using the expert associated with $\phi_{g,I}^-$. See Appendix B.2.1 for details. $\square$

Lemma 3 (Depth-wise calibration accounting). *Let $\mathcal{V}_d := \mathcal{V}_T \cap \mathcal{D}_d$ be the set of depth- d intervals that are active in at least one prediction round, and let $m_d := |\mathcal{V}_d|$. Assume $L \leq T$. On $\mathcal{E}_{\text{samp}}$, for every group $g \in \mathcal{G}$ and every depth d ,*

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

Proof Sketch. Apply Lemma 2 to each interval $I \in \mathcal{V}_d$ on its active lifetime and combine it with the sampling event. The split rule implies $N(I)w_d^2 \leq CL$ for every interval that ever appears: a non-max-depth interval can overshoot the split threshold by at most one round of mass, while a max-depth interval has $N(I) \leq T$ and $Tw_d^2 = O(1)$. Hence the deterministic discretization term is absorbed by the square-root term. Summing $\sqrt{N_T^g(I)L}$ over $I \in \mathcal{V}_d$ gives the first bound by Cauchy-Schwarz and the second bound using $N_T^g(I) \leq N(I) \leq CL/w_d^2$. See Appendix B.2.2. $\square$

6.2 Controlling Splits From Approximation Structure

We now connect the approximation costs to the adaptive tree. The key point is that if an interval midpoint is far from all score values on a block, then the interval can be played often only when a significant amount of residual approximation error is assigned to it.

Lemma 4 (Intervals far from the score are rarely played). *Condition on $\mathcal{E}_{\text{conc}}$. Fix a depth d , write $w = w_d$, let S be a 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*

$$\delta_I := \text{dist}(r_I, C_f(S)), \quad D_{S,I} := \sum_{t \in S \cap A(I)} \pi_{t,I} (|\mu_t - f(x_t)| - w)_+.$$

There is a universal constant C_0 such that, whenever $\delta_I > C_0 Bw$,

$$N_S(I) \leq C \left[\frac{B^2 L}{(\delta_I - C_0 Bw)^2} + \frac{BD_{S,I}}{\delta_I - C_0 Bw} \right].$$

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

Proof Sketch. Let $h_{r_I} = \sum_g \alpha_g g$ be the threshold representation of f at threshold r_I . The margin condition gives $h_{r_I}(x_t)(f(x_t) - r_I) \geq (3/4)|f(x_t) - r_I|$, while $|h_{r_I}(x_t)| \leq B$. Therefore

$$h_{r_I}(x_t)(\mu_t - r_I) \geq (3/4)\delta_I - Bw - B(|\mu_t - f(x_t)| - w)_+.$$

After summing with weights $\pi_{t,I}$ over $S \cap A(I)$, this lower bounds the expected signed bias by a linear term in $N_S(I)$ minus $BD_{S,I}$. The algorithmic upper bound follows by expanding h_{r_I} in the group basis and applying Lemma 2 and $\mathcal{E}_{\text{conc}}$ to each group. Combining the two bounds and solving the resulting quadratic gives the claim. See Appendix C.1. $\square$

We next convert control of the total play on a single block into a bound on how much that block can contribute toward splits at depth d .

Lemma 5 (Split count on one block). *Condition on $\mathcal{E}_{\text{conc}}$. Fix depth d , write $w = w_d$, and let S be a contiguous block. If f is $(K, B, \frac{1}{4})$ -threshold representable on S , then*

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

Proof Sketch. We group depth- d intervals according to their distance from the nearest realized score value. Intervals within distance $O(Bw_d)$ are few, contributing at most $O(BK)$ to $\Xi_d(S)$. For farther intervals, Lemma 4 gives a bound with two terms: a threshold-cost term that decays quadratically in the distance, and a residual term depending on the residual mass assigned to that interval. Summing over distance ranges and optimizing the cutoff between nearby and far intervals gives the claimed bound. See Appendix C.2. $\square$

Let $\mathcal{A}_d$ denote the set of depth- d intervals that split by time T .

Lemma 6 (Active intervals from split bounds). *Condition on $\mathcal{E}_{\text{conc}}$ and let $m_d = |\mathcal{V}_d|$. Then*

$$m_0 = 1, \quad m_d \leq C \min \left\{ 2^d, \frac{Tw_d^2}{L}, \Psi_{d-1}^{\text{res}}(\mu_{1:T}; \mathcal{G}) \right\} \quad (d \geq 1).$$

Proof Sketch. Every active depth- d interval is generated by a depth- $(d-1)$ parent that split, so $m_d \leq 2|\mathcal{A}_{d-1}|$. The deterministic caps $m_d \leq 2^d$ and $m_d \leq CTw_d^2/L$ follow from the dyadic grid and the total play mass. For the instance-dependent bound, split mass is additive over any contiguous temporal partition, and Lemma 5 bounds the resulting blockwise split costs. Taking the infimum over partitions and score representations gives the bound in terms of Ψ_{d-1}^{res} . See Appendix C.3. $\square$

Lemma 7 (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 Sketch. The deterministic caps give the worst-case contribution $O(T^{2/3}L^{1/3})$. For the instance-dependent contribution, split $M + A\sqrt{w_d/L}$ into an M part and an A part. The M part sums to $O(\sqrt{TLM})$ by a dyadic crossover between Tw_d and LM/w_d . The A part sums to $O(T^{1/3}L^{1/3}A^{2/3})$ by the crossover $w_* = (A\sqrt{L}/T)^{2/3}$. See Appendix C.4 for the details. $\square$

6.3 Proof of the Main Theorem

Proof of Theorem 1. If $L > T$, then the trivial bound $\text{MCerr}_{\mathcal{G}}(T) \leq T$ 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 . Hence assume $L \leq T$ and condition on $\mathcal{E}_{\text{global}}$.

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 $g \in \mathcal{G}$,

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

Applying Lemma 3 at each depth and then maximizing over g gives

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

Lemma 6 gives $m_d \leq C\Theta_d$ for every depth d . Absorbing constants into C proves the depth-wise bound in the theorem. The deterministic parts of Θ_d also give the worst-case branch. Indeed, for $d \geq 1$ we use only

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

Thus the d th 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.

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 $R_{S,d-1}(f_S) \leq R_S(f_S)$,

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

where the last step uses $w_{d-1} = 2w_d$. Combining this with Lemma 6, the active interval counts satisfy the hypothesis of Lemma 7 with $M = M(\mathcal{P}, f)$ and A equal to an absolute-constant multiple of $A(\mathcal{P}, f)$. Applying that lemma gives the segmented bound. The worst-case cap is the first branch proved above. $\square$

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

$$\sqrt{TL} + \sum_{d=1}^{d_{\max}} \sqrt{TL M_{d-1}^{\text{thr}}} \leq \tilde{O} \left(\sqrt{TL (1 + \Gamma_T^{\text{thr}}(\mu_{1:T}; \mathcal{G}))} \right),$$

where the last step uses Cauchy-Schwarz over the $O(\log T)$ 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 $M_d^{\text{thr}} \leq B_d K_d \leq BK$. Hence $\Gamma_T^{\text{thr}} \leq O(BK \log T)$, and the logarithmic factor is hidden in $\tilde{O}(\cdot)$. $\square$

Proof of Corollary 2. Let the stationary segments be $S_1, \dots, S_J$, and write $q_t = q^{(j)}$ on S_j . Let $\mathcal{P} = \{S_1, \dots, S_J\}$. On each segment S_j , use the constant score $f_{S_j} \equiv q^{(j)}$. In the singleton-context setting, each such score is $(1, 1, \frac{1}{4})$ -threshold representable and has zero residual. Thus $M(\mathcal{P}, f) = 1 + J$ and $A(\mathcal{P}, f) = 0$. The segmented part of Theorem 1 gives

$$\text{CalErr}(T) \leq \tilde{O} \left(\min \left\{ T^{2/3}, \sqrt{JT} \right\} \right).$$

$\square$

Proof of Corollary 3. Let $c^* \in \arg \min_{c \in [0,1]} \sum_{t=1}^T |q_t - c|$. Use the one-block partition $\mathcal{P} = \{[T]\}$ and the constant score $f \equiv c^*$. In the singleton-context setting this score is $(1, 1, \frac{1}{4})$ -threshold representable. Therefore $M(\mathcal{P}, f) = O(1)$ and $A(\mathcal{P}, f) = \sqrt{C_{\text{stat}}(q)}$. The segmented part of Theorem 1 gives

$$\text{CalErr}(T) \leq \tilde{O}\left(\sqrt{T} + T^{1/3} C_{\text{stat}}(q)^{1/3}\right) = \tilde{O}\left(\sqrt{T} + (T C_{\text{stat}}(q))^{1/3}\right).$$

□

7 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 $\min\{\sqrt{T}\Gamma_T^{\text{thr}}, T^{2/3}\}$. 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 $\Gamma_T^{\text{thr}} \geq \tilde{\Omega}(m)$, while the Walsh family has $\Gamma_T^{\text{thr}} = \tilde{\Theta}(m)$. The lower-bound proof of Collina et al. [2026b] then gives multicalibration error $\tilde{\Omega}(\sqrt{mT})$ for every $m \leq T^{1/3}$.

7.1 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 $2 \leq m \leq T^{1/3}$. 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., $x_t := x_{1+((t-1) \bmod m)}$. The outcomes y_t are generated as:

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

Under this construction, the predictable mean is simply the context itself, $\mu_t = x_t$, and the optimal regression function in the class is the identity $f^*(x) = x$.

7.2 Fundamental Limits of Threshold Complexity

Lemma 8 (Generic fine-scale threshold lower bound). *For the ordered-grid instance above and any group family $\mathcal{G}$,*

$$\Gamma_T^{\text{thr}}(\mu_{1:T}; \mathcal{G}) \geq c' m \log T = \tilde{\Omega}(m),$$

for some constant $c' > 0$.

Lemma 9 (Threshold complexity of the Walsh family). *There exists a binary group family $\mathcal{G}_{T,m}$ of size $O(\log^3 m)$ such that for the ordered-grid instance, the threshold complexity satisfies:*

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

7.3 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 $2 \leq m \leq T^{1/3}$, every (possibly randomized) online forecaster on the ordered-grid instance above and the Walsh family $\mathcal{G}_{T,m}$ satisfies*

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

Moreover, by Lemma 9, this same instance has

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

so the lower bound can equivalently be written as

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

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

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

Proofs of Lemmas 8, 9, and Theorem 2 appear in Appendix D. Conversely, plugging Lemma 9 into Corollary 1, and converting the high-probability bound to expectation using the failure-event argument from Theorem 1, gives the matching upper bound

$$\mathbb{E}[\text{MCerr}_{\mathcal{G}_{T,m}}(T)] \leq \tilde{O}(\sqrt{mT}).$$

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. This is consistent with the residual bound in Theorem 1: the Walsh instances have exact low-residual scores, and the hardness lies in the threshold cost of representing their thresholds rather than in residual drift.

Acknowledgments

We gratefully acknowledge support from the Simons Collaboration on Algorithmic Fairness and the NSF EnCoRE Tripods institute.

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. 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. 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(1):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 *2006 IEEE Information Theory Workshop-ITW’06 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 and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.
- Dean P Foster, Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Complexity-based approach to calibration with checking rules. In *Proceedings of the 24th Annual Conference on Learning Theory*, 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 *Proceedings of the 2024 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 *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pages 82–1. 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. *Advances in Neural Information Processing Systems*, 36:72464–72506, 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. In *Electronic colloquium on computational complexity (ECCC)*, volume 14, 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 forecasters. 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. Annual 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: Multicalibating and other applications. *Advances in Neural Information Processing Systems*, 35:29051–29063, 2022.
- Ehud Lehrer. Any inspection is manipulable. *Econometrica*, 69(5):1333–1347, 2001.
- Junyan Liu, Haipeng Luo, and Lillian J Ratliff. Adaptive calibration in non-stationary environments. *arXiv preprint arXiv:2605.11490*, 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 *International Conference on Machine Learning*, 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*, 11:30–31, 2022.

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, page 133, 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 Concentration Inequality

A.1 Two-stage Freedman Inequality (Lemma 10)

Lemma 10 (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*

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

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 the standard Freedman inequality by constructing an interleaved filtration. Set $\tilde{\mathcal{F}}_0 := \mathcal{F}_0$, and define $(\tilde{\mathcal{F}}_k)_{k=1}^{2n}$ as follows:

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

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 $\tilde{\mathcal{F}}_k$ -measurable and satisfies:

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

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

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

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]. $\square$

A.2 The Global Concentration Event

A.2.1 Sampling Error (Lemma 11)

Lemma 11 (Sampling error). *There exists an event $\mathcal{E}_{\text{samp}}$ of probability at least $1 - \frac{1}{4T}$ such that, on $\mathcal{E}_{\text{samp}}$, for every $g \in \bar{\mathcal{G}}$ and every interval $I \in \mathcal{D}$,*

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

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

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

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 $|Z_t^g(I)| \leq 1$. The conditional variance is bounded as follows:

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

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

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

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

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

$$\mathcal{U}_m := \{2^{m-1} < N_T^g(I) \leq 2^m\},$$

with the convention $\mathcal{U}_0 = \{N_T^g(I) \leq 1\}$. On $\mathcal{U}_m$ we have $V_T^g(I) \leq 2^m$, so Lemma 10 gives, for every $x \geq 0$,

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

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 $\sqrt{2^{m+1}\Lambda} \leq \sqrt{2} \cdot 2^{m/2} \Lambda \leq 2\sqrt{(N_T^g(I) \vee 1)\Lambda}$. Summing over $m = 0, \dots, m_{\max}$, we obtain that with probability at least $1 - \delta$:

$$|M_T^g(I)| \leq 2\sqrt{(N_T^g(I) \vee 1)\Lambda} + \frac{2}{3}\Lambda.$$

Using the inequality $\sqrt{N_T^g(I) \vee 1} \leq \sqrt{N_T^g(I)} + 1$, and simplifying constants, we have:

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

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

$$|M_T^g(I)| \leq C \left(\sqrt{N_T^g(I)L} + L \right)$$

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

A.2.2 Grouped Martingale Concentration

Lemma 12 (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 $S \subseteq [T]$,*

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

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 11.

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

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

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:

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

The increments are bounded by $|X_t^g(I, S)| \leq 1$. The predictable quadratic variation of the sum $M_S^g(I) := \sum_{t=1}^T X_t^g(I, S)$ satisfies:

$$V_S^g(I) := \sum_{t=1}^T \mathbb{E}[(X_t^g(I, S))^2 \mid \mathcal{H}_{t-1}] \leq \sum_{t \in S} g(x_t) \pi_{t,I} = N_S^g(I),$$

where we used the facts that $(y_t - \mu_t)^2 \leq 1$, $g(x_t) \in \{0, 1\}$, 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 $m_{\text{max}} := \lceil \log_2 T \rceil$. For each $m = 0, 1, \dots, m_{\text{max}}$, 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 10), for any $x \geq 0$:

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

Step 3: Union bound over buckets and (g, I) pairs. Fix $\delta \in (0, 1)$ and let $\Lambda := \log(2(m_{\text{max}} + 1)/\delta)$. Setting the threshold $x_m := \sqrt{2^{m+1}\Lambda} + \frac{2}{3}\Lambda$ and summing over all $m_{\text{max}} + 1$ 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. $\square$

B Algorithmic Properties

B.1 Weighted Feasibility (Lemma 1)

Proof of Lemma 1. 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

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

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

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

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:

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

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

B.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 $|\mathcal{D}| = \sum_{d=0}^{d_{\max}} 2^d \leq 4\sqrt{T}$, the number of coordinates is bounded by $M \leq 8\sqrt{T}$.

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

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

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 $|\phi_{g,I}^\pm(\pi_t, y_t)|$. 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

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

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 $\hat{\ell}_t := \sum_{e: I_{t,e}=1} \omega_{t,e} \ell_{t,e}$. By our normalization, this corresponds to the predicted violation $\hat{\phi}_t = \sum_{I \in B_t} \pi_{t,I} (a_{t,I} (y_t - r_I) - b_{t,I} w_I)$ from the potential-based update:

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

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

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

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 $C_{\text{anh}} > 0$ such that, simultaneously for every expert e and every horizon $b \leq T$,*

$$\sum_{t=1}^b I_{t,e} (\hat{\ell}_t - \ell_{t,e}) \leq C_{\text{anh}} \left(\sqrt{\tilde{L}_{b,e} L} + L \right),$$

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

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

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

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

The confidence-rated AdaNormalHedge guarantee gives

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

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

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

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

$$R_{b,e} \leq \sqrt{2\tilde{L}_{b,e} A_{b,e}} + A_{b,e},$$

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

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

where $B_b \leq \frac{5}{2} + \frac{3}{2} \ln(1 + T) = O(\log T)$. 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 $\log B_b = O(\log \log T)$ and $\log(1 + \log N) = O(\log \log(T|\bar{\mathcal{G}}|))$, we obtain $A_{b,e} \leq C' L$ for a universal constant C' . Substituting into the regret inequality yields the claim. $\square$

B.2.1 Bias Control via Regret (Lemma 2)

Proof of Lemma 2. 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

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

Using the wrapper identity from Section 2,

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

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

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

Proposition 1 yields

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

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

$$\tilde{L}_{b,e} = \sum_{t \in S} \left[\frac{\hat{\phi}_t - \phi_{g,I}^+(\pi_t, y_t)}{2G} \right]_+ \leq \sum_{t \in S} \frac{(-\phi_{g,I}^+(\pi_t, y_t))_+}{2G} \leq \frac{1}{2} \sum_{t \in S} g(x_t) \pi_{t,I} = \frac{1}{2} N_S^g(I),$$

because $\hat{\phi}_t \leq 0$ and $|\phi_{g,I}^+(\pi_t, y_t)| \leq G g(x_t) \pi_{t,I}$. Rearranging gives

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

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 $\tilde{L}_{b,e}$. Thus the regret term scales with $N_S^g(I) = \sum_{t \in S} g(x_t) \pi_{t,I}$ rather than with $|S|$ or the active lifetime of I . Adding back the width term yields

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

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

B.2.2 Depth-Wise Calibration Accounting (Lemma 3)

Proof of Lemma 3. Fix a group g and a depth d , and let C be the absolute universal constant whose exact values may change from line to line. For each $I \in \mathcal{V}_d$, apply Lemma 2 on $A(I)$ and then use the sampling event $\mathcal{E}_{\text{samp}}$:

$$\left| \sum_{t:p_t=r_I} g(x_t)(y_t - r_I) \right| \leq C \left(N_T^g(I)w_d + \sqrt{N_T^g(I)L} + L \right).$$

By the splitting rule, every interval that ever appears satisfies $N(I)w_d^2 \leq CL$, including max-depth intervals. Since $N_T^g(I) \leq N(I)$, the deterministic term is absorbed:

$$N_T^g(I)w_d \leq C\sqrt{N_T^g(I)L}.$$

Thus

$$\left| \sum_{t:p_t=r_I} g(x_t)(y_t - r_I) \right| \leq C \left(\sqrt{N_T^g(I)L} + L \right).$$

Summing over $I \in \mathcal{V}_d$ and using Cauchy–Schwarz gives

$$\sum_{I \in \mathcal{V}_d} \sqrt{N_T^g(I)L} \leq \sqrt{Lm_d} \sum_{I \in \mathcal{V}_d} N_T^g(I) \leq \sqrt{TLm_d},$$

because at each round the total mass placed on depth- d active intervals is at most one. If $L > T$, the main theorem uses the trivial worst-case bound; the depth-wise bound is only invoked in the case $L \leq T$. When $L \leq T$, the deterministic active-count cap $m_d \leq CTw_d^2/L$ for $d \geq 1$ (see more details about this in Appendix C.3 proof of Lemma 3) and $m_0 = 1$ imply

$$Lm_d = \sqrt{TLm_d} \cdot \sqrt{\frac{Lm_d}{T}} \leq \sqrt{TLm_d} \cdot \sqrt{Cw_d^2} \leq C\sqrt{TLm_d},$$

so the additive term is absorbed into the first branch. For the second bound, use $N_T^g(I) \leq CL/w_d^2$ for every $I \in \mathcal{V}_d$:

$$\sum_{I \in \mathcal{V}_d} \sqrt{N_T^g(I)L} \leq \sum_{I \in \mathcal{V}_d} \frac{CL}{w_d} = \frac{CLm_d}{w_d}.$$

Together with $Lm_d \leq Lm_d/w_d$, this proves the second branch. $\square$

C Bounding Splits From Approximation Structure

C.1 Bins Far From Score Values Are Rarely Played (Lemma 4)

Proof of Lemma 4. Fix I and assume $\delta_I > C_0Bw$; otherwise only the trivial bound is claimed. Let $S_I := S \cap A(I)$. Since $\pi_{t,I} = 0$ outside $A(I)$, we have $N_S(I) = N_{S_I}(I)$. If $S_I = \emptyset$, there is nothing to prove.

Let $h_{r_I} = \sum_{g \in \mathcal{G}} \alpha_g g$ be the threshold representation of f at threshold r_I , so $\sum_g |\alpha_g| \leq B$. For every $t \in S$,

$$h_{r_I}(x_t)(f(x_t) - r_I) \geq \frac{3}{4}|f(x_t) - r_I| \geq \frac{3}{4}\delta_I.$$

Also $|h_{r_I}(x_t)| \leq B$ because the groups are binary-valued. Therefore, with

$$\rho_t := (|\mu_t - f(x_t)| - w)_+,$$

we have

$$\begin{aligned}
h_{r_I}(x_t)(\mu_t - r_I) &= h_{r_I}(x_t)(f(x_t) - r_I) + h_{r_I}(x_t)(\mu_t - f(x_t)) \\
&\geq \frac{3}{4}\delta_I - B|\mu_t - f(x_t)| \\
&\geq \frac{3}{4}\delta_I - Bw - B\rho_t.
\end{aligned}$$

Summing over S_I gives the lower bound

$$A_\mu := \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(\mu_t - r_I) \geq \left(\frac{3}{4}\delta_I - Bw\right) N_S(I) - BD_{S,I}. \quad (2)$$

For the upper bound,

$$A_\mu = \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - r_I) - \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - \mu_t). \quad (3)$$

Invoking Lemma 2 and then using $N_{S_I}^g(I) \leq N_S(I)$ gives

$$\begin{aligned}
\left| \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - r_I) \right| &\leq \sum_{g \in \mathcal{G}} |\alpha_g| \left| \sum_{t \in S_I} g(x_t) \pi_{t,I} (y_t - r_I) \right| \\
&\leq \sum_g |\alpha_g| N_{S_I}^g(I)w + C_{\text{loc}} \sum_g |\alpha_g| \sqrt{N_{S_I}^g(I)L} + BC_{\text{loc}}L \\
&\leq BN_S(I)w + C_{\text{loc}}B\sqrt{N_S(I)L} + C_{\text{loc}}BL,
\end{aligned} \quad (4)$$

where the square-root term uses $\sum_g |\alpha_g| \sqrt{N_{S_I}^g(I)} \leq B\sqrt{N_S(I)}$. Similarly, by $\mathcal{E}_{\text{conc}}$,

$$\begin{aligned}
\left| \sum_{t \in S_I} \pi_{t,I} h_{r_I}(x_t)(y_t - \mu_t) \right| &\leq \sum_{g \in \mathcal{G}} |\alpha_g| \left| \sum_{t \in S_I} g(x_t) \pi_{t,I} (y_t - \mu_t) \right| \\
&\leq \sum_g |\alpha_g| C_{\text{conc}} \left(\sqrt{N_{S_I}^g(I)L} + L \right) \\
&\leq C_{\text{conc}}B\sqrt{N_S(I)L} + C_{\text{conc}}BL.
\end{aligned} \quad (5)$$

Substituting (5) and (4) into (3) gives

$$|A_\mu| \leq BN_S(I)w + CB(\sqrt{N_S(I)L} + L),$$

where C is an absolute universal constant whose exact values may change from line to line.

Combining this upper bound with (2) and choosing C_0 large enough to absorb all $BwN_S(I)$ terms gives, for $\Delta_I := \delta_I - C_0Bw > 0$,

$$\Delta_I N_S(I) \leq CB \left(D_{S,I} + \sqrt{N_S(I)L} + L \right).$$

Let $z := \sqrt{N_S(I)}$. Then

$$\Delta_I z^2 \leq CBD_{S,I} + CB\sqrt{L}z + CBL.$$

Young's inequality gives $CB\sqrt{L}z \leq \frac{1}{2}\Delta_I z^2 + CB^2L/\Delta_I$, hence

$$\frac{1}{2}\Delta_I z^2 \leq CBD_{S,I} + C\frac{B^2L}{\Delta_I} + CBL,$$

and therefore

$$N_S(I) = z^2 \leq C \left[\frac{B^2L}{\Delta_I^2} + \frac{BD_{S,I}}{\Delta_I} + \frac{BL}{\Delta_I} \right].$$

Since $\Delta_I \leq 1$ and $B \geq 3/4$ for nonempty blocks, the last term is absorbed into CB^2L/Δ_I^2 . This proves the claim. $\square$

C.2 Split Count on One Block (Lemma 5)

Proof of Lemma 5. Let $C_f(S) = \{c_1, \dots, c_K\}$ be the realized score values on S , and write $R := R_{S,d}(f)$. Assign each depth- d interval I to a nearest score value $c_{k(I)}$, and set $\delta_I := |r_I - c_{k(I)}|$. Let C_0 be the constant from Lemma 4. Choose

$$Q := C_1 B + \sqrt{\frac{BRw}{KL}},$$

where C_1 is a sufficiently large universal constant, so that $Q \geq 2C_0 B$.

The near region consists of intervals with $\delta_I \leq Qw$. Since depth- d midpoints are spaced by w , each score value has $O(Q)$ nearby intervals. Hence the near-region contribution to $\Xi_d(S)$ is $O(KQ)$.

For the far region, define distance ranges

$$\mathcal{S}_{k,\ell} := \{I : k(I) = k, \ell w < |r_I - c_k| \leq (\ell + 1)w\}, \quad \ell \geq \lceil Q \rceil.$$

Each range contains $O(1)$ intervals. Since $Q \geq 2C_0 B$, every interval in these ranges has

$$\Delta_I := \delta_I - C_0 Bw \geq c\ell w$$

for a universal constant $c > 0$. Lemma 4 implies

$$\min \left\{ 1, \frac{N_S(I)w^2}{L} \right\} \leq C \left[\frac{B^2}{\ell^2} + \frac{BD_{S,I}w}{\ell L} \right].$$

Summing the first term over all far ranges gives

$$K \sum_{\ell \geq \lceil Q \rceil} O\left(\frac{B^2}{\ell^2}\right) \leq C \frac{KB^2}{Q} \leq CBK,$$

using $Q \geq C_1 B$. For the residual term, use $\ell \geq Q$ and

$$\sum_{I \in \mathcal{D}_d} D_{S,I} = \sum_{I \in \mathcal{D}_d} \sum_{t \in S \cap A(I)} \pi_{t,I} (|\mu_t - f(x_t)| - w)_+ \leq R,$$

because at a fixed round the total mass assigned to active depth- d intervals is at most one. Thus the far residual contribution is at most $CBRw/(QL)$.

Combining the near and far terms,

$$\Xi_d(S) \leq C \left[KQ + \frac{BRw}{QL} \right].$$

The chosen value of Q gives

$$KQ \leq CKB + C\sqrt{\frac{BKRw}{L}}, \quad \frac{BRw}{QL} \leq C\sqrt{\frac{BKRw}{L}}.$$

This proves the lemma. $\square$

C.3 Active Intervals From Split Bounds (Lemma 6)

Proof of Lemma 6. The root is the only active depth-0 interval, so $m_0 = 1$.

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

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

It remains to prove the instance-dependent bound. 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 Ψ_{d-1}^{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,

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

Applying Lemma 5 to each block at depth $d-1$ gives

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

Taking the infimum over $\mathcal{P}_{d-1}$ and the scores gives $|\mathcal{A}_{d-1}| \leq C\Psi_{d-1}^{\text{res}}(\mu_{1:T}; \mathcal{G})$, and hence $m_d \leq C\Psi_{d-1}^{\text{res}}(\mu_{1:T}; \mathcal{G})$. Combining the three bounds proves the lemma. $\square$

C.4 Dyadic Summation (Lemma 7)

Proof of Lemma 7. It is enough to prove the bound for $d \geq 1$; the root contributes $O(\sqrt{TL})$, which is absorbed by the claimed bound since $M \geq 1$.

For fixed d , define

$$F_d(x) := \min \left\{ \sqrt{TLx}, \frac{Lx}{w_d} \right\}.$$

The function F_d is nonnegative, increasing, concave on $\mathbb{R}_+$, and satisfies $F_d(0) = 0$. Hence it is subadditive:

$$F_d(x+y) \leq F_d(x) + F_d(y) \quad \text{for all } x, y \geq 0.$$

Indeed, concavity and $F_d(0) = 0$ imply $F_d(\lambda z) \geq \lambda F_d(z)$ for $\lambda \in [0, 1]$; applying this with $z = x+y$ and $\lambda = x/(x+y)$ and $\lambda = y/(x+y)$ gives the claim.

Let C_* denote the constant in the hypothesis and set

$$u_d := C_* \min \left\{ 2^d, \frac{Tw_d^2}{L}, M \right\}, \quad v_d := C_* \min \left\{ 2^d, \frac{Tw_d^2}{L}, A\sqrt{\frac{w_d}{L}} \right\}.$$

Since

$$m_d \leq C_* \min \left\{ 2^d, \frac{Tw_d^2}{L}, M + A\sqrt{\frac{w_d}{L}} \right\} \leq u_d + v_d,$$

monotonicity and subadditivity give

$$F_d(m_d) \leq F_d(u_d) + F_d(v_d).$$

Thus it suffices to bound the two sums separately.

For the M part, the grid cap can only reduce the sum, so it is enough to bound

$$\sum_d \min \left\{ Tw_d, \sqrt{TLM}, \frac{LM}{w_d} \right\},$$

up to constants depending only on C_* . The terms Tw_d and LM/w_d cross at

$$w_* = (LM/T)^{1/2}.$$

The dyadic tails on either side are geometric, and the value at the crossover is $\sqrt{TLM}$. Endpoint cases only make the sum smaller. Therefore the M part is $O(\sqrt{TLM})$.

If $A = 0$, then the A part vanishes. Otherwise, the grid cap and mass cap imply

$$F_d(v_d) \leq C \min \left\{ Tw_d, \sqrt{TA} L^{1/4} w_d^{1/4}, A\sqrt{L} w_d^{-1/2} \right\}.$$

In particular,

$$F_d(v_d) \leq C \min \left\{ Tw_d, A\sqrt{L} w_d^{-1/2} \right\}.$$

These two terms cross at

$$w_* = \left(\frac{A\sqrt{L}}{T} \right)^{2/3},$$

where both are equal to $T^{1/3} L^{1/3} A^{2/3}$. The dyadic tails on both sides of this crossover are geometric. If w_* lies outside the available depth range, the nearest endpoint gives a smaller sum. Hence the A part is $O(T^{1/3} L^{1/3} A^{2/3})$.

Combining the two sums proves the claim. $\square$

D Proofs for Tightness of the Threshold-Complexity Bound

D.1 Proof of Lemma 8

Proof of Lemma 8. 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:

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

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 $|\mu - f_{S,d}| \leq w_d$. Thus,

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

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 $K_{S,d} \geq |X(S)|$.

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

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

Because the context sequence is round-robin and $m \leq T^{1/3} \leq T$, 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 $\sum_{S \in \mathcal{P}_d} |X(S)| \geq m$. This yields the lower bound for a single fine-scale depth:

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

Step 3: Summation over depths. The condition $w_d \leq c/m$ is satisfied for all depths $d \geq \log_2(m/c)$. Given that

$$d_{\max} = \frac{1}{2} \log_2 T + O(1) \quad \text{and} \quad \log_2 m \leq \frac{1}{3} \log_2 T,$$

there are $\Omega(\log T)$ depths with $w_d \leq c/m$. Summing the above lower bound over those depths gives

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

□

D.2 Proof of Lemma 9

Proof of Lemma 9. 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 $\text{idx} : [0, 1] \rightarrow [m]$ such that $\text{idx}(x_i) = i$ for each grid point. Let $\psi_\ell^{\text{Wal}} : \{0, \dots, m-1\} \rightarrow \{\pm 1\}$ be the standard Walsh system. The signed Walsh features are then defined as:

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

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

The subsampled Walsh family is defined as $\mathcal{G}_{T,m} := \{\mathbf{1}\} \cup \{g_\ell^{\text{Wal},+}, g_\ell^{\text{Wal},-} : \ell \in S\}$, where $S \subseteq \{1, \dots, m-1\}$ is a subset of indices of size $|S| = O(\log^3 m)$.

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

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

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

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

for some constant $C > 0$. This implies that every threshold of f^* has representation norm $B \leq 1 + 2 \sum_{\ell \in S} |c_\ell| \leq C \log m$ 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 takes values on a grid of spacing α_d , is nondecreasing in x , and satisfies $|f_d(x) - x| \leq \alpha_d/2 = w_d$ for every $x \in [0, 1]$. Thus the single-block score f_d has $\mathbf{R}_{[T],d}(f_d) = 0$. 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 $K_d \leq \min \left\{ \left\lceil \frac{1}{\alpha_d} \right\rceil + 1, m \right\} = \min \{2^d + 1, m\} \leq C \min\{2^d, m\}$ 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

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

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

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

Applying Lemma 8 with $\mathcal{G} = \mathcal{G}_{T,m}$ gives

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

□

D.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

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

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

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

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

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

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 $L_0 := \log(T+1)$,

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

These imported inequalities apply for every power-of-two $m \leq T^{1/3}$; the statement in Collina et al. [2026b] specializes them to the optimized choice $m \asymp T^{1/3}$.

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

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

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

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

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

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

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

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

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

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

for universal constants $c, C > 0$.

The equivalence with $\sqrt{T\Gamma_T^{\text{thr}}}$ follows from Lemma 9. Given any target complexity budget $\Gamma_\star \geq 2$, choose m to be the largest power of two such that $m \leq \min\{\Gamma_\star, T^{1/3}\}$. Then

$$\Gamma_T^{\text{thr}}(\mu_{1:T}; \mathcal{G}_{T,m}) \leq \tilde{O}(\Gamma_\star),$$

and the lower bound above gives

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

□