# INSTANCE-ADAPTIVE ONLINE MULTICLASS MULTICALIBRATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study online multiclass multicalibration over $T$ rounds, and give an algorithm with beyond-worst-case guarantees. With $n$ possible outcomes, forecasts lie in a simplex of dimension $d = n - 1$. In the worst case, cumulative $\ell_1$ multicalibration error scales with $\widetilde{\Theta}(T^{(d+1)/(d+2)})$, which becomes untenable once $d$ exceeds a small constant. Our algorithm adapts to the geometry of the realized sequence: if the sequence of adversarially selected outcome means lies in an affine subspace of dimension $s$, it obtains error $\widetilde{O}(n^{1/(s+2)}T^{(s+1)/(s+2)})$ with high probability. The algorithm adapts to this structure without knowing the subspace or its dimension. This rate is minimax optimal up to polylogarithmic factors. For a stochastic instance with a fixed conditional mean, the bound specializes to $\widetilde{O}(\sqrt{nT})$, giving a square-root dependence on the horizon.

## 1 INTRODUCTION

Consider forecasting which of $n$ possible decision-relevant states will occur, based on an observed context $x \in \mathcal{X}$. For example, predicting which token will occur next in a sequence, where the context is the preceding token sequence. A forecast is a probability vector within the simplex of dimension $d = n - 1$; in our example, this would correspond to a vector of probabilities of the subsequent token. Calibration requires that, conditional on a forecast, the frequencies of the realized states match the forecast probabilities (Dawid, 1982). Calibrated forecasts form a transparent interface between predictions and decision making: for every downstream decision maker (whatever their utility function or action set), their best prediction-to-action policy is to take the forecasts at face value and best respond to them as if they are correct (Foster and Vohra, 1997; Kleinberg et al., 2023; Noarov et al., 2025). Multicalibration requires calibration simultaneously within a given collection of possibly overlapping groups of contexts in $\mathcal{X}$ (Hébert-Johnson et al., 2018), and extends this guarantee to policies whose context dependence is captured by those groups.

Much of the algorithmic work on (multi)calibration focuses on forecasting probabilities of *binary* outcomes, for which $d = 1$. In high-dimensional settings, worst-case multicalibration rates deteriorate sharply, both in the batch and online cases (Collina et al., 2026a;b). Our error measure is the natural multiclass generalization of expected calibration error (ECE), in cumulative form: within each group, sum the $\ell_1$ deviations between each forecast and its empirical outcome frequencies, weighted by how often that forecast occurs, then take the maximum over groups. The online worst-case rate for this multicalibration error is $\widetilde{\Theta}(T^{(d+1)/(d+2)})$ in the number of rounds $T$. This is dispiriting both because many prediction tasks we care about are fundamentally multiclass, like language prediction, and because the decision-making motivation for calibration is much stronger when forecasts can describe a richer state space. This motivates us to ask:

> *Can an online algorithm make predictions that are guaranteed to be multicalibrated against adversarial sequences, yet adapt to the structure of the realized sequence to obtain substantially better rates on easier instances?*

We give a positive answer: our algorithm adapts to the geometry of the sequence of outcome distributions. At every round, the learner observes a context and commits to a distribution over predictions. The adversary chooses an outcome distribution that may depend on the history, the current context, and the learner's distribution over predictions, but not on its freshly sampled prediction. The



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

learner observes the sampled outcomes, but our error guarantee adapts to the geometric structure of their conditional means, which are not directly observed. For categorical outcomes, each mean is the vector of outcome probabilities; our model also permits observations that are themselves probability vectors. We say the sequence of conditional means has affine dimension at most $s$ if, after a possible additive shift, it lies in a linear subspace of dimension $s$. Whenever this is the case, our algorithm achieves multicalibration error

$$\widetilde{O}\Big(n^{1/(s+2)}T^{(s+1)/(s+2)}\Big)$$

with high probability (Theorem 1).

Importantly, our algorithm adapts to this structure as the sequence is realized, without knowing the subspace or its dimension. In the special case where the conditional outcome mean is fixed, the said outcome mean lies at a single point and $s = 0$. The bound becomes $\widetilde{O}(\sqrt{nT})$, giving a square root dependence on the horizon even when the ambient dimension is large. More generally, suppose each round's conditional outcome distribution is a mixture of $k$ fixed, possibly full-support distributions, with adversarially varying mixture weights. The conditional means lie in the convex hull of the component means, so $s \leq k - 1$, regardless of $n$. At $s = n - 1$, the bound recovers the worst-case horizon exponent. The guarantee also tolerates means lying close to a low-dimensional affine set, with the approximation tradeoff specified in Remark 1. For every $0 \leq s \leq n - 1$ and $T \geq n \geq 2$, Theorem 2 gives a lower bound matching this rate in expectation, up to polylogarithmic factors, on an instance whose means have affine dimension at most $s$, with a suitably chosen family of $O(\log^3(T + 1))$ many groups. Its outcome distributions and groups are fixed before play.

Our algorithm learns its effective prediction range from observed outcomes. For each group and available forecast, the algorithm estimates the frequencies of outcomes it observes, with each round contributing in proportion to the probability assigned to that forecast and its membership weight in the group. It stores estimates at several levels of precision, adding a forecast to the range when no value already stored at that level is sufficiently close.

The algorithm combines this range discovery with machinery for ensuring second-order swap-regret minimization under a logarithmic proper scoring rule. We require the pairwise comparison bounds defining swap regret to hold within every group and on every interval during which both forecasts are available. The guarantee of low swap regret for a forecast close to the estimated outcome frequencies then allows us to bound calibration error. The formal argument requires more finesse as the relevant comparator (prediction) may only be discovered after the rounds on which it would have helped. We form estimates from nested stretches of observations: longer stretches support finer distinctions, while shorter ones let the analysis use comparators available at their start. With enough data, these estimates lie close to averages of the adversary's conditional means. Their separation at each level means that each set forms a packing; this then implies that low-dimensional geometry bounds both the number of forecasts discovered and the error accumulated during discovery. Finally, the algorithm uses its observed calibration error to choose how finely to distinguish forecasts.

### 1.1 Related work

**Instance-adaptive calibration.** The most closely related works are Huang et al. (2026) and Liu et al. (2026), who obtain instance-adaptive calibration guarantees for one-dimensional predictions. Huang et al. (2026) preserve the worst-case $\widetilde{O}(T^{2/3})$ cumulative error bound for binary multicalibration while improving on structured instances, including a $\widetilde{O}(\sqrt{T})$ bound for i.i.d. outcomes in the marginal setting, without context-defined groups. Liu et al. (2026) give scalar marginal calibration bounds that improve when the outcome mean stays nearly constant or changes only a few times. We study multiclass multicalibration, where forecasts are probability vectors and worst-case rates deteriorate with the number of outcomes. This is not a coordinatewise extension of scalar calibration: our guarantee conditions on the entire predicted vector, and the low-dimensional structure need not align with the coordinate axes.

**Multiclass and high-dimensional calibration.** For online marginal $\ell_1$ calibration, Peng (2025) and Fishelson et al. (2025a) achieve any constant average calibration error after a number of rounds polynomial in dimension. Like us, they construct forecasts from averages of past observations. Our guarantees additionally hold within groups and adapt to the affine dimension of the conditional means. For instances with low dimensional structure, we give improved rates.

In the batch setting, Gopalan et al. (2024) achieve projected smooth calibration with both sample size and running time polynomial in the number of classes at fixed accuracy. This guarantee ensures that, for every subset of labels, the total probability assigned to that subset is close to a calibrated binary predictor for membership in the subset. This is weaker than full calibration.

**Worst-case rates.** Even for binary marginal calibration, Qiao and Valiant (2021) rule out a worst-case $\widetilde{O}(\sqrt{T})$ cumulative error bound, while Dagan et al. (2025) obtain an upper bound below $T^{2/3}$. For multicalibration, Collina et al. (2026a) establish a binary $\widetilde{\Omega}(T^{2/3})$ cumulative error lower bound, with a forthcoming revision extending the bound to $\widetilde{\Omega}(T^{(d+1)/(d+2)})$ in ambient dimension $d$. Our lower bound builds on their construction while separating the affine dimension of the conditional means from the number of outcome coordinates (Theorem 2).

**Other online calibration methods.** Randomization makes calibration possible against arbitrary outcome sequences (Foster and Vohra, 1998), with internal- and swap-regret reductions providing an algorithmic route (Foster and Vohra, 1997; Blum and Mansour, 2007). Gupta et al. (2022) give online group-conditional guarantees for mean predictions, moments, and prediction intervals. For scalar outcomes, Noarov et al. (2025); Ghuge et al. (2025) obtain $\widetilde{O}(T^{2/3})$ cumulative ECE for finite group families, while Garg et al. (2024) reduce multicalibration over infinite benchmark classes to online learning. For scalar predictions, Hu et al. (2026) use online-learning benchmarks to control squared calibration error; Fishelson et al. (2025b) identify probability-weighted squared marginal calibration error with full swap regret. Our prediction component combines the subsequence-regret framework of Lee et al. (2022) with the second-order comparison update of Gaillard et al. (2014).

## 2 Problem and Results

### 2.1 Online multicalibration

We denote $[q] = \{1, \dots, q\}$ for any positive integer $q$. Fix a horizon $T \in \mathbb{N}_{\geq 1}$, confidence level $\delta \in (0, 1)$, and number of outcome coordinates or classes $n \geq 1$. A $n$-multiclass prediction lies in the simplex $\Delta_{n-1} = \left\{p \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^n p_i = 1\right\}$, whose vertices are the standard basis vectors $e_1, \dots, e_n$. One such prediction represents a distribution over the $n$ classes, naturally interpreted as the probability of a given class label.

At each round $t$, the learner observes a context $x_t \in \mathcal{X}$ and announces a probability distribution $\pi_t$ over $\Delta_{n-1}$. It samples and announces a prediction $p_t \sim \pi_t$ using fresh private randomness, and then observes an outcome $y_t \in \Delta_{n-1}$. Categorical observations are the special case $y_t \in \{e_1, \dots, e_n\}$; both the upper and lower bounds below use the model allowing any $y_t \in \Delta_{n-1}$.

We assume an adversary generates the observations $y_t$ above, in a way which can depend upon the past, on $x_t$, and on the announced mixture $\pi_t$, but not on the fresh sample. Let $(\mathcal{F}_t^-)_{t=1}^T$ be the filtration, where $\mathcal{F}_t^-$ is generated by the information available immediately before $p_t$ is sampled, including the past, the current context $x_t$, the learner's internal state, and the announced distribution $\pi_t$. Write

$$\mu_t := \mathbb{E}[y_t \mid \mathcal{F}_t^-].$$

The fresh private prediction $p_t$ is conditionally independent of $y_t$ given $\mathcal{F}_t^-$.

Fix a known finite set $\mathcal{G}$ which we will call *groups*, where any $g \in \mathcal{G}$ is a function $g : \mathcal{X} \to [0, 1]$. For any $x_t \in \mathcal{X}$, the learner can evaluate $g(x_t)$ exactly for every $g \in \mathcal{G}$. The realized multiclass multicalibration error of an algorithm on a sequence is defined as

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) := \max_{g \in \mathcal{G}} \sum_{r \in \{p_1, \dots, p_T\}} \left\| \sum_{t : p_t = r} g(x_t)(y_t - r) \right\|_1 .$$

For a group of total weight $W_g = \sum_t g(x_t) > 0$, its term in the maximum is $W_g$ times its weighted multiclass expected calibration error (ECE). We augment the set of groups to contain the constant group $\mathbf{1}(x) = 1$ and denote it as $\overline{\mathcal{G}} = \mathcal{G} \cup \{\mathbf{1}\}$.

### 2.2 Adaptation to affine dimension

We now introduce the machinery necessary to define when the sequence of conditional means has low effective dimension. This roughly translates to belonging to a low dimensional linear subspace, though we allow for the subspace to be offset from the origin.

**Definition 1** (Affine dimension)**.** *Let $S \subseteq \mathbb{R}^n$ be nonempty. Its affine dimension, denoted* $\mathrm{affdim}(S)$*, is the dimension of the smallest affine subspace containing $S$. An affine subspace is a translate $a + V$ of a linear subspace $V$, and has dimension* $\dim V$.

For example, a singleton has affine dimension 0, a nondegenerate line segment has affine dimension 1, and a planar set not contained in a line has affine dimension 2. The simplex $\Delta_{n-1}$ has affine dimension $n-1$. These sets can also be translated away from the origin by any constant vector and have the same affine dimension.

The following theorem bounds our algorithm's multicalibration error in terms of the affine dimension of the conditional means.

**Theorem 1** (Uniform affine-dimensional upper bound)**.** *With the adaptive rule described in Section 5, Algorithm 1 takes inputs $T, n, \delta, \mathcal{G}$ and has the following guarantee against every adaptive adversary. Suppose after the realization of any adaptive sequence, $s = \mathrm{affdim}\{\mu_1, \dots, \mu_T\}$. Then with probability at least $1-\delta$, the realized predictions satisfy*

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le C \left[\log\left(\frac{eTn|\overline{\mathcal{G}}|}{\delta}\right)\right]^2 n^{1/(s+2)} T^{(s+1)/(s+2)}. \tag{1}$$

*where $C > 0$ is a universal constant.*

The algorithm does not need to know $s$ or the affine subspace containing the conditional means. The guarantee also extends to conditional means that lie near a low-dimensional affine set, with an additional error term measuring their distance from that set.

**Remark 1** (Approximate low-dimensional structure)**.** *For a nonempty closed convex set $S \subseteq \Delta_{n-1}$, write $\rho_\infty(S) := \max_{t \le T} \inf_{u \in S} \|\mu_t - u\|_1$. On the same event, simultaneously for every such $S$ and integer $s$ with $\mathrm{affdim}(S) \le s \le n-1$,*

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le \widetilde{O}\left(n^{1/(s+2)} T^{(s+1)/(s+2)} + T\sqrt{\rho_\infty(S)}\right).$$

*The $\widetilde{O}$ hides only the squared logarithm in* (1)*; $s, S$ may be chosen after play.*

We complement this guarantee with a worst-case lower bound with matching dependence on $n, T$, and the affine dimension of the conditional means, up to polylogarithmic factors.

**Theorem 2** (Affine-dimensional lower bound)**.** *There are universal constants $c, C > 0$ such that, for all integers $n \ge 2$, $T \ge n$, and $0 \le s \le n-1$, there is a stochastic instance with a family $\mathcal{G}$ of at most $C \log^3(T+1)$ binary groups. Its outcome means satisfy $\mathrm{affdim}\{\mu_1, \dots, \mu_T\} \le s$, and every possibly randomized online forecaster incurs*

$$\mathbb{E}\left[\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T)\right] \ge c \frac{n^{1/(s+2)} T^{(s+1)/(s+2)}}{\log^2(T+1)}.$$

*The context sequence is deterministic, and the groups and outcome distributions are fixed before play. For $s = 0$, categorical outcomes and the constant group suffice, with lower bound $c\sqrt{nT}$.*

The main construction averages each point of a low-dimensional simplex grid with a uniformly chosen simplex vertex. The conditional means lie in an affine image of the grid, but the centered noise uses all $n$ coordinates. Appendix D gives the proof, building on Collina et al. (2026a).

## 3 The algorithm

We now present the algorithm which proves Theorem 1. In order to naturally adapt to low-dimensional instances, the algorithm combines a swap-regret prediction rule with discovery of new forecasts from

weighted outcome averages. This discovery protocol maintains collections of predictions at various discretizations of accuracy levels. At each accuracy level, the predictions are separated from each other, leading them to form a *packing*. As weighted observations accumulate, the discovery procedure may add new prediction values.

We first describe the algorithm with a fixed number $1 \le K \le K_{\max} := 2 + \lceil \log_2 T \rceil$ of accuracy levels, while allowing the stored prediction sets to grow through discovery. Section 5 then explains how to adjust $K$ using the observed calibration error. The algorithm has runtime polynomial in $T, n, |\mathcal{G}|, \log(1/\delta)$, as formalized in Appendix C.4.

### 3.1 Prediction rule, smoothing, and loss

Let $\mathcal{R}_t$ be the set of available prediction values at round $t$, initialized as $\mathcal{R}_0 = \{q_0\} = \{(1/n, \dots, 1/n)\}$. The next subsection describes how our discovery protocol maintains this set. Based on the history and current context $x_t$, a second-order swap-regret forecaster chooses a distribution $\pi_t$ over $\mathcal{R}_t$. The learner samples and announces $p_t \sim \pi_t$ using fresh private randomness.

After observing $y_t$, the learner forms the smoothed outcome $y'_t = (1-\beta)y_t + \beta q_0$, where $\beta = 1/(2T^2)$, and updates the forecaster using the logarithmic losses[1]

$$\ell(a, y'_t) = -\sum_{i=1}^{n} y'_{t,i} \log a_i, \qquad a \in \mathcal{R}_t.$$

The discovery protocol construction in Section 3.2 ensures that every available prediction has coordinates bounded from below by $\beta/n$, keeping these losses bounded while changing the final calibration error by at most $2\beta T$.

The forecaster controls the loss advantage of switching between two predictions, within each group and over every interval during which both predictions are available. Appendix A.1 specifies the prediction rule, and Lemma 1 states its guarantee.

### 3.2 Discovering prediction values at several accuracies

The discovery procedure generates additional prediction values from weighted averages of smoothed outcomes. We first describe how these averages are formed, then explain when they are added to the available predictions.

**Weighted observations and blocks.** For each group $g \in \overline{\mathcal{G}}$ and prediction $a \in \mathcal{R}_t$, define the round-$t$ weight $\alpha_t^{g,a} = g(x_t)\pi_{t,a}$, where $\pi_{t,a}$ is the probability assigned to $a$. We call $a$ the source prediction and write $\alpha_t$ when $g, a$ are clear from context. For each pair $(g, a)$, we collect the smoothed outcomes $y'_t$, with weights $\alpha_t^{g,a}$, into *blocks* of rounds. Each block $E$ spans a contiguous interval of rounds. For a block $E$ associated with $(g, a)$, define its total weight and weighted outcome sum by

$$N_E := \sum_{t \in E} \alpha_t^{g,a}, \qquad V_E := \sum_{t \in E} \alpha_t^{g,a} y'_t.$$

Both sums are zero for an empty block. When $N_E > 0$, the block's weighted mean is $z_E = V_E/N_E$.

**Combining observations into blocks.** To form candidate prediction values at different accuracies, we maintain blocks for each pair $(g, a)$ at $K$ levels, indexed by $j = 0, \dots, K-1$ from coarsest to finest. For each pair $(g, a)$ and level $j$, we keep at most one *open block*, which is still collecting weight until reaching a completion threshold $m_j$, and then we close it and call it *completed*.

At level zero, the open block directly takes weight $\alpha_t$ and smoothed outcomes $y'_t$. At level $j \ge 1$, the open block can only collect past observations $(N_E, V_E)$ from a *completed* block $E$ at level $j-1$. Once a block closes, the next input starts a new block $E'$ with $N_{E'}$ and $V_{E'}$ initialized at zero, as shown in Figure 1.

[1] Excess logarithmic loss controls $\ell_1$ error through Pinsker's inequality without a dimension factor; a direct squared-loss argument controls $\ell_2$ error, whose conversion to $\ell_1$ can cost $\sqrt{n}$.

Appendix F specifies the exact choice of $L, \kappa$. The completion threshold at level $j$ is $m_j = 1024nL/w_j^2$, where $w_j = 2^{1-j}$ is the separation threshold used below. The thresholds are chosen so that exactly four completed blocks, called *children*, form a completed block at the next level, their *parent*. Every completed level-$j$ block has weight in $[m_j, 2m_j)$.

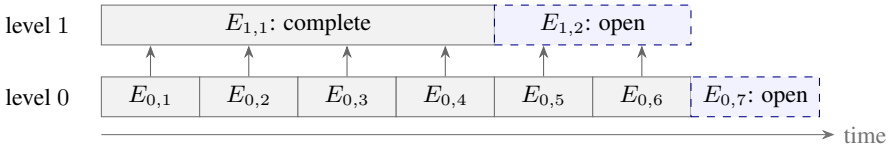


Figure 1: Blocks for one $(g, a)$ pair, where $E_{j,r}$ denotes the $r$-th block at level $j$ in chronological order. Completed blocks $E_{0,1}, \ldots, E_{0,4}$ form $E_{1,1}$. The open block $E_{1,2}$ has received the summaries of $E_{0,5}, E_{0,6}$. The block $E_{0,7}$ is still collecting observations and has not been forwarded. Arrows add both weight and outcome sum.

**Algorithm 1** Learning predictions with $K$ accuracy levels

**Require:** $T, n, \delta, \mathcal{G}$, number of levels $K \in \{1, \ldots, K_{\max}\}$.
Initialize $P_0 = \{q_0\}$, $P_j = \varnothing$ for $1 \le j \le K_{\max}$, empty blocks, and the prediction rule of Appendix A.1.
Order groups as supplied, appending $\mathbf{1}$ if absent; order prediction values by first discovery, starting with $q_0$, and use this order for ties. Repeated values keep their original position.
**for** $t = 1, \ldots, T$ **do**
  Observe $x_t$ and set $\mathcal{R}_t \leftarrow P_0 \cup \bigcup_{j=1}^{K-1} P_j$.
  Use the prediction rule to choose $\pi_t$ on $\mathcal{R}_t$; sample and announce $p_t \sim \pi_t$.
  Observe $y_t$, form $y'_t$, and update the swap-regret forecaster using the losses $\ell(a, y'_t)$ for $a \in \mathcal{R}_t$.
  **for all** $g \in \overline{\mathcal{G}}$, $a \in \mathcal{R}_t$ with $\alpha_t^{g,a} = g(x_t)\pi_{t,a} > 0$ **do**
    Start a zero-level block $E$ with $(N_E, V_E) = (0, 0)$ if absent.
    $N_E \leftarrow N_E + \alpha_t^{g,a}$, $V_E \leftarrow V_E + \alpha_t^{g,a} y'_t$.
  **end for**
  **while** a zero-level block has $N_E \ge m_0$ **do**
    Choose the first such block in source-prediction order, then group order.
    Call ProcessThresholdBlock$(E, 0, g, a)$.
  **end while**
**end for**
**procedure** ProcessThresholdBlock$(E, j, g, a)$
  Close $E$ and set $z_E \leftarrow V_E/N_E$.
  **if** $j \ge 1$ and $H(b, z_E) > w_j$ for every $b \in P_j$ **then**
    **for** $k = 1, \ldots, j$ **do**
      If $H(b, z_E) > w_k$ for every $b \in P_k$, insert $z_E$ into $P_k$.
    **end for**
  **end if**
  **if** $j < K - 1$ **then**
    Get the open level-$(j+1)$ block $E'$ of $(g, a)$, starting it at $(0, 0)$ if absent.
    $N_{E'} \leftarrow N_{E'} + N_E$, $V_{E'} \leftarrow V_{E'} + V_E$.
    If $N_{E'} \ge m_{j+1}$, call ProcessThresholdBlock$(E', j+1, g, a)$.
  **end if**
**end procedure**

**Stored predictions and insertion.** To decide which completed-block means to store as new candidate prediction values, we maintain a set of stored predictions at each level shared across all $(g, a)$ pairs. At level 0, we always have a fixed set $P_0 = \{q_0\}$. At each positive level $j$, all groups and source predictions share an initially empty set $P_j$ of stored prediction vectors. Distinct values in $P_j$ are kept more than $w_j$ apart in the Hellinger distance $H(p, q) = \|\sqrt{p} - \sqrt{q}\|_2$, with coordinatewise square roots.

When a block $E$ at a positive level $j$ is completed, compare its mean $z_E$ with the values in $P_j$. If some value in $P_j$ is within Hellinger distance $w_j$ of $z_E$, no insertion is made. Otherwise, consider every lower level $k = 1, \ldots, j$. At each level $k$, insert $z_E$ into $P_k$ if $H(b, z_E) > w_k$ for every $b \in P_k$.

**Available predictions.** The forecaster uses the union of these stored sets. We write $P_j(t)$ for the contents of $P_j$ at the start of round $t$, so the available prediction set is

$$\mathcal{R}_t = P_0 \cup \bigcup_{j=1}^{K-1} P_j(t).$$

In what follows, we write $P_j$ for the stored set at the time under consideration, making the time explicit only when needed. After all discovery updates on round $t$, the resulting sets are $P_j(t+1)$. Since the stored sets only grow, $\mathcal{R}_t$ grows monotonically while $K$ remains fixed. Algorithm 1 specifies these operations.

## 4 Bounding the Algorithm's calibration error

We bound the calibration error of Algorithm 1 in three steps. First, we bound a block's weighted error when a prediction close to its empirical mean is available throughout the block. Next, we partition the observations into residual pieces to account for predictions discovered later. Finally, we use the low-dimensional geometry of the conditional means to bound the resulting discovery cost.

The rule in Section 5 may decrease $K$, clearing blocks but retaining stored prediction sets. Call a maximal stretch with the same $K$ a *depth-$K$ interval*.

### 4.1 An accurate available comparator controls error

We now present a lemma which shows that the weighted difference of loss between two values $a, q$ is controlled by a quadratic function of that same difference in losses. This implies that if there exists a low loss prediction, the predictions made by our algorithm must also have low loss.

**Lemma 1** (Interval loss-comparison bound)**.** *The prediction rule in Appendix A.1 satisfies the following on every realization. For every $g \in \overline{\mathcal{G}}$, pair $a, q$, and contiguous interval $I \subseteq [T]$ on which both predictions are available, let $\alpha_t = g(x_t)\pi_{t,a}$ and $X_t = \ell(a, y'_t) - \ell(q, y'_t)$. Then*

$$\sum_{t \in I} \alpha_t X_t \le \frac{1}{8\kappa} \sum_{t \in I} \alpha_t X_t^2 + 8\kappa L. \tag{2}$$

With sums over $I$, write $N = \sum_t \alpha_t > 0$, $z = N^{-1} \sum_t \alpha_t y'_t$, and $D(z\|a) = \sum_i z_i \log(z_i/a_i)$. Substituting in logarithmic loss gives the identity $\sum_t \alpha_t X_t = N[D(z\|a) - D(z\|q)]$. The quadratic terms and thus the right-hand side of the inequality can be bounded using these same divergences: $\sum_t \alpha_t X_t^2 \le 2\kappa N[D(z\|a) + D(z\|q)]$. Substitution into (2) lets us absorb the source divergence into the left-hand side, yielding the multiplicative comparison $D(z\|a) \le 3D(z\|q) + 16\kappa L/N$.

An accurate comparator therefore forces the source prediction to be accurate, with an additive divergence cost of order $\kappa L/N$. The following lemma converts this divergence comparison into bounds on the distance between the source prediction $a$ and the weighted outcome mean $z$, when the comparator $q$ is available throughout the interval.

**Lemma 2** (Available-comparator bound)**.** *The two bounds below hold on every realization. For every group and contiguous round interval on which $a, q$ are both available, with $N, z$ defined above,*

$$H(z, a) \le \sqrt{3\kappa} H(z, q) + 4\sqrt{\kappa L/N}. \tag{3}$$

*If $N \ge m_j$ and $H(z, q) \le 2w_j$, then*

$$\|z - a\|_1 \le 6\sqrt{\kappa}\, w_j. \tag{4}$$

### 4.2 Accounting for comparators discovered later

The preceding argument controls a block whenever a prediction close to its empirical mean is available throughout the block. The main complication is that such a comparator may be discovered only when the block is completed.

Let $I$ consist of all rounds from the start of a depth-$K$ interval through any round within it. We handle this by recursively decomposing the weighted observations in $I$:

- For each completed block $E$ at a level $j > 0$, let $q_E$ be the first stored candidate prediction in $P_j$ within Hellinger distance $w_j$ of its empirical mean $z_E$, inserting $z_E$ itself if no such prediction exists. If $q_E$ was available throughout $E$, we analyze $E$ directly. Otherwise, we replace $E$ by its children and repeat. At level zero the fixed prediction $q_0$ is always available, so this recursion terminates.
- For each unfinished block $E$ at a level $j > 0$ at the end of $I$, we apply the same rule to its completed children. Any unfinished level-zero block with positive weight is retained as a residual piece.

The resulting *residual pieces* partition the weighted observations, as shown in Fig. 2.

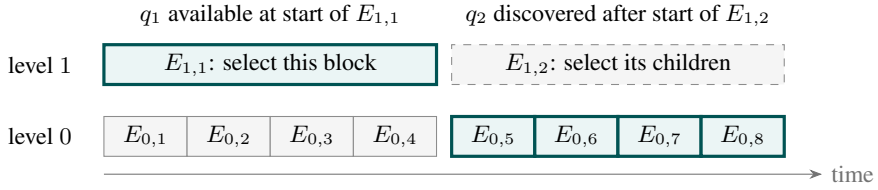


Figure 2: Selecting pieces for one group–source pair. Both parents are complete, with comparators $q_i = q_{E_{1,i}}$ satisfying $H(q_i, z_{E_{1,i}}) \le w_1$. Select $E_{1,1}$ because $q_1$ was available throughout it; replace $E_{1,2}$ by its children because $q_2$ was unavailable at its start. The blocks with thick outlines partition the observations. This selection is used only in the proof.

For each pair $(g, a)$, the triangle inequality gives

$$\left\| \sum_{t \in I} \alpha_t^{g,a} (y'_t - a) \right\|_1 \le \sum_E N_E \|z_E - a\|_1,$$

where $E$ ranges over all residual pieces in $I$. The sampling and smoothing corrections given below then convert this into a bound on the realized calibration error. For every completed level-$j$ piece, the available-comparator bound gives

$$N_E \|z_E - a\|_1 \le 6\sqrt{\kappa}\, N_E w_j. \tag{5}$$

Thus, calibration error is small on fine pieces, and the pieces at the finest level $K-1$ contribute at most $6\sqrt{\kappa}|I|w_{K-1}$ due to $\sum_E N_E \le |I|$.

We next control the contribution of coarser pieces by bounding how many are selected. Call a block a *boundary block* if its useful comparator was discovered after the block began. A completed piece below the finest level must have a boundary or unfinished parent, which can be bounded by the following two observable quantities. Let $M_I$ be the number of stored prediction values available by the end of an interval $I$, and let $B_I$ be the maximum between 1 and the largest number of boundary blocks for any fixed group, source prediction, and level.

As shown in Lemma 10, each group has $O(M_I B_I)$ completed pieces at each coarser level. Combining these counts with the fact $N_E \le 2m_j$ and (5) gives the coarse-scale term $O(\sqrt{\kappa}\, nLM_I B_I / w_{K-1})$.

We also need to account for the sampling and smoothing cost. On the common event $\mathcal{E}_{\text{global}}$ of Lemma 5, which has probability at least $1-\delta$, the sampling correction on each interval $I$ is at most $4\sqrt{nLM_I T}$. Smoothing adds at most $2\beta|I|$. Write $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I)$ for the realized multicalibration error restricted to rounds in $I$. Combining these corrections with the residual-piece bounds gives the following corollary.

**Corollary 1** (Calibration error from discovery counts)**.** *There is a universal constant $C$ such that, on $\mathcal{E}_{\text{global}}$, every interval $I$ defined above, with $K \ge 2$ and terminal resolution $w = w_{K-1}$, satisfies*

$$\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I) \le C\sqrt{\kappa}\left(|I|w + \frac{nLM_I B_I}{w}\right) + 4\sqrt{nLM_I T} + 2\beta|I|.$$

The term $|I|w$ is the ordinary approximation cost at the terminal resolution. The term $nLM_I B_I / w$ is the price of discovering useful prediction values online. The formal residual decomposition and piece-counting argument appear in Appendices B.5–B.7.

### 4.3 LOW-DIMENSIONAL MEANS LIMIT DISCOVERY

Next, we bound $M_I$ and $B_I$ using the low-dimensional geometry of the conditional means. Suppose the means lie in, or sufficiently near, a convex set $S$ of affine dimension $s$. On the event $\mathcal{E}_{\mathrm{global}}$, every prediction stored at a positive level is close to the smoothed set $S_\beta = (1-\beta)S + \beta q_0$.

To bound $M_I$, we use the $w_j$-separation of predictions in $P_j$. A packing argument near $S_\beta$ bounds their number at each level. Summing over levels gives the bound on $M_I$.

To bound $B_I$, we exploit the chronological order of boundary blocks. Fix a group, source prediction $a$, and positive level $j$. Near $a$, the $w_j$-separation of boundary comparators gives a local packing bound. Far from $a$, every earlier comparator is available throughout each later block. The Hellinger bound (3) then gives separation proportional to the later comparator's distance from $a$, up to $\kappa$-dependent factors. Within distance ranges with doubling radii, this gives a local packing bound independent of the radius. Summing over the near region and these ranges yields a bound on $B_I$ with no explicit dependence on $w$. The following lemma states both bounds.

**Lemma 3** (Affine prediction and boundary-block counts)**.** *There are universal constants $a, b \geq 1$ such that, on $\mathcal{E}_{\mathrm{global}}$, for every nonempty closed convex set $S \subseteq \Delta_{n-1}$ of affine dimension $s$, every $2 \leq K \leq K_{\max}$ with $w = w_{K-1} \geq 64\sqrt{\rho_\infty(S)}$, and every interval $I$ at depth $K$ as defined in Section 4.2,*

$$M_I \leq K_{\max}(a\sqrt{\kappa}/w)^s, \qquad B_I \leq K_{\max}(b\kappa)^s.$$

Appendix B.4 establishes the global and local Hellinger packing bounds used to prove Lemma 3. Combining Corollary 1 and Lemma 3 gives a calibration bound in terms of the resolution $w$ for a depth-$K$ interval. We next show how to adjust this depth $K$ using the observed calibration error, without knowing $S$ or $s$.

## 5 CHOOSING THE RESOLUTION

Start at the finest permitted resolution, $K = K_{\max}$. Define calibration error budget for each resolution $\Gamma_K = C_{\mathrm{test}}\sqrt{\kappa}\,T w_{K-1}$, where $C_{\mathrm{test}}$ is the fixed universal constant in Appendix F. Let $I$ be the rounds spent at the current $K$ so far. After each round, including all discovery updates, compute $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I)$ from the realized predictions and outcomes in $I$. If $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I) > \Gamma_K$, $K > 1$, and $t < T$, end $I$, decrease $K$ by one, and begin a new interval with empty block accumulators and calibration error reset to zero. Retain all stored prediction sets, but only levels $0, \ldots, K-1$ for the new value of $K$ contribute to the next round's prediction range. A value becomes unavailable only if it belongs to none of these levels, even if it is still stored at a higher level. Comparison states for predictions that remain available are preserved. Appendix C.1 gives the pseudocode.

**Why the test finds a suitable resolution.** The exponents of $n$ and $T$ in the final error bound arise by balancing the finest-level error against the cost of discovery:

$$Tw \asymp \frac{n}{w^{s+1}}, \qquad w \asymp (n/T)^{1/(s+2)}.$$

The omitted coefficients add only logarithmic factors to the balancing width after taking their $(s+2)$-th roots. The sampling term is bounded by the geometric mean of the two full error terms (Appendix C.3). For analysis, choose a comparison resolution by rounding the larger of a sufficiently large constant multiple of this width and $64\sqrt{\rho_\infty(S)}$ up to a dyadic value. If it exceeds one, the bound $2T$ suffices. Otherwise every prefix at this resolution has error below its budget, so the algorithm cannot move past it. This resolution depends on $s, S$ only in the analysis.

Each interval's error is at most $\Gamma_K + 2$: one round changes $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I)$ by at most two. The triangle inequality over the intervals gives $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(T) \leq \sum_I \mathrm{MCerr}^{(1)}_{\mathcal{G}}(I)$. Successive budgets double, so their sum is at most twice the budget at the comparison resolution; the overshoots add at most $2K_{\max}$. Solving the balance gives Theorem 1 and Remark 1; Appendix C.3 supplies the rounding, constant, and small-horizon calculations.

## AI USE STATEMENT

In this work, generative AI tools were used to polish the manuscript, assist in drafting and revising proofs, and prepare the initial draft. We have not used generative AI tools for retrieval and discovery of related work, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted content, checked it for correctness, and extensively revised it for clarity and readability. We take responsibility for the final content of this work, including text and theoretical claims.

## REFERENCES


Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007. URL https://www.jmlr.org/papers/v8/blum07a.html.

Sergey G. Bobkov. On Milman's ellipsoids and M-position of convex bodies. In *Concentration, Functional Inequalities and Isoperimetry*, volume 545 of *Contemporary Mathematics*, pages 23–33. American Mathematical Society, 2011. https://doi.org/10.1090/conm/545/10762.

Natalie Collina, Jiuyao Lu, Georgy Noarov, and Aaron Roth. Optimal lower bounds for online multicalibration. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS)*, 2026a. URL https://arxiv.org/abs/2601.05245. To appear.

Natalie Collina, Jiuyao Lu, Georgy Noarov, and Aaron Roth. The sample complexity of multicalibration. arXiv preprint arXiv:2604.21923, 2026b. URL https://arxiv.org/abs/2604.21923.

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. https://doi.org/10.1145/3717823.3718178.

A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379): 605–610, 1982. https://doi.org/10.1080/01621459.1982.10477856.

Maxwell Fishelson, Noah Golowich, Mehryar Mohri, and Jon Schneider. High-dimensional calibration from swap regret. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 50433–50465, 2025a. https://doi.org/10.52202/085713-1682.

Maxwell Fishelson, Robert Kleinberg, Princewill Okoroafor, Renato Paes Leme, Jon Schneider, and Yifeng Teng. Full swap regret and discretized calibration. In *Proc. International Conference on Algorithmic Learning Theory (ALT)*, volume 272, pages 444–480. PMLR, 2025b. URL https://proceedings.mlr.press/v272/fishelson25a.html.

Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1–2):40–55, 1997. https://doi.org/10.1006/game.1997.0595.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998. https://doi.org/10.1093/biomet/85.2.379.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. https://doi.org/10.1214/aop/1176996452.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proc. Conference on Learning Theory (COLT)*, volume 35, pages 176–196. PMLR, 2014. URL https://proceedings.mlr.press/v35/gaillard14.html.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proc. Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2725–2792, 2024. https://doi.org/10.1137/1.9781611977912.98.

Rohan Ghuge, Vidya Muthukumar, and Sahil Singla. Improved and oracle-efficient online $\ell_1$-multicalibration. In *Proc. International Conference on Machine Learning (ICML)*, volume 267, pages 19437–19457. PMLR, 2025. URL https://proceedings.mlr.press/v267/ghuge25a.html.

Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In *Proc. Conference on Learning Theory (COLT)*, volume 247, pages 1983–2026. PMLR, 2024. URL https://proceedings.mlr.press/v247/gopalan24a.html.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *Proc. Innovations in Theoretical Computer Science Conference (ITCS)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 82:1–82:24, 2022. https://doi.org/10.4230/LIPIcs.ITCS.2022.82.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proc. International Conference on Machine Learning (ICML)*, volume 80, pages 1939–1948. PMLR, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Lunjia Hu, Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Efficient swap multicalibration of elicitable properties. In *Proc. Conference on Learning Theory (COLT)*, volume 336, pages 3314–3348. PMLR, 2026. URL https://proceedings.mlr.press/v336/hu26b.html.

Zhiming Huang, Jamie Morgenstern, Aaron Roth, and Claire Jie Zhang. Instance-adaptive online multicalibration. arXiv preprint arXiv:2605.09273, 2026. URL https://arxiv.org/abs/2605.09273.

Robert Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *Proc. Conference on Learning Theory (COLT)*, volume 195, pages 5143–5145. PMLR, 2023. URL https://proceedings.mlr.press/v195/kleinberg23a.html.

Daniel D. Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 29051–29063, 2022. https://doi.org/10.52202/068431-2106.

Junyan Liu, Haipeng Luo, and Lillian J. Ratliff. Adaptive calibration in non-stationary environments. arXiv preprint arXiv:2605.11490, 2026. URL https://arxiv.org/abs/2605.11490.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. In *Proc. International Conference on Machine Learning (ICML)*, volume 267, pages 46762–46783. PMLR, 2025. URL https://proceedings.mlr.press/v267/noarov25b.html.

Binghui Peng. High dimensional online calibration in polynomial time. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS)*, 2025. URL https://arxiv.org/abs/2504.09096.

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. https://doi.org/10.1145/3406325.3451050.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. https://doi.org/10.1214/ECP.v16-1624.

Joel A. Tropp. Probability theory & computational mathematics. CMS/ACM 117 lecture notes, California Institute of Technology, 2024. URL https://tropp.caltech.edu/notes/Tro24-Probability-Theory-LN.pdf. Fall 2024 course, version typeset January 23, 2025.

# A Prediction and block-processing details

This appendix specifies the prediction rule and proves the block-weight and discovery-count bounds for Algorithm 1.

## A.1 Prediction component

We give a prediction rule based on second-order swap-regret minimization under the logarithmic proper scoring rule.

The parameters in the update rules below are fixed before play as functions of the inputs, with one feasible choice given in Appendix F.

The discovery analysis uses one property of the prediction component. For every group $g$, source prediction $a$, comparator prediction $b$, and interval on which both predictions are active, it needs a pathwise bound on the logarithmic-loss advantage of switching from $a$ to $b$, weighted by $g(x_t)\pi_{t,a}$. The bound must remain valid as prediction values become active or inactive and must depend on the squared comparison increments actually incurred, since a full-horizon bound would be too large for pairs that receive little weight. Lemma 1 states this pathwise comparison bound.

**Relation to subsequence regret.** This is a group-weighted, interval-wise internal-regret comparison, following the subsequence-regret viewpoint (Lee et al., 2022). On an interval $I$, the function $g(x_t)$ weights the rounds of interest, and the ordered pair $(a, b)$ specifies moving the probability mass $\pi_{t,a}$ from prediction $a$ to prediction $b$. The comparison accrues only on rounds when both predictions are active. The guarantee holds simultaneously over the prescribed group functions, ordered prediction pairs, and contiguous intervals, rather than arbitrary subsets of rounds selected in hindsight. We first specify the prediction values and loss, then give the prediction rule.

**Prediction values.** Each prediction is a probability vector $a \in \Delta_d$. The initial value is $q_0 = (1/n, \ldots, 1/n)$. We order distinct values by their first discovery, using this order to resolve ties. Rediscovering an available value preserves its comparison history. The number of distinct values is at most $A_{\max} = 1 + |\overline{\mathcal{G}}|T$, as shown in Appendix A.2.

Set $P_0 = \{q_0\}$ permanently and initialize $P_j(1) = \varnothing$ for $1 \le j \le K_{\max}$. At the start of round $t$ with current depth $K$, the active prediction values are

$$\mathcal{R}_t = P_0 \cup \bigcup_{j=1}^{K-1} P_j(t).$$

The discovery component maintains the stored prediction sets $P_j$ at positive levels. A value appears once in $\mathcal{R}_t$, even if it belongs to several of these sets. Write $\pi_{t,a} = \pi_t(a)$, with $\pi_{t,a} = 0$ when $a \notin \mathcal{R}_t$.

Once a value becomes unavailable, it never becomes available again. Apart from the permanent value $q_0$, each value is tested at every coarser positive level on first discovery; a value that prevents insertion at any such level remains there permanently. Thus its coarsest membership is fixed at discovery, and decreasing $K$ cannot restore availability. This justifies discarding comparison aggregates as soon as either prediction becomes unavailable.

**Fixed logarithmic loss.** Set $\beta = 1/(2T^2)$ and use $y'_t = (1-\beta)y_t + \beta q_0$ for comparisons and block means. The loss is

$$\ell(a, y') = -\sum_{i=1}^{n} y'_i \log a_i.$$

It compares predictions by how much probability they assign to the observed outcome.

**Prediction rule and comparison state.** Fix the learning rate at $\eta = \bar{D}/(8\kappa)$, where $\kappa = \bar{D} + 2$ and $\bar{D} = \log(n/\beta)$. For each group $g$ and ordered pair of distinct prediction values $a, b$, maintain an aggregate $B_t^{g,a,b}$. Initialize it to $B_t^{g,a,b} = t$ on the first round when both values are active. Retain it

while both values remain active. After observing $x_t$, but before the fresh draw or the outcome, form the directed edge weights on $\mathcal{R}_t$:

$$W_{t,a,b} = \sum_{g \in \overline{\mathcal{G}}} g(x_t) B_t^{g,a,b}, \qquad a \neq b, \qquad W_{t,a,a} = 0. \tag{6}$$

Choose the unique probability vector satisfying the stationarity equations

$$\pi_{t,b} \sum_{c \neq b} W_{t,b,c} = \sum_{a \neq b} \pi_{t,a} W_{t,a,b} \quad (b \in \mathcal{R}_t), \qquad \sum_{a \in \mathcal{R}_t} \pi_{t,a} = 1. \tag{7}$$

With one available prediction, give it probability one. With more than one, all off-diagonal weights are strictly positive because $\mathbf{1} \in \overline{\mathcal{G}}$ and all aggregates remain positive. Existence and uniqueness are proved in Appendix B.1. The learner samples and announces $p_t \sim \pi_t$. After observing $y_t$, set $c_t^{g,a,b} = g(x_t)\pi_{t,a}(\ell(a, y_t') - \ell(b, y_t'))/\bar{D}$ if both predictions are active, and set it to zero otherwise. Update the comparison state by

$$B_{t+1}^{g,a,b} = 1 + \left(1 + \eta c_t^{g,a,b}\right) B_t^{g,a,b}. \tag{8}$$

A positive increment increases the weight of switching from $a$ to $b$, while the added one includes a new interval starting on the next round. These updates affect only later predictions. A pair with no previous active round has $B_t = t$. For analysis, the recurrence also defines unstored aggregates, with zero increments while either value is inactive.

### A.2 Block weights and discovery counts

The discovery update forms blocks from weighted observations and processes completed blocks to update the stored prediction sets.

Scale 0 is the coarsest estimation level. Increasing the scale index halves the resolution $w_\ell$ and quadruples the block weight threshold, up to the finest stored level $K - 1$. Thus a finer level asks for a more accurate empirical average and waits for more weight before processing it. The scale index does not identify a predetermined region of the simplex.

**Why Hellinger distance?** The logarithmic-loss comparison is governed by Kullback–Leibler divergence, which smoothing bounds by a logarithmic factor times squared Hellinger distance. Combining Hellinger proximity to an available prediction with this loss comparison controls the KL divergence from the block mean to the source prediction. This divergence bounds their squared $\ell_1$ distance with a dimension-independent constant. The triangle inequality for Hellinger distance also lets us combine empirical-mean concentration with packing bounds on the unknown approximating sets to control the number of discovered predictions.

Fix one pair $(g, a)$ and temporarily ignore the hierarchies for all other pairs. On round $t$, this pair receives the outcome with weight $g(x_t)\pi_{t,a}$. Its scale-zero block receives weighted observations directly, whereas every positive-scale block receives only completed blocks from the preceding scale. The full algorithm runs this construction for every pair, with all pairs sharing the stored prediction sets $P_\ell$.

**Forming blocks.** For this pair, there is at most one open block at each scale. A block $E$ stores its weight $N_E$ and weighted outcome sum $V_E$, both initialized at zero. The empirical average is $z_E = V_E/N_E$ when $N_E > 0$. For analysis, write $a(E) = a$ for its source prediction and $s(E), e(E)$ for the endpoints of its interval $E$ when the block is nonempty. Zero-weight rounds inside this interval contribute zero to all block sums. The constituent times, completed child lists, and parent links are not stored. After $y_t$ is observed, the source receives the atom

$$\alpha_t^{g,a} = g(x_t)\pi_{t,a}, \qquad (N, V) \leftarrow (N + \alpha_t^{g,a}, V + \alpha_t^{g,a} y_t')$$

in its unique open scale-zero block whenever $\alpha_t^{g,a} > 0$. At scale $\ell$, the weight threshold is

$$m_\ell = 1024 \frac{nL}{w_\ell^2}.$$

At every scale $j$, a block closes immediately when $N_E \geq m_j$.

The thresholds $m_j = 4^j m_0$ imply that every completed positive-scale block has exactly four children. Indeed, a completed zero-level block has weight in $[m_0, m_0 + 1)$. Inductively, three level-$(j-1)$ children have total weight less than $3 \cdot 4^{j-1}(m_0+1) \leq 4^j m_0$, since the constants give $m_0 \geq 3$. Four children have weight at least $4^j m_0$ and less than $4^j(m_0+1)$. Thus the fourth child is exactly the first threshold crossing.

**Processing completed blocks.** Algorithm 1 gives the complete update, including its recursive subroutine. The source list $\mathcal{R}_t$ is fixed before the outcome. All zero-level observations are added before any closure is processed, and each recursive chain finishes before the next source–group pair is processed. Thus new predictions affect later insertion tests on the same round, but receive no observations or prediction probability until the next round. The sets $P_j$ are append-only and persist across changes in $K$.

**Bounding the number of discovered predictions.** For each fixed group $g$ and scale $j$, completed blocks across sources have total weight at most

$$\sum_{t=1}^{T} \sum_{a \in \mathcal{R}_t} g(x_t)\pi_{t,a} = \sum_{t=1}^{T} g(x_t) \leq T.$$

Indeed, at a fixed scale each weighted observation belongs to at most one completed block of its source, and observations are not replayed across depth intervals. Each completed scale-$j$ block has weight at least $m_j$, so there are at most $T/m_j$ such blocks per group. Every first discovery other than $q_0$ is caused by a positive-scale closure, and each closure introduces at most one new numerical prediction value. Since $m_j = 4^j m_0$, the total number of distinct predictions is therefore at most

$$1 + |\overline{\mathcal{G}}|T \sum_{j \geq 1} \frac{1}{m_j} = 1 + \frac{|\overline{\mathcal{G}}|T}{3m_0} \leq 1 + |\overline{\mathcal{G}}|T = A_{\max}.$$

Reusing a numerical value or inserting it into additional stored prediction sets does not create another first-discovery index.

# B Analysis at a fixed resolution

## B.1 Interval loss comparison

*Proof of Lemma 1.* The local residual bound uses the interval comparison to control accumulated logarithmic-loss differences. We derive it by using the stationarity equations to bound the total comparison weight, then bounding the product of updates for a fixed pair on an arbitrary available interval.

**Step 1. The stationarity equations control the total comparison weight.** By the definition of smoothing,

$$y'_{t,i} = (1-\beta)y_{t,i} + \frac{\beta}{n} \geq \frac{\beta}{n} =: \epsilon.$$

Every prediction $a \in \mathcal{R}_t$ is either $q_0$ or a weighted mean of smoothed outcomes, so each coordinate is at least $\epsilon$. Hence, by the fact that $y'_t \in \Delta_{n-1}$,

$$0 \leq \ell(a, y'_t) = \sum_i y'_{t,i} \log(1/a_i) \leq \log(1/\epsilon) \sum_i y'_{t,i} = \log(1/\epsilon) =: \bar{D}$$

and

$$|c_t^{g,a,b}| \leq g(x_t)\pi_{t,a} \frac{|\ell(a, y'_t) - \ell(b, y'_t)|}{\bar{D}} \leq g(x_t)\pi_{t,a} \leq 1.$$

Since $\eta < 1/8$, all factors $1 + \eta c_t$ are positive and all aggregates are at least one. The stationarity equations have a unique, strictly positive probability solution, as proved in Details of Step 1.

For analysis, extend the comparison aggregates to all $|\overline{\mathcal{G}}|A_{\max}^2$ group–prediction pairs, setting $c_t = 0$ whenever either prediction is unavailable or the two predictions coincide. Now by definition of $B_t^{g,a,b}$ and $c_t^{g,a,b}$, we have

$$\begin{aligned}\sum_{g,a,b} B^{g,a,b} c^{g,a,b} &= \sum_{a,b}\sum_{g} B^{g,a,b}\frac{g(x_t)\pi_a(\ell_a-\ell_b)}{\bar{D}} \\ &= \frac{1}{\bar{D}}\sum_{a,b}\pi_a(\ell_a-\ell_b)\underbrace{\sum_g g(x_t)B^{g,a,b}}_{=W_{a,b}} \\ &= \frac{1}{\bar{D}}\sum_{a,b}\pi_a W_{a,b}(\ell_a-\ell_b).\end{aligned}$$

Since $W_{a,a} = 0$ and by (7):

$$\begin{aligned}\sum_{a,b}\pi_a W_{a,b}(\ell_a-\ell_b) &= \sum_a \ell_a\pi_a\sum_b W_{a,b} - \sum_b \ell_b\sum_a \pi_a W_{a,b} \\ &= \sum_b \ell_b\left(\pi_b\sum_{c\neq b} W_{b,c} - \sum_{a\neq b}\pi_a W_{a,b}\right) = 0,\end{aligned}$$

we have $\sum_{g,a,b} B^{g,a,b}c^{g,a,b} = 0$. Thus by (8), we obtain

$$\begin{aligned}\sum_{g,a,b} B_{t+1}^{g,a,b} &= |\overline{\mathcal{G}}|A_{\max}^2 + \sum_{g,a,b} B_t^{g,a,b} + \eta\underbrace{\sum_{g,a,b} B_t^{g,a,b}c_t^{g,a,b}}_{=0} \\ &= |\overline{\mathcal{G}}|A_{\max}^2 + \sum_{g,a,b} B_t^{g,a,b}.\end{aligned}$$

Since $B_1^{g,a,b} = 1$ for every $g, a, b$, we have

$$\sum_{g,a,b} B_t^{g,a,b} = |\overline{\mathcal{G}}|A_{\max}^2 + (t-1)|\overline{\mathcal{G}}|A_{\max}^2 = |\overline{\mathcal{G}}|A_{\max}^2 t. \tag{9}$$

**Step 2. Complete the interval loss-comparison bound.** Fix $g, a, b$ and a contiguous interval $I = [u, v]$ on which both predictions are available. Write $X_t = \ell(a, y_t') - \ell(b, y_t')$. If $a = b$, then $X_t = 0$ and the claim is immediate. Otherwise iterating (8) gives

$$B_{v+1}^{g,a,b} \geq B_u^{g,a,b}\prod_{t=u}^{v}(1+\eta c_t^{g,a,b}) \geq \prod_{t=u}^{v}(1+\eta c_t^{g,a,b}),$$

where the last inequality is due to $B_u^{g,a,b} \geq 1$. Thus, by (9), we obtain

$$\prod_{t=u}^{v}(1+\eta c_t^{g,a,b}) \leq B_{v+1}^{g,a,b} \leq \sum_{g',a',b'} B_{v+1}^{g',a',b'} = |\overline{\mathcal{G}}|A_{\max}^2(v+1) \leq |\overline{\mathcal{G}}|A_{\max}^2(T+1).$$

Taking logarithms gives

$$\sum_{t\in I}\log(1+\eta c_t^{g,a,b}) \leq \log\big(|\overline{\mathcal{G}}|A_{\max}^2(T+1)\big).$$

Since $|\eta c_t^{g,a,b}| \leq \eta < 1/8 < 1/2$, applying $\log(1+x) \geq x - x^2$ to each summand yields

$$\eta\sum_{t\in I} c_t^{g,a,b} - \eta^2\sum_{t\in I}(c_t^{g,a,b})^2 \leq \log\big(|\overline{\mathcal{G}}|A_{\max}^2(T+1)\big).$$

Substituting $c_t^{g,a,b} = \alpha_t X_t/\bar{D}$, rearranging, and multiplying by $\bar{D}/\eta$, we obtain

$$\sum_{t\in I}\alpha_t X_t \leq \frac{\eta}{\bar{D}}\sum_{t\in I}\alpha_t^2 X_t^2 + \frac{\bar{D}}{\eta}\log\big(|\overline{\mathcal{G}}|A_{\max}^2(T+1)\big).$$

Finally, $0 \le \alpha_t = g(x_t)\pi_{t,a} \le 1$ implies $\alpha_t^2 \le \alpha_t$. Using $\eta = \bar{D}/(8\kappa)$ and $\log(|\overline{\mathcal{G}}|A_{\max}^2(T+1)) \le L$, we conclude that

$$\sum_{t\in I} \alpha_t X_t \le \frac{1}{8\kappa} \sum_{t\in I} \alpha_t X_t^2 + 8\kappa L.$$

This proves (2) on every path for all such intervals, including intervals selected after observation.

**Details of Step 1: Existence and uniqueness of the prediction distribution.** If there is one available prediction, its probability is one. Otherwise $B_t > 0$, and the constant group makes every off-diagonal $W_{t,a,b}$ positive. For the proof only, set $Z_t = \max_a \sum_{b\neq a} W_{t,a,b} > 0$, $M_t(a,b) = W_{t,a,b}/Z_t$ for $a \neq b$, and $M_t(a,a) = 1 - \sum_{b\neq a} M_t(a,b)$. This is a finite row-stochastic matrix. For any probability row vector $v$, take the averages $\bar{v}_m = \frac{1}{m}\sum_{k=0}^{m-1} vM_t^k$. They have a convergent subsequence, and $\bar{v}_m M_t - \bar{v}_m \to 0$. The limit is therefore a stationary probability vector. If $M_t x = x$, a maximal coordinate of $x$ is a convex combination with positive coefficients on every other coordinate, so all coordinates of $x$ are equal. Thus the right nullspace of $I - M_t$, and hence its left nullspace, has dimension one. The stationary probability vector is unique. If one of its coordinates were zero, stationarity and the positive off-diagonal entries would force all the other coordinates to be zero, contradicting normalization. Thus every coordinate is strictly positive. Eliminating $Z_t$ from its equations gives exactly (7). The algorithm need not construct $Z_t$, $M_t$, or a communicating class.

□

### B.2 Local comparison

We first establish the distance inequalities used in the available-comparator bound and the subsequent affine packing analysis.

**Lemma 4** (KL–Hellinger comparison)**.** *Let $p, q \in \Delta_{n-1}$ and $0 < \epsilon \le 1$, with $q_i \ge \epsilon$ for every $i$. Then*

$$H(p,q)^2 \le D(p\|q) \le \big(2 + \log(1/\epsilon)\big) H(p,q)^2.$$

*Proof.* For $p_i > 0$, the inequality $\log x \le x - 1$, applied to $x = \sqrt{q_i/p_i}$, gives $p_i \log(p_i/q_i) \ge 2p_i - 2\sqrt{p_i q_i}$. It also holds at $p_i = 0$ by continuity. Summing gives $D(p\|q) \ge 2 - 2\sum_i \sqrt{p_i q_i} = H(p,q)^2$.

For the upper bound, $\sum_i (p_i - q_i) = 0$ gives

$$D(p\|q) = \sum_i q_i f\big(\sqrt{p_i/q_i}\big), \qquad f(x) = 2x^2 \log x - x^2 + 1.$$

Here $f(1) = 0$, $f'(x) = 4x \log x$, and $f(0) = 1$ by continuity. For $0 < x \le 1$, use $\log s \ge 1 - 1/s$; for $x \ge 1$, use $s \log s \le (s-1)(1 + \log x)$ on $1 \le s \le x$. Integrating yields

$$f(x) = -4\int_x^1 s \log s \, ds \le 2(1-x)^2 \qquad (0 \le x \le 1),$$
$$f(x) = 4\int_1^x s \log s \, ds \le 2(1 + \log x)(x-1)^2 \quad (x \ge 1).$$

For $x = \sqrt{p_i/q_i} \ge 1$, we have $2\log x \le \log(1/\epsilon)$. Thus in both cases $f(x) \le (2 + \log(1/\epsilon))(x-1)^2$. Multiply by $q_i$ and sum to obtain the upper bound. □

*Proof of Lemma 2.* Fix a group $g$ and a contiguous interval $I$ on which predictions $a, q$ are both available. Let $\alpha_t = \pi_{t,a} g(x_t)$, $N = \sum_{t\in I} \alpha_t > 0$, and $z = N^{-1} \sum_{t\in I} \alpha_t y'_t$. We convert the interval loss-comparison bound of Lemma 1 into a bound on $D(z\|a)$ in terms of $D(z\|q)$. To do so, we first control the squared loss differences by these two divergences, then substitute this estimate into the comparison bound and rearrange. The residual and Hellinger bounds follow from inequalities relating KL divergence to $\ell_1$ and Hellinger distances.

**Step 1. Bound the squared loss difference between two predictions.** For the two predictions $a, q$, we define

$$X_t = \ell(a, y'_t) - \ell(q, y'_t) = \sum_i y'_{t,i} \log(q_i/a_i).$$

Jensen inequality therefore gives

$$\sum_{t\in I}\alpha_t X_t^2 \le \sum_{t\in I}\alpha_t \sum_i y'_{t,i}\log^2(q_i/a_i)$$
$$= N\sum_i z_i \log^2(a_i/q_i),$$

where the equality uses the definition $z_i = \sum_{t\in I}\alpha_t y'_{t,i}/N$. To bound the right-hand side, it therefore suffices to prove

$$\sum_i z_i \log^2(a_i/q_i) \le 2\kappa[D(z\|a) + D(z\|q)].$$

To see which coordinatewise inequality is needed, first expand the two divergences. Since $a, q, z$ are probability vectors, $\sum_i(a_i + q_i - 2z_i) = 0$, and hence

$$D(z\|a) + D(z\|q) = \sum_i \left[a_i + q_i - 2z_i + z_i \log\frac{z_i^2}{a_i q_i}\right],$$

where the logarithmic term is interpreted as zero when $z_i = 0$. For $z_i > 0$, let $u = a_i/z_i$ and $v = q_i/z_i$. The corresponding summand is $z_i[u + v - 2 - \log(uv)]$, whereas the summand on the left of the desired bound is $z_i \log^2(u/v)$. It therefore suffices to prove

$$\log^2(u/v) \le 2\kappa[u + v - 2 - \log(uv)].$$

The left-hand side depends only on the ratio $u/v$. Let $t = \log(u/v)$. We use the following decomposition to obtain a lower bound depending only on $t$:

$$u + v - 2 - \log(uv) = 2\left(\frac{u+v}{2} - 1 - \log\frac{u+v}{2}\right) + 2\log\frac{u+v}{2\sqrt{uv}}$$
$$\ge 2\log\cosh(t/2).$$

Here the inequality follows from $x - 1 - \log x \ge 0$ for $x > 0$, applied with $x = (u + v)/2$, and $\frac{u+v}{2\sqrt{uv}} = \cosh(t/2)$. We next bound $2\log\cosh(t/2)$ below in terms of $t^2$. For $x \ge 0$, the inequality $e^{2x} \ge 1 + 2x$ implies

$$\tanh x = \frac{e^{2x} - 1}{e^{2x} + 1} \ge \frac{x}{1 + x}.$$

Integrating, we obtain

$$\log\cosh x = \int_0^x \tanh s\, ds \ge \int_0^x \frac{s}{1+s}\, ds \ge \int_0^x \frac{s}{1+x}\, ds = \frac{x^2}{2(1+x)}.$$

Taking $x = |t|/2$ and using $\cosh(t/2) = \cosh(|t|/2)$, we get

$$2\log\cosh(t/2) \ge \frac{t^2}{4 + 2|t|}.$$

Because $t = \log(a_i/q_i)$ and $a_i, q_i \in [\epsilon, 1]$, we have $|t| \le \log(1/\epsilon) = \bar{D}$. Thus $4 + 2|t| \le 4 + 2\bar{D} = 2\kappa$, giving

$$\frac{a_i}{z_i} + \frac{q_i}{z_i} - 2 - \log\left(\frac{a_i q_i}{z_i^2}\right) \ge \frac{t^2}{4 + 2|t|} \ge \frac{t^2}{2\kappa} = \frac{\log^2(a_i/q_i)}{2\kappa}.$$

Multiplying by $z_i$, we obtain the coordinatewise bound

$$z_i \log^2(a_i/q_i) \le 2\kappa\left[a_i + q_i - 2z_i + z_i \log\frac{z_i^2}{a_i q_i}\right].$$

When $z_i = 0$, the left-hand side is zero and the right-hand side is $2\kappa(a_i + q_i) \ge 0$, so the same bound holds. Summing over all coordinates and using the divergence expansion above yields

$$\sum_i z_i \log^2(a_i/q_i) \le 2\kappa[D(z\|a) + D(z\|q)].$$

Substituting into the initial Jensen bound gives

$$\sum_{t\in I} \alpha_t X_t^2 \le 2\kappa N[D(z\|a) + D(z\|q)].$$

**Step 2. Convert the interval comparison into a divergence bound.** Both predictions are available on the whole interval, so Lemma 1 applies. Using the definition of $z$, this comparison bound, and then the squared loss bound from Step 1, we obtain

$$\begin{aligned} N[D(z\|a) - D(z\|q)] &= N\sum_i z_i \log(q_i/a_i) = \sum_{t\in I} \alpha_t X_t \\ &\le \frac{1}{8\kappa}\sum_{t\in I} \alpha_t X_t^2 + 8\kappa L \\ &\le \tfrac{1}{4} N[D(z\|a) + D(z\|q)] + 8\kappa L. \end{aligned}$$

Since both divergences are nonnegative, replacing $1/4$ by $1/2$ for simpler constants preserves the upper bound:

$$N[D(z\|a) - D(z\|q)] \le \tfrac{1}{2} N[D(z\|a) + D(z\|q)] + 8\kappa L.$$

Rearranging the last inequality gives

$$D(z\|a) \le 3D(z\|q) + \frac{16\kappa L}{N}.$$

**Step 3. Complete the residual and Hellinger bounds.** We apply Lemma 4 to the divergence bound from Step 2, using $2 + \log(1/\epsilon) = \kappa$.

- *Derive the Hellinger bound* (3). Since every coordinate of $a$ and $q$ is at least $\epsilon$, by Lemma 4, we have

$$\begin{aligned} H(z,a)^2 &\le D(z\|a) \\ &\le 3D(z\|q) + \frac{16\kappa L}{N} \\ &\le 3\kappa H(z,q)^2 + \frac{16\kappa L}{N}. \end{aligned}$$

  Taking square roots and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for $x, y \ge 0$ gives

$$H(z,a) \le \sqrt{3\kappa} H(z,q) + 4\sqrt{\kappa L/N},$$

  which proves (3).

- *Derive the residual bound* (4). Fix a scale $j$ such that $N \ge m_j$ and $H(z,q) \le 2w_j$. By Pinsker's inequality, the bound from Step 2, and Lemma 4,

$$\begin{aligned} \|z-a\|_1^2 &\le 2D(z\|a) \\ &\le 6D(z\|q) + \frac{32\kappa L}{N} \\ &\le 6\kappa H(z,q)^2 + \frac{32\kappa L}{N}. \end{aligned}$$

  Taking square roots gives

$$\|z-a\|_1 \le \sqrt{6\kappa} H(z,q) + \sqrt{32\kappa L/N}.$$

  If $H(z,q) \le 2w_j$ and $N \ge m_j = 1024nL/w_j^2$, then

$$\begin{aligned} \|z-a\|_1 &\le 2\sqrt{6\kappa}\, w_j + \sqrt{\frac{32\kappa L w_j^2}{1024nL}} \\ &= \left(2\sqrt{6} + \sqrt{\frac{32}{1024n}}\right)\sqrt{\kappa}\, w_j \\ &\le 6\sqrt{\kappa}\, w_j. \end{aligned}$$

  This proves (4). For a completed scale-zero block, the same argument applies with $q = q_0$, since $N \ge m_0$ and $H(z, q_0) \le \sqrt{2} \le 2w_0$.

□

### B.3 A COMMON HIGH-PROBABILITY EVENT

Recall $\mathcal{F}_t^-$ is the filtration before prediction is sampled from $\pi_t$ as defined in Section 2.1. Let $\mathcal{F}_t$ contain the prediction, outcome, and all updates through round $t$. Then $\mathcal{F}_t^- \subseteq \mathcal{F}_t$, and $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}^-$ for $t < T$.

Let $\mu'_t = \mathbb{E}[y'_t \mid \mathcal{F}_t^-] = (1-\beta)\mu_t + \beta q_0$. For a group $g \in \overline{\mathcal{G}}$, discovered prediction value $a$, and contiguous interval $I \subseteq [T]$, let $\alpha_t = g(x_t)\pi_{t,a}$ and $N = \sum_{t\in I} \alpha_t > 0$, and write

$$z = \frac{1}{N}\sum_{t\in I} \alpha_t y'_t, \qquad \bar{\mu}' = \frac{1}{N}\sum_{t\in I} \alpha_t \mu'_t.$$

**Lemma 5** (Outcome noise and prediction sampling)**.** *There are events $\mathcal{E}_{\mathrm{conc}}$ and $\mathcal{E}_{\mathrm{samp}}$, each with failure probability at most $\delta/2$, on which the following bounds hold simultaneously on the actual algorithmic run.*

*(i) On $\mathcal{E}_{\mathrm{conc}}$, for every $g, a, I$ with $N > 0$,*

$$H(z, \bar{\mu}') \le 2\sqrt{nL/N}.$$

*In particular, a completed scale-$j$ block $E$ satisfies*

$$H(z_E, \bar{\mu}'_E) \le w_j/16, \tag{10}$$

*where $\bar{\mu}'_E$ is the average above with the weights and observation interval of $E$.*

*(ii) On $\mathcal{E}_{\mathrm{samp}}$, for every contiguous interval $I$, every group $g \in \overline{\mathcal{G}}$, and every set $Q$ of $m \ge 1$ distinct prediction values discovered during the run,*

$$\sum_{a\in Q} \left\| \sum_{t\in I} g(x_t)(\mathbf{1}\{p_t = a\} - \pi_{t,a})(y'_t - a) \right\|_1 \le 4\sqrt{nmLT}. \tag{11}$$

*Thus $\mathcal{E}_{\mathrm{global}} := \mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{samp}}$ has probability at least $1 - \delta$.*

We use the following two-sided consequence of the scalar Freedman inequality (Freedman, 1975; Tropp, 2011).

**Lemma 6** (Freedman's inequality)**.** *For a positive integer $q$, let $(\mathcal{H}_t)_{t=0}^q$ be a filtration and let $S_t = \sum_{v=1}^t \xi_v$, where $\xi_v$ is $\mathcal{H}_v$-measurable, $\mathbb{E}[\xi_v \mid \mathcal{H}_{v-1}] = 0$, and $|\xi_v| \le 1$ almost surely. Define $W_t = \sum_{v=1}^t \mathbb{E}[\xi_v^2 \mid \mathcal{H}_{v-1}]$. For every $b, x > 0$,*

$$\Pr\left(\exists t \in [q] :\ |S_t| \ge \sqrt{2bx} + \frac{2x}{3},\ W_t \le b\right) \le 2e^{-x}.$$

*Proof.* The scalar inequality in Tropp (2011, Theorem 1.1), applied to $S_t$ and $-S_t$, gives

$$\Pr(\exists t \in [q] :\ |S_t| \ge a,\ W_t \le b) \le 2\exp\left(-\frac{a^2}{2(b + a/3)}\right).$$

For $a = \sqrt{2bx} + 2x/3$, we have $a^2 - 2x(b + a/3) = (2x/3)\sqrt{2bx} \ge 0$, proving the claim. □

We also need a one-sided, conditional exponential-moment version of the Hoeffding–Azuma maximal inequality (see Tropp (2024, Theorem 26.9) for the bounded-increment formulation). We state and prove the version needed here to make its assumptions and constants explicit.

**Lemma 7** (Maximal Azuma–Hoeffding inequality)**.** *For a positive integer $q$, let $(S_t)_{t=0}^q$ be a martingale with respect to a filtration $(\mathcal{H}_t)_{t=0}^q$, with $S_0 = 0$. Suppose that, for a deterministic $c > 0$, its increments $\xi_t = S_t - S_{t-1}$ satisfy*

$$\mathbb{E}[e^{\lambda \xi_t} \mid \mathcal{H}_{t-1}] \le e^{\lambda^2 c^2/8} \qquad (\lambda \in \mathbb{R},\ t \in [q]).$$

*Then, for every $x > 0$,*

$$\Pr\left(\max_{1\le t\le q} S_t \ge c\sqrt{\frac{qx}{2}}\right) \le e^{-x}.$$

*The exponential-moment assumption holds whenever each increment has conditional mean zero and, given $\mathcal{H}_{t-1}$, lies in an interval of length at most $c$.*

*Proof.* Condition on $\mathcal{H}_{t-1}$ and let $f(\lambda) = \log \mathbb{E}[e^{\lambda \xi_t} \mid \mathcal{H}_{t-1}]$. If $\xi_t$ lies in an interval of length $c$, then $f''(\lambda)$ is its variance under exponential reweighting, at most $c^2/4$: squared distance from the interval's midpoint is at most $c^2/4$, and variance is no larger. Since $f(0) = f'(0) = 0$, integration gives $f(\lambda) \le \lambda^2 c^2/8$ for every real $\lambda$, proving the stated sufficient condition. Under the exponential-moment assumption, for each $\lambda > 0$,

$$Z_t = \exp\left\{ \lambda S_t - \frac{t\lambda^2 c^2}{8} \right\}, \qquad t = 0, \dots, q,$$

is a nonnegative supermartingale with $Z_0 = 1$. Fix $a > 0$ and define the bounded stopping time

$$\tau = \min\big(\{t \in [q] : S_t \ge a\} \cup \{q\}\big).$$

On $\{\max_{1 \le t \le q} S_t \ge a\}$, we have $S_\tau \ge a$ and $\tau \le q$. Consequently,

$$1 \ge \mathbb{E}[Z_\tau] \ge \exp\left\{ \lambda a - \frac{q\lambda^2 c^2}{8} \right\} \Pr\left( \max_{1 \le t \le q} S_t \ge a \right).$$

Taking $\lambda = 4a/(qc^2)$ gives a probability bound of $\exp\{-2a^2/(qc^2)\}$. Substituting $a = c\sqrt{qx/2}$ proves the claim. □

*Proof of Lemma 5.* Appendix A.2 bounds the number of distinct predictions by $A_{\max}$ on every run. We prove the two concentration bounds separately. For the shifted martingales in Steps 1 and 2, define $\mathcal{F}^-_{T+1} := \mathcal{F}_T$.

The probability proofs use permanent first-discovery indices in $[A_{\max}]$, one for each distinct numerical prediction. On active rounds the value and its mixture weight are measurable before prediction sampling.

**Step 1. Outcome-noise event $\mathcal{E}_{\mathrm{conc}}$.**

(a) *Define and normalize the weighted sums.* For each group $g \in \overline{\mathcal{G}}$, index $k \in [A_{\max}]$, and round $u \in [T]$, consider the following weighted sums. Writing $a_k$ for the value at index $k$, use predictable weights $\alpha_v = \mathbf{1}\{v \ge u\}\pi_{v,a_k} g(x_v)$, which is zero before that value is discovered or while it is unavailable. For each endpoint $t \in \{0, \dots, T\}$, define

$$N_t = \sum_{v=1}^{t} \alpha_v, \qquad V_{t,i} = \sum_{v=1}^{t} \alpha_v y'_{v,i}, \qquad U_{t,i} = \sum_{v=1}^{t} \alpha_v \mathbb{E}[y'_{v,i} \mid \mathcal{F}^-_v] \quad (i \in [n]).$$

Write $V_t = (V_{t,i})_{i=1}^n$ and $U_t = (U_{t,i})_{i=1}^n$. Since $\alpha_v \ge 0$ and $y'_v \in \Delta_{n-1}$, both vectors have nonnegative coordinates and

$$\sum_{i=1}^{n} V_{t,i} = \sum_{v=1}^{t} \alpha_v \sum_{i=1}^{n} y'_{v,i} = \sum_{v=1}^{t} \alpha_v = N_t,$$
$$\sum_{i=1}^{n} U_{t,i} = \sum_{v=1}^{t} \alpha_v \mathbb{E}\left[ \sum_{i=1}^{n} y'_{v,i} \,\middle|\, \mathcal{F}^-_v \right] = \sum_{v=1}^{t} \alpha_v = N_t.$$

Thus $V_t/N_t, U_t/N_t \in \Delta_{n-1}$ whenever $N_t > 0$.

(b) *Apply Freedman to the weighted coordinate sums.* For each coordinate $i$, the increments $\xi_v := \alpha_v(y'_{v,i} - \mathbb{E}[y'_{v,i} \mid \mathcal{F}^-_v])$ are conditionally centered given $\mathcal{F}^-_v$, i.e., $\mathbb{E}[\xi_v \mid \mathcal{F}^-_v] = 0$ and have absolute value bounded by 1:

$$|\xi_v| = \alpha_v \left| y'_{v,i} - \mathbb{E}[y'_{v,i} \mid \mathcal{F}^-_v] \right| \le \alpha_v \le 1.$$

To use the notation of Lemma 6, define

$$S_t := V_{t,i} - U_{t,i} = \sum_{v=1}^{t} \xi_v, \qquad t = 0, \dots, T.$$

Then $S_0 = 0$, and $S_t$ is measurable with respect to $\mathcal{H}_t := \mathcal{F}^-_{t+1}$. For $t \in [T]$,

$$\mathbb{E}[S_t - S_{t-1} \mid \mathcal{H}_{t-1}] = \mathbb{E}[\xi_t \mid \mathcal{F}^-_t] = 0.$$

Thus $(S_t)_{t=0}^T$ is a martingale with respect to $(\mathcal{H}_t)_{t=0}^T$. Its cumulative conditional variance after round $t$ satisfies

$$W_{t,i} := \sum_{v=1}^{t} \mathbb{E}[\xi_v^2 \mid \mathcal{H}_{v-1}] = \sum_{v=1}^{t} \mathbb{E}\Big[\alpha_v^2 \big(y'_{v,i} - \mathbb{E}[y'_{v,i} \mid \mathcal{F}_v^-]\big)^2 \,\Big|\, \mathcal{F}_v^-\Big]$$
$$\le \sum_{v=1}^{t} \alpha_v^2 \mathbb{E}[(y'_{v,i})^2 \mid \mathcal{F}_v^-] \le \sum_{v=1}^{t} \alpha_v \mathbb{E}[y'_{v,i} \mid \mathcal{F}_v^-] = U_{t,i}.$$

Thus, for every $b, x > 0$, Lemma 6 gives

$$\Pr\left(\exists t \in [T] : |V_{t,i} - U_{t,i}| \ge \sqrt{2bx} + \frac{2x}{3},\ U_{t,i} \le b\right) \le 2e^{-x}.$$

(c) *Obtain coordinate bounds uniformly over all intervals.* We apply geometric peeling to the cumulative conditional mean $U_{t,i} \in [0, T]$, using the intervals

$$[0, 1], \qquad \big((4/3)^{h-1}, (4/3)^h\big], \quad h = 1, \dots, J-1,$$

where $J = 1 + \lceil \log_{4/3} T \rceil$. For each interval, we apply Freedman's inequality with its deterministic upper endpoint $b_h = (4/3)^h$, and then take a union bound over the intervals. Then, let

$$L_0 = \log\left(\frac{4n|\overline{\mathcal{G}}|A_{\max}TJ}{\delta}\right).$$

Note that $L_0 > 1$. Apply the preceding Freedman bound with $b = b_h$ and $x = L_0$, and take a union bound over all $n$ coordinates, $|\overline{\mathcal{G}}|$ groups, $A_{\max}$ prediction indices, $T$ starts, and $J$ variance bounds. The total failure probability is at most

$$2n|\overline{\mathcal{G}}|A_{\max}TJe^{-L_0} = \delta/2.$$

Let $\mathcal{E}_{\text{conc}}$ be the simultaneous success event. On this event, for every group, prediction index, start, endpoint $t$, coordinate $i$, and $h$, if $U_{t,i} \le b_h$, then

$$|V_{t,i} - U_{t,i}| \le \sqrt{2b_hL_0} + \frac{2L_0}{3}.$$

(d) *Convert coordinate bounds to a Hellinger bound.* Fix an endpoint $t$ with $N_t > 0$ and suppress its subscript in $N_t, V_t, U_t$ and their coordinates. The empirical mean and weighted predictable mean are $z = V/N$ and $\bar{\mu}' = U/N$. Our next goal is to prove

$$H(z, \bar{\mu}')^2 = \frac{1}{N}\sum_{i=1}^{n}(\sqrt{V_i} - \sqrt{U_i})^2 \le \frac{4nL_0}{N}.$$

It suffices to prove $|\sqrt{V_i} - \sqrt{U_i}| \le 2\sqrt{L_0}$ for every $i$.

If $U_i \le 1$, use $b_0 = 1$. Since $L_0 > 1$,

$$0 \le V_i \le 1 + \sqrt{2L_0} + \frac{2L_0}{3} < 4L_0,$$

so both square roots lie in $[0, 2\sqrt{L_0}]$.

If $U_i > 1$, choose the smallest $h$ with $b_h \ge U_i$. Such an $h$ exists because $U_i \le T$, and $U_i \le b_h < (4/3)U_i$. Hence

$$|V_i - U_i| \le \sqrt{\frac{8}{3}U_iL_0} + \frac{2L_0}{3}.$$

For the upper deviation, this implies

$$V_i \le U_i + \sqrt{\frac{8}{3}U_iL_0} + \frac{2L_0}{3} \le (\sqrt{U_i} + 2\sqrt{L_0})^2,$$

and therefore $\sqrt{V_i} - \sqrt{U_i} \le 2\sqrt{L_0}$. For the lower deviation, consider $V_i < U_i$. If $U_i \le 4L_0$, then $\sqrt{U_i} - \sqrt{V_i} \le \sqrt{U_i} \le 2\sqrt{L_0}$. Otherwise,

$$\begin{aligned}\sqrt{U_i} - \sqrt{V_i} &= \frac{U_i - V_i}{\sqrt{U_i} + \sqrt{V_i}} \\ &\le \frac{\sqrt{\frac{8}{3}U_i L_0} + \frac{2L_0}{3}}{\sqrt{U_i}} \\ &\le \sqrt{\frac{8}{3}L_0} + \frac{2L_0}{3\sqrt{U_i}} \\ &\le \left(\sqrt{\frac{8}{3}} + \frac{1}{3}\right)\sqrt{L_0} \\ &< 2\sqrt{L_0}.\end{aligned}$$

Combining the two directions and summing over coordinates gives

$$H(z, \bar{\mu}')^2 = \frac{1}{N}\sum_{i=1}^{n}(\sqrt{V_i} - \sqrt{U_i})^2 \le \frac{4nL_0}{N}.$$

To apply this bound at the block weight threshold $m_\ell = 1024nL/w_\ell^2$, we check that $L_0 \le L$. Since $(4/3)^3 > 2$, we have $J \le 3(1 + \lceil \log_2 T \rceil)$. The definition of $L$ in Appendix F therefore gives

$$L_0 \le \log\left(\frac{12n|\overline{\mathcal{G}}|A_{\max}T(1 + \lceil \log_2 T \rceil)}{\delta}\right) \le L.$$

Consequently, for $N \ge m_\ell$,

$$H(z, \bar{\mu}') \le 2\sqrt{\frac{nL_0}{N}} \le \frac{w_\ell}{16}\sqrt{\frac{L_0}{L}} \le \frac{w_\ell}{16}.$$

All starts and endpoints were covered before choosing a block, so the bound includes blocks ending at a depth change or the horizon.

**Step 2. Prediction-sampling event $\mathcal{E}_{\text{samp}}$.**

(a) *Express sampling residuals as centered scalar increments.* Fix a deterministic start $u \in [T]$, group $g \in \overline{\mathcal{G}}$, subset $Q \subseteq [A_{\max}]$ of size $m$, and signs $\sigma_i \in \{-1, 1\}^n$ for $i \in Q$. For this proof, let $a_i$ denote the prediction with index $i$. Set its summand to zero until it is discovered and whenever it is unavailable. For each $t \in [T]$, the scalar increment

$$\xi_t = g(x_t)\sum_{i \in Q}\left(\mathbf{1}\{p_t = a_i\} - \pi_{t,a_i}\right)\sigma_i^{\mathsf{T}}(y'_t - a_i)$$

is conditionally centered given $\mathcal{F}_t^-$ and $y'_t$.

(b) *Construct the martingale and its exponential-moment bound.* Given $\mathcal{F}_t^-$ and $y'_t$, define

$$Y_t = g(x_t)\sum_{i \in Q}\mathbf{1}\{p_t = a_i\}\sigma_i^{\mathsf{T}}(y'_t - a_i).$$

At most one summand is nonzero, and $\|y'_t - a_i\|_1 \le 2$, so $Y_t \in [-2, 2]$. Private-draw independence gives

$$\xi_t = Y_t - \mathbb{E}[Y_t \mid \mathcal{F}_t^-, y'_t].$$

By the conditional Hoeffding bound established in the proof of Lemma 7, the value-range length $c = 4$ gives

$$\mathbb{E}[e^{\lambda\xi_t} \mid \mathcal{F}_t^-, y'_t] \le e^{2\lambda^2}.$$

Taking conditional expectations again yields

$$\mathbb{E}[\xi_t \mid \mathcal{F}_t^-] = 0, \qquad \mathbb{E}[e^{\lambda\xi_t} \mid \mathcal{F}_t^-] \le e^{2\lambda^2}.$$

As in Step 1, use the pre-sampling filtration with a one-round shift and define

$$S_t = \sum_{v=1}^{t} \mathbf{1}\{v \geq u\}\xi_v, \qquad t = 0, \ldots, T.$$

Define $\mathcal{H}_t := \mathcal{F}_{t+1}^-$ for $t = 0, \ldots, T$. Then $S_0 = 0$, $S_t$ is $\mathcal{H}_t$-measurable, and, for $t \in [T]$,

$$\mathbb{E}[S_t - S_{t-1} \mid \mathcal{H}_{t-1}] = 0, \qquad \mathbb{E}[e^{\lambda(S_t - S_{t-1})} \mid \mathcal{H}_{t-1}] \leq e^{2\lambda^2}.$$

Thus $(S_t)_{t=0}^T$ is a martingale with respect to $(\mathcal{H}_t)_{t=0}^T$, and $S_t = \sum_{v=u}^t \xi_v$ for $t \geq u$.

(c) *Obtain a simultaneous bound over intervals, groups, subsets, and signs.* Set $A_m = nm \log 2 + m \log A_{\max} + \log(2|\overline{\mathcal{G}}|TA_{\max}/\delta)$. Apply Lemma 7 to this martingale with $q = T$, $c = 4$, and $x = A_m$. This gives

$$\Pr\left( \max_{u \leq t \leq T} \sum_{v=u}^{t} \xi_v \geq \sqrt{8TA_m} \right) \leq e^{-A_m}.$$

For fixed $m$, there are at most $\binom{A_{\max}}{m} \leq A_{\max}^m$ index subsets, $2^{nm}$ sign choices, $T$ starts, and $|\overline{\mathcal{G}}|$ groups. The union bound over these choices gives a total failure probability at most

$$|\overline{\mathcal{G}}|TA_{\max}^m 2^{nm} e^{-A_m} = \frac{\delta}{2A_{\max}}.$$

Summing over $m = 1, \ldots, A_{\max}$ gives total failure probability at most $\delta/2$.

(d) *Recover the $\ell_1$ sampling-residual bound.* On the simultaneous success event from part (c), maximize over the sign vector for each prediction and use $A_m \leq 2nmL$. For every interval $[u, t] \subseteq [T]$, group $g \in \overline{\mathcal{G}}$, and set $Q$ of $m \geq 1$ distinct prediction values discovered during the run, this gives

$$\sum_{a \in Q} \left\| \sum_{v=u}^{t} g(x_v)(\mathbf{1}\{p_v = a\} - \pi_{v,a})(y'_v - a) \right\|_1 \leq \sqrt{8TA_m} \leq 4\sqrt{nmLT}.$$

Denote this event by $\mathcal{E}_{\text{samp}}$, which has probability at least $1 - \delta/2$.

By a union bound, $\mathcal{E}_{\text{global}} = \mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{samp}}$ has probability at least $1 - \delta$.

□

### B.4 Affine-dimensional Hellinger geometry

For the remaining packing and residual analysis, fix a realization in $\mathcal{E}_{\text{global}}$, whose definition is independent of the choice of $s, S$. Taking coordinatewise square roots turns Hellinger distance into Euclidean distance, but the image of a set of affine dimension $s$ need not lie in an $s$-dimensional affine subspace. We instead parameterize the set in its own affine span and combine these coordinates with the gradient of negative entropy. The resulting map keeps points in a bounded union of Euclidean balls and increases their separation, so a Euclidean packing bound applies in dimension $s$.

**Definition 2** (Hellinger packing numbers)**.** *Let $S \subseteq \Delta_{n-1}$, $v > 0$, and $R > 0$. Define*

$$\mathsf{G}_H(S; v) := \sup\{|Q| : \ Q \subseteq S, \ H(q, q') > v/2 \text{ for all distinct } q, q' \in Q\} \qquad \textit{(global packing)},$$

*and*

$$\mathsf{D}_H(S; v, R) := \sup_{p \in \Delta_{n-1}} \mathsf{G}_H(S \cap \{q \in \Delta_{n-1} : H(p, q) \leq Rv\}; v) \qquad \textit{(local packing)}.$$

*Both numbers use separation $v/2$. The global number counts separated points across $S$. The local number is the largest such count within any ball of radius $Rv$, where $R$ specifies the ball radius relative to the scale $v$.*

**Lemma 8** (Ellipsoid covers of a convex body and its polar)**.** *There is a universal constant $C_0 \geq 2$ such that, for every integer $s \geq 1$ and every origin-symmetric convex body $B \subseteq \mathbb{R}^s$, there is an origin-centered ellipsoid $E$ such that $B$ can be covered by at most $C_0^s$ translates of $E$, and $B^\circ$ can be covered by at most $C_0^s$ translates of $E^\circ$. Here, $B^\circ$ and $E^\circ$ denote the polar bodies of $B$ and $E$, respectively, where $K^\circ := \{y \in \mathbb{R}^s : \langle x, y\rangle \leq 1 \text{ for all } x \in K\}$.*

*Proof.* We use Milman's ellipsoid theorem in the formulation of Bobkov (2011, Theorem 1.1). There is an origin-centered ellipsoid $E$ such that

$$\frac{|B+E|}{|B\cap E|}\frac{|B^\circ+E^\circ|}{|B^\circ\cap E^\circ|} \leq C_M^s$$

for a universal $C_M$, where $|\cdot|$ denotes $s$-dimensional volume and $+$ denotes Minkowski addition. Let $x_1, \ldots, x_m \in B$ be the centers of a maximal family of translates $x_i + E/2$ with pairwise disjoint interiors. Each translate lies in $B + E$, and volume becomes $2^{-s}|E|$. Since their boundaries have zero volume,

$$m\,2^{-s}|E| = \left|\bigcup_{i=1}^m (x_i + E/2)\right| \leq |B+E|.$$

Moreover, each factor in Milman's inequality is at least one, and $B \cap E \subseteq E$, so

$$\frac{|B+E|}{|E|} \leq \frac{|B+E|}{|B\cap E|} \leq C_M^s.$$

Consequently,

$$m \leq 2^s \frac{|B+E|}{|E|} \leq (2C_M)^s.$$

Maximality makes the corresponding translates of $E$ cover $B$. The same argument applies to the polar pair. Set $C_0 = \max(2, 2C_M)$. □

**Lemma 9** (Dimension-uniform Hellinger packing)**.** *There is a universal constant $C_0 \geq 2$ such that, for every positive compact convex probability set $S$ of affine dimension $s$ and every $D_0 > 0$ satisfying $D(p\|q) \leq D_0$ for all $p, q \in S$, the following bound holds for every $v > 0$:*

$$\mathsf{G}_H(S; v) \leq [C_0^2(1 + 2\sqrt{2D_0}/v)]^s.$$

*In particular, for coordinate floor $\epsilon$,*

$$\mathsf{G}_H(S; v) \leq [C_0^2(1 + 2\sqrt{2\kappa}/v)]^s, \qquad \mathsf{D}_H(S; v, R) \leq [C_0^2(1 + 4\sqrt{2\kappa}R)]^s.$$

*Proof.* We reduce Hellinger packing in $S$ to Euclidean packing in $\mathbb{R}^s$. Step 1 maps points at Hellinger distance greater than $v/2$ to points at Euclidean distance greater than $2v$. Step 2 uses Lemma 8 to cover the image by at most $C_0^{2s}$ balls of radius $2\sqrt{2D_0}$, then counts the separated image points by comparing volumes. Step 3 specializes the divergence diameter to obtain the global and local estimates. If $s = 0$, then $S$ is a singleton and all the asserted packing bounds hold. Henceforth assume $s \geq 1$.

**Step 1. Construct a map that expands Hellinger separation.** The construction uses the fact that symmetrized KL divergence can be written as an inner product between a difference of probabilities and a difference of their logarithms. We first express this identity in $s$-dimensional coordinates, then turn the inner product into a lower bound on a Euclidean distance squared.

Choose $p_0 \in S$. Since $S$ has affine dimension $s$, the linear space $\operatorname{span}\{p - p_0 : p \in S\}$ has dimension $s$. Choose a basis $a_1, \ldots, a_s$ of this space and define $A := [a_1 \ \cdots \ a_s] \in \mathbb{R}^{n\times s}$. Define

$$p(x) := p_0 + Ax \quad (x \in \mathbb{R}^s), \qquad X := \{x \in \mathbb{R}^s : p(x) \in S\}.$$

Then every point in $S$ has a unique representation $p(x)$ with $x \in X$.

The set $X$ is compact, convex, full-dimensional in $\mathbb{R}^s$, and contains zero. Define $g(x) = A^\mathsf{T} \log p(x)$, where the logarithm is taken coordinatewise. This is the gradient of negative entropy $\sum_i p_i(x) \log p_i(x)$, since each column of $A$ sums to zero. Since $A(x - y) = p(x) - p(y)$, we

have

$$\begin{aligned}\langle x-y, g(x)-g(y)\rangle &= \sum_i \big(p_i(x)-p_i(y)\big)\log\frac{p_i(x)}{p_i(y)}\\ &= D(p(x)\|p(y)) + D(p(y)\|p(x)).\end{aligned}$$

To turn this inner product into a distance bound, we will use $\|r+s\|^2 \geq 4\langle r, s\rangle$, which follows from $\|r-s\|^2 \geq 0$. Thus we want to add transformed versions of $x$ and $g(x)$ whose differences preserve the inner product above.

Define $B = X - X := \{x-y : x \in X,\ y \in X\}$, an origin-symmetric convex body. Apply Lemma 8 to $B$, and let $E$ be the resulting ellipsoid, with covering constant $C_0$. Choose an invertible linear map $L$ taking $E$ to the Euclidean unit ball. Then $L^{-\mathsf{T}}$ takes $E^\circ$ to the same unit ball, and the two transformations preserve inner products: $\langle Lr, L^{-\mathsf{T}}s\rangle = \langle r, s\rangle$. This choice will also let us use the ellipsoid covers to bound the image in Step 2. Define

$$F(x) := \sqrt{2D_0}\, Lx + (2D_0)^{-1/2} L^{-\mathsf{T}}(g(x)-g(0)),$$

where the two scaling factors multiply to one, so they preserve the inner product. As shown in Step 2, they also give the two summands the same covering radius. Applying the preceding norm inequality to the two summands of $F(x)-F(y)$ gives

$$\begin{aligned}\|F(x)-F(y)\|^2 &\geq 4\langle x-y, g(x)-g(y)\rangle\\ &= 4[D(p(x)\|p(y)) + D(p(y)\|p(x))]\\ &\geq 16H(p(x),p(y))^2.\end{aligned}$$

The last inequality follows by summing $(a-b)\log(a/b) \geq 4(\sqrt{a}-\sqrt{b})^2$ over coordinates. For completeness, when $r \geq 1$, the function $\log r - 2(r-1)/(r+1)$ vanishes at 1 and has derivative $(r-1)^2/[r(r+1)^2] \geq 0$. Taking $r = \sqrt{a/b}$ proves the scalar inequality for $a \geq b$, and symmetry gives the other case.

Let $Q \subseteq S$ be any finite set with pairwise Hellinger distances greater than $v/2$. Its corresponding image points under $F$ therefore have pairwise Euclidean distances greater than $2v$. It remains to bound how many such points the image can contain.

**Step 2. Cover the image and count the separated points.** We will cover each of the two summands of $F$ by at most $C_0^s$ balls of radius $\sqrt{2D_0}$. Adding the centers of these two covers will then cover $F(X)$ by at most $C_0^{2s}$ balls of radius $2\sqrt{2D_0}$.

Since $0 \in X$, we have $X \subseteq B$, so the first summand is controlled by the cover of $B$. For the second summand, we first show that $g(x)-g(0) \in 2D_0B^\circ$. For $u, v, x \in X$, expanding the divergences gives

$$\begin{aligned}\langle g(x)-g(0), u-v\rangle &= D(p(u)\|p_0) + D(p(v)\|p(x))\\ &\quad - D(p(u)\|p(x)) - D(p(v)\|p_0).\end{aligned}$$

Each divergence lies in $[0, D_0]$, so the absolute value of this expression is at most $2D_0$. Since every element of $B = X - X$ has the form $u - v$, the definition of the polar body $B^\circ := \{y \in \mathbb{R}^s : \langle y, b\rangle \leq 1 \text{ for every } b \in B\}$ implies

$$g(x) - g(0) \in 2D_0B^\circ \qquad (x \in X).$$

The cover of $B$ by at most $C_0^s$ translates of $E$, after applying $\sqrt{2D_0}L$, covers the first summand by at most $C_0^s$ Euclidean balls of radius $\sqrt{2D_0}$. Similarly, cover $B^\circ$ by at most $C_0^s$ translates of $E^\circ$, scale by $2D_0$, and apply $(2D_0)^{-1/2}L^{-\mathsf{T}}$. The inclusion just proved shows that these balls cover the second summand, and their radius is $2D_0(2D_0)^{-1/2} = \sqrt{2D_0}$. Let $c_1, \ldots, c_M$ and $d_1, \ldots, d_N$ be the centers of the covers for the first and second summands, respectively, where $M, N \leq C_0^s$. For every $x \in X$, there exist $i$ and $j$ such that

$$\Big\|\sqrt{2D_0}Lx - c_i\Big\| \leq \sqrt{2D_0}, \qquad \Big\|(2D_0)^{-1/2}L^{-\mathsf{T}}(g(x)-g(0)) - d_j\Big\| \leq \sqrt{2D_0}.$$

The triangle inequality therefore gives

$$\begin{aligned}\|F(x)-(c_i+d_j)\| &\leq \Big\|\sqrt{2D_0}Lx - c_i\Big\| + \Big\|(2D_0)^{-1/2}L^{-\mathsf{T}}(g(x)-g(0)) - d_j\Big\|\\ &\leq 2\sqrt{2D_0}.\end{aligned}$$

Thus the balls centered at all pairwise sums $c_i + d_j$, each of radius $2\sqrt{2D_0}$, cover $F(X)$. There are at most $MN \le C_0^{2s}$ such balls, which proves the desired covering bound.

We now count the image points corresponding to $Q$ within each ball of this cover. Fix one such covering ball, whose radius is $2\sqrt{2D_0}$, and suppose it contains $m$ image points. The Euclidean balls of radius $v$ centered at these points are pairwise disjoint. By the triangle inequality, all these smaller balls lie in the concentric ball of radius $2\sqrt{2D_0} + v$. Writing $\omega_s$ for the volume of the Euclidean unit ball in $\mathbb{R}^s$, comparison of volumes gives

$$m\,\omega_s v^s \le \omega_s\big(2\sqrt{2D_0} + v\big)^s, \qquad m \le \left(1 + \frac{2\sqrt{2D_0}}{v}\right)^s.$$

Since at most $C_0^{2s}$ covering balls contain all the image points, we obtain

$$|Q| \le C_0^{2s}\left(1 + \frac{2\sqrt{2D_0}}{v}\right)^s.$$

Taking the supremum over such $Q$ proves the first packing bound.

**Step 3. Specialize the divergence diameter.** *Global bound.* When every coordinate is at least $\epsilon$, we have $D(p\|q) \le \log(1/\epsilon) \le \kappa$ for all $p, q \in S$. Taking $D_0 = \kappa$ in the bound just proved gives

$$\mathsf{G}_H(S; v) \le [C_0^2(1 + 2\sqrt{2\kappa}/v)]^s.$$

*Local bound.* Fix a center $p \in \Delta_{n-1}$ and consider $K = S \cap \{q : H(p, q) \le Rv\}$. This set is compact and convex, because $H(p, q)^2$ is continuous and convex in $q$. For any $q, q' \in K$, the triangle inequality and Lemma 4 give

$$H(q, q') \le 2Rv, \qquad D(q\|q') \le \kappa H(q, q')^2 \le 4\kappa R^2 v^2.$$

If $K$ is empty or a singleton, its packing count is at most one. Otherwise $R > 0$, and we apply the bound from Step 2 to $K$ in its own affine span, with $D_0 = 4\kappa R^2 v^2$ and the same separation $v/2$. Since $\mathrm{affdim}(K) \le s$, this gives

$$\mathsf{G}_H(K; v) \le [C_0^2(1 + 4\sqrt{2\kappa}R)]^{\mathrm{affdim}(K)} \le [C_0^2(1 + 4\sqrt{2\kappa}R)]^s.$$

Taking the supremum over centers $p$ proves the local bound. □

### B.5 Partitioning observations and counting pieces

**The partition rule.** The local comparison requires an associated comparator that was available throughout the block. For a completed scale-zero threshold block, set $\mathcal{P}(E) = \{E\}$. For a completed positive-level block with children $E_1, \ldots, E_m$ in chronological order, set

$$\mathcal{P}(E) = \begin{cases} \{E\}, & E \text{ is not a boundary block,} \\ \bigcup_{i=1}^m \mathcal{P}(E_i), & E \text{ is a boundary block.} \end{cases}$$

At a depth-interval endpoint, apply this map to the completed scale-$(K-1)$ roots and apply the child union to the unfinished positive-scale roots of the remaining forest. Use only maximal roots, and add the unique unfinished positive-weight scale-zero successor of each $(g, a)$, if present. Empty blocks contribute nothing. This is the parent-or-children selection rule from Section 4.2. It is defined from the chronological updates for analysis and is not maintained by the algorithm.

For $0 \le j < K - 1$ and each target group $g$, count its completed scale-$j$ residual pieces over all source predictions. Let $Q_j(I)$ be the largest of these counts.

**Counting each observation once.** The parent-or-children recursion selects a disjoint partition. Each weighted observation belongs to one depth interval and one source; each selected completed piece below the finest level has a unique parent, which is boundary or unfinished.

**Lemma 10** (Partition and piece counts)**.** *On $\mathcal{E}_{\mathrm{global}}$, let $I$ be any nonempty prefix of a depth-$K$ interval. Let $M_I$ and $B_I$ be the counts defined in Section 4.2. Moreover, $B_I \le M_I$, and for every level $0 \le j < K - 1$,*

$$Q_j(I) \le \min\left\{8M_I B_I, \frac{|I|}{m_j}\right\}.$$

*Proof.* We first prove that $B_I \le M_I$. Fix a group, a source prediction, and a positive level. Each completed boundary block has a comparator that was unavailable at its start but is available by its completion. Since blocks with these fixed indices are processed consecutively, this comparator is available at the start of every later block and cannot serve as the comparator of a later boundary block. Thus these boundary blocks have distinct comparators, all counted by $M_I$, so their number is at most $M_I$. Taking the maximum over groups, sources, and positive levels, and using $M_I \ge 1$, gives $B_I \le M_I$.

Next, we bound $Q_j(I)$ for $0 \le j < K-1$. Fix a group and a source prediction. Every selected completed level-$j$ piece has a unique level-$(j+1)$ parent, which is either boundary or unfinished. For this group and source, there are at most $B_I$ completed boundary parents at level $j+1$, each with four completed children, and at most one unfinished parent, with at most three completed children. Thus there are at most $4B_I + 3$ selected completed level-$j$ pieces for this group and source. Summing over at most $M_I$ sources and then taking the maximum over groups gives

$$Q_j(I) \le (4B_I + 3)M_I \le 8M_I B_I \le 8M_I^2.$$

For each group, the selected completed level-$j$ pieces have total weight at most $\sum_{t\in I} g(x_t) \sum_a \pi_{t,a} \le |I|$, and each has weight at least $m_j$. Taking the maximum of their counts over groups gives $Q_j(I) \le |I|/m_j$.

□

### B.6 Proof of the residual-piece bound

**Lemma 11** (Calibration error from residual-piece counts)**.** *On $\mathcal{E}_{\mathrm{global}}$, for every depth $K$ and every nonempty prefix $I$ of a depth-$K$ interval, as defined in Section 4.2, we have*

$$\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I) \le \underbrace{6\sqrt{\kappa}\,|I| w_{K-1}}_{\substack{\text{completed pieces}\\ \text{at scale } K-1}} + \underbrace{12288\sqrt{\kappa}\,nL \sum_{j=0}^{K-2} \frac{Q_j(I)}{w_j}}_{\substack{\text{completed pieces}\\ \text{below scale } K-1}} + \underbrace{2\min(|I|, m_0 M_I)}_{\substack{\text{unfinished}\\ \text{scale-zero pieces}}} + 4\sqrt{nLM_I T} + 2\beta|I|.$$

*Proof of Lemma 11.* On $\mathcal{E}_{\mathrm{global}}$, the sampling bound (11), the smoothing correction $\|y_t - y'_t\|_1 \le 2\beta$, and the block partition defined in Section 4.2 give

$$\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I) \le \max_{g\in\mathcal{G}} \sum_a \sum_{E\in\mathcal{P}_{g,a}(I)} N_E \|z_E - a\|_1 + 4\sqrt{nLM_I T} + 2\beta|I|.$$

Fix a group $g$. Here the sampling bound uses the $M_I$ stored prediction values counted at the end of $I$, which include every prediction value used in $I$. The partition includes unfinished scale-zero pieces of positive weight, so their means are also well-defined. We bound the piece costs separately for completed pieces at the top scale, completed pieces at lower scales, and unfinished scale-zero pieces.

For each selected completed scale-$j$ residual piece, Lemma 2 gives $N_E\|z_E - a\|_1 \le 6\sqrt{\kappa}\, w_j N_E$. The completed scale-$(K-1)$ residual pieces have total weight at most $|I|$, so their total cost is bounded by

$$\sum_a \sum_{\substack{E\in\mathcal{P}_{g,a}(I)\\ E \text{ completed at scale } K-1}} N_E\|z_E - a\|_1 \le 6\sqrt{\kappa}\, w_{K-1}|I|.$$

Now fix a lower scale $0 \le j < K-1$. Each completed scale-$j$ piece has $N_E < 2m_j$, and there are at most $Q_j(I)$ such selected pieces over all sources for this group. Using $m_j = 1024nL/w_j^2$, their total cost therefore satisfies

$$\sum_a \sum_{\substack{E\in\mathcal{P}_{g,a}(I)\\ E \text{ completed at scale } j}} N_E\|z_E - a\|_1 \le 12\sqrt{\kappa}\, w_j m_j Q_j(I) = \frac{12288\sqrt{\kappa}\, nL Q_j(I)}{w_j}.$$

There is at most one unfinished scale-zero piece per source, each with $N_E < m_0$. Since $\|z_E - a\|_1 \le \|z_E\|_1 + \|a\|_1 = 2$, each such piece has cost at most $2N_E$. Their total weight is also at most $|I|$, by disjointness and $\sum_a \pi_{t,a} = 1$. There are at most $M_I$ sources, so

$$\sum_a \sum_{\substack{E \in \mathcal{P}_{g,a}(I) \\ E \text{ unfinished at scale } 0}} N_E \|z_E - a\|_1 \le 2\min(|I|, m_0 M_I).$$

Summing over the lower scales, taking the maximum over groups, and substituting the three cost bounds into the opening inequality gives

$$\begin{aligned}\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \le 6\sqrt{\kappa}\,|I| w_{K-1} + 12288\sqrt{\kappa}\,nL \sum_{j=0}^{K-2} \frac{Q_j(I)}{w_j} + 2\min(|I|, m_0 M_I) \\ + 4\sqrt{nLM_I T} + 2\beta|I|.\end{aligned}$$

The event is uniform over intervals and prediction subsets, so the same argument holds simultaneously at every permitted endpoint. □

### B.7 Proof of the discovery-count corollary

*Proof of Corollary 1.* On $\mathcal{E}_{\mathrm{global}}$, start from Lemma 11. Write $w = w_{K-1}$, so $w \le 1$ because $K \ge 2$.

For the lower-scale term, the proof of Lemma 10 gives $Q_j(I) \le 8M_I B_I$. Since $w_j = 2^{1-j}$, we have $\sum_{j=0}^{K-2} w_j^{-1} = w^{-1} - 1/2 < w^{-1}$, and hence

$$\begin{aligned}12288\sqrt{\kappa}\,nL \sum_{j=0}^{K-2} \frac{Q_j(I)}{w_j} &\le 12288\sqrt{\kappa}\,nL \cdot 8M_I B_I \sum_{j=0}^{K-2} \frac{1}{w_j} \\ &\le 98304\sqrt{\kappa}\,\frac{nLM_I B_I}{w}.\end{aligned}$$

For the unfinished scale-zero term, $m_0 = 1024nL/w_0^2 = 256nL$ gives

$$\begin{aligned}2\min(|I|, m_0 M_I) &\le 2m_0 M_I \\ &= 512nLM_I \\ &\le 512\sqrt{\kappa}\,\frac{nLM_I B_I}{w},\end{aligned}$$

where the last inequality uses $w \le 1$ and $B_I, \kappa \ge 1$. Substituting these two bounds into Lemma 11 and using $98304 + 512 = 98816$ gives

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \le 6\sqrt{\kappa}\,|I| w + 98816\sqrt{\kappa}\,\frac{nLM_I B_I}{w} + 4\sqrt{nLM_I T} + 2\beta|I|.$$

Bounding 6 and 98816 by $10^6$ in this estimate proves the corollary. □

## C Choosing the resolution from the data

### C.1 Adaptive procedure

Algorithm 2 implements Section 5, with $\Gamma_K$ as defined there. The interval residual includes all rounds through the test, including weight in unfinished blocks. For its incremental computation, maintain

$$Y_{I,g,a} = \sum_{t \in I} g(x_t)\mathbf{1}\{p_t = a\}(y_t - a), \qquad \mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) = \max_{g \in \mathcal{G}} \sum_a \|Y_{I,g,a}\|_1.$$

**Algorithm 2** Choosing the number of accuracy levels

**Require:** Inputs of Algorithm 1, except the prescribed depth $K$.
Set $K \leftarrow K_{\max}$ and initialize the state of Algorithm 1 once.
Set $I \leftarrow \varnothing$ and clear all calibration-residual accumulators. Any previously uninitialized $Y_{I,g,a}$ is interpreted as the zero vector in $\mathbb{R}^n$.
**for** $t = 1, \dots, T$ **do**
  Execute round $t$ of Algorithm 1 at depth $K$, including its comparison update and all discovery updates.
  Extend $I \leftarrow I \cup \{t\}$, carrying forward its residual vectors.
  **for all** $g \in \mathcal{G}$ **do**
    Set $Y_{I,g,p_t} \leftarrow Y_{I,g,p_t} + g(x_t)(y_t - p_t)$, leaving other prediction entries unchanged.
  **end for**
  Compute $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I) = \max_{g\in\mathcal{G}} \sum_a \|Y_{I,g,a}\|_1$, summing over all discovered prediction values.
  **if** $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(I) > \Gamma_K$, $K > 1$, and $t < T$ **then**
    End $I$ at $t$ and set $K \leftarrow K - 1$.
    Clear all local block accumulators and residual vectors, and set $I \leftarrow \varnothing$.
    Retain all discovered prediction values, their memberships in the sets $P_j$, and comparison aggregates whose endpoints remain active.
  **end if**
**end for**

### C.2 Proof of the affine prediction and boundary-block counts

*Proof of Lemma 3.* Fix an interval $I$ at depth $K$, and abbreviate $P_j(\max I + 1)$ by $P_j$ for $j \geq 1$ throughout this proof. Fix $w = w_{K-1} \geq 64\sqrt{\rho_\infty(S)}$ and let $S_\beta = (1-\beta)S + \beta q_0$. Smoothing multiplies $\ell_1$ distances by $1-\beta$, so $H^2 \leq \|\cdot\|_1$ gives $\mathrm{dist}_H((1-\beta)\mu_t + \beta q_0, S_\beta) \leq \sqrt{\rho_\infty(S)} \leq w/64$. By convexity of $S_\beta$ and joint convexity of $H^2$, the same bound holds for every weighted predictable mean. Each value inserted into $P_j$ comes from a completed scale-$\ell$ block $E$ with $\ell \geq j$. Applying (10) at scale $\ell$ and using the triangle inequality gives

$$\mathrm{dist}_H(z_E, S_\beta) \leq H(z_E, \bar{\mu}'_E) + \mathrm{dist}_H(\bar{\mu}'_E, S_\beta) \leq \frac{w_\ell}{16} + \frac{w}{64} \leq \frac{w_j}{8}.$$

This includes values inherited from earlier depths and those inserted on the final round of $I$.

**Step 1. Count prediction values.** Values in $P_j$ are more than $w_j$ apart, so their representatives in $S_\beta$ are more than $w_j/2$ apart. To see this, for each $a \in P_j$, choose a representative $\tilde{a} \in S_\beta$ with $H(a, \tilde{a}) \leq w_j/8$. For distinct $a, b \in P_j$, the insertion rule gives $H(a,b) > w_j$, so

$$H(\tilde{a}, \tilde{b}) \geq H(a,b) - H(a,\tilde{a}) - H(b,\tilde{b}) > \frac{3w_j}{4} > \frac{w_j}{2}.$$

Thus, these representatives form a $w_j/2$-separated subset of $S_\beta$, and $|P_j| \leq \mathsf{G}_H(S_\beta; w_j)$.

The global bound in Lemma 9 gives $|P_j| \leq \left[C_0^2\left(1 + \frac{2\sqrt{2\kappa}}{w_j}\right)\right]^s \leq \left(\frac{a\sqrt{\kappa}}{w_j}\right)^s$ for a universal $a = C_0^2(1+2\sqrt{2}) \geq 1$. Since $w_j \geq w$,

$$M_I = \left|P_0 \cup \bigcup_{j=1}^{K-1} P_j\right| \leq 1 + \sum_{j=1}^{K-1} |P_j| \leq K_{\max}(a\sqrt{\kappa}/w)^s.$$

**Step 2. Count boundary comparators in chronological order.** Recall that $B_I$ is the maximum, over groups, sources, and positive levels, of the number of completed boundary blocks in $I$, with a minimum value of 1. For each fixed group, source, and level, these blocks have distinct comparators: each earlier comparator remains available throughout every later block and therefore cannot serve as the comparator of a later boundary block. It therefore suffices to bound the number of distinct comparators for each fixed group, source, and positive level. Fix a group, source $a$, and positive level $j$, and list the completed boundary blocks' distinct comparators $q_1, q_2, \dots \in P_j$ in closure order. Let $E_m$ be the block associated with $q_m$, with mean $z_m$ and weight $N_m$. For $i < m$, the earlier comparator $q_i$ is available throughout $E_m$, so applying (3) with source $a$ and comparator $q_i$ gives

$$H(z_m, a) \leq \sqrt{3\kappa}\, H(z_m, q_i) + 4\sqrt{\frac{\kappa L}{N_m}}.$$

By the definition of $q_m$, we have $H(q_m, z_m) \le w_j$. Applying the triangle inequality twice therefore yields

$$\begin{aligned} H(q_m, a) &\le H(q_m, z_m) + H(z_m, a) \\ &\le H(q_m, z_m) + \sqrt{3\kappa}\big(H(z_m, q_m) + H(q_m, q_i)\big) + 4\sqrt{\frac{\kappa L}{N_m}} \\ &\le \sqrt{3\kappa}\, H(q_m, q_i) + (1 + \sqrt{3\kappa}) w_j + 4\sqrt{\frac{\kappa L}{N_m}}. \end{aligned}$$

Since $E_m$ is completed, $N_m \ge m_j = 1024 nL / w_j^2$, and hence

$$4\sqrt{\frac{\kappa L}{N_m}} \le \frac{\sqrt{\kappa}}{8\sqrt{n}}\, w_j.$$

Using $n, \kappa \ge 1$, both $\sqrt{3\kappa}$ and $1 + \sqrt{3\kappa} + \sqrt{\kappa}/(8\sqrt{n})$ are at most $16\sqrt{\kappa}$. Thus

$$H(q_m, a) \le A\big(H(q_m, q_i) + w_j\big), \qquad A = 16\sqrt{\kappa}.$$

We count these comparators by partitioning them according to their Hellinger distance from the source $a$. For each $q_i$, choose a nearby representative $\tilde{q}_i \in S_\beta$ with $H(q_i, \tilde{q}_i) \le w_j/8$. We bound the separation and radius of the representatives in each region:

- **Near region:** $H(q_i, a) \le 4A w_j$**.** By the separation argument in Step 1, $H(\tilde{q}_i, \tilde{q}_k) > w_j/2$. Each representative also satisfies

  $$H(\tilde{q}_i, a) \le H(\tilde{q}_i, q_i) + H(q_i, a) \le \frac{w_j}{8} + 4A w_j \le 5A w_j.$$

  Thus the representatives are separated by more than $w_j/2$ and lie in the ball of radius $5A w_j$ centered at $a$.

- **Far regions.** For each integer $h \ge 0$, consider the comparators satisfying $4A2^h w_j < H(q_i, a) \le 8A2^h w_j$. For two comparators in the same far region, order them as $i < m$. The preceding chronological comparison gives

  $$H(q_m, q_i) \ge \frac{H(q_m, a)}{A} - w_j > (4 \cdot 2^h - 1) w_j.$$

  Replacing both comparators by their representatives loses at most $w_j/4$, so

  $$H(\tilde{q}_m, \tilde{q}_i) > (4 \cdot 2^h - 1) w_j - \frac{w_j}{4} > 2 \cdot 2^h w_j.$$

  Each representative lies within distance

  $$H(\tilde{q}_i, a) \le \frac{w_j}{8} + 8A2^h w_j \le 9A2^h w_j$$

  of $a$. Thus, the representatives in the $h$-th far region are separated by more than $2 \cdot 2^h w_j$ and lie in the ball of radius $9A2^h w_j$ centered at $a$.

Apply the local packing bound in Lemma 9 with $v = w_j$ for the near region and $v = 2^h w_j$ for the $h$-th far region. In both cases, the representatives are separated by more than $v/2$ and lie within radius $Rv$ of $a$, where $R = 300\sqrt{\kappa} > 9A$. The bound therefore limits each region to

$$[C_0^2(1 + 1200\sqrt{2}\,\kappa)]^s \le (b\kappa)^s$$

comparators, for a universal $b \ge 1$. It remains to count the nonempty regions. For any probability vectors $p, q$, we have $H(p, q) \le \sqrt{2}$. If $h \ge j$, a comparator in the $h$-th far region would have to satisfy

$$H(q_i, a) > 4A2^h w_j = 8A2^{h-j} \ge 8A > \sqrt{2},$$

which is impossible. Thus only the far regions indexed by $h = 0, \dots, j-1$ can be nonempty. There are at most $j$ such regions and one near region, giving at most $j + 1 \le K \le K_{\max}$ regions in total. Each region contains at most $(b\kappa)^s$ comparators, so there are at most $K_{\max}(b\kappa)^s$ comparators in total. Since $K_{\max}(b\kappa)^s \ge 1$, this bound also covers the minimum value of 1 in the definition of $B_I$. By the reduction at the start of this step,

$$B_I \le K_{\max}(b\kappa)^s.$$

□

### C.3 Proof of Theorem 1

*Proof.* Work on $\mathcal{E}_{\text{global}}$ , we prove the stronger guarantee in Remark 1 and then specialize it to the exact affine dimension of the means. Write $\rho = \rho_\infty(S)$. The algorithm adjusts its current depth $K$ using only the observed residual tests. For any visited depth $K \geq 2$ with width $w = w_{K-1} \geq 64\sqrt{\rho}$, Lemma 3 bounds $M_I$ and $B_I$ for every nonempty prefix $I$ of the depth-$K$ interval. Using $|I| \leq T$ and substituting these counts into Corollary 1 gives

$$\begin{aligned}\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \leq 10^6\sqrt{\kappa}\Bigg[ &\underbrace{Tw}_{\text{top-piece cost}} + \underbrace{\frac{nLK_{\max}^2(ab\kappa^{3/2})^s}{w^{s+1}}}_{\text{discovery cost}} \\ &+ \underbrace{\sqrt{nLK_{\max}(a\sqrt{\kappa}/w)^s T}}_{\text{sampling correction}}\Bigg] + \underbrace{2\beta|I|}_{\text{smoothing correction}}.\end{aligned} \tag{12}$$

We first choose a comparison width, then show that this error bound prevents the algorithm from moving to a coarser scale, and finally sum the interval errors. The comparison width depends on $s, S$ only in the analysis and is not an input to the algorithm.

**Step 1. Choose a comparison width.** We seek a small width $w$ for which the discovery cost is controlled by the top-piece cost $Tw$. Increasing $w$ decreases the discovery cost but increases $Tw$. The width at which these two costs are equal is

$$Tw_\star^{s+2} = nLK_{\max}^2(ab\kappa^{3/2})^s.$$

Since $LK_{\max}^2(ab\kappa^{3/2})^s \geq 1$, we have

$$w_\star \geq (n/T)^{1/(s+2)} \geq T^{-1/2}. \tag{13}$$

To ensure that Lemma 3 applies, we also require $w \geq 64\sqrt{\rho}$. Accordingly, set

$$v_\star = \max\{2w_\star, 64\sqrt{\rho}\}.$$

If $v_\star > 1$, the deterministic bound gives $\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \leq 2T < 2Tv_\star$. We include this case in the final parameter substitution in Step 3. For the arguments below, assume $v_\star \leq 1$. The available widths at depths $2, \ldots, K_{\max}$ are

$$1,\ \frac{1}{2},\ \ldots,\ 2^{-\lceil \log_2 T\rceil}.$$

Since

$$2^{-\lceil \log_2 T\rceil} \leq T^{-1} \leq 2T^{-1/2} \leq 2w_\star \leq v_\star \leq 1,$$

$v_\star$ lies within this range. Round $v_\star$ up to the smallest available width that is at least $v_\star$, and let $K^\star$ be the depth with its finest width used denoted by $w_{K^\star-1} \geq v_\star$. Consecutive widths differ by a factor of two, so

$$2 \leq K^\star \leq K_{\max}, \qquad v_\star \leq w_{K^\star-1} < 2v_\star.$$

Thus, $w_{K^\star-1} \geq 64\sqrt{\rho}$, as required by Lemma 3. Moreover, $w_{K^\star-1} \geq 2w_\star \geq T^{-1/2}$ by (13), which implies $Tw_{K^\star-1} \geq 1$ since $n, T \geq 1$ and $s \geq 0$. The discovery term at this width satisfies

$$\frac{nLK_{\max}^2(ab\kappa^{3/2})^s}{w_{K^\star-1}^{s+1}} = Tw_{K^\star-1}\left(\frac{w_\star}{w_{K^\star-1}}\right)^{s+2} \leq \frac{Tw_{K^\star-1}}{4},$$

because $w_{K^\star-1} \geq 2w_\star$ and $s \geq 0$.

**Step 2. The algorithm never moves to a coarser scale.** If the algorithm reaches depth $K^\star$, (12) applies with $w = w_{K^\star-1}$ to every nonempty prefix $I$ of the resulting depth-$K^\star$ interval, since Step 1 ensures $w_{K^\star-1} \geq 64\sqrt{\rho}$. Step 1 bounds the discovery term by $Tw_{K^\star-1}/4$. Using this bound and $K_{\max}(b\kappa)^s \geq 1$, the sampling correction satisfies

$$\sqrt{nLK_{\max}\left(\frac{a\sqrt{\kappa}}{w_{K^\star-1}}\right)^s T} \leq \sqrt{Tw_{K^\star-1}\,\frac{nLK_{\max}^2(ab\kappa^{3/2})^s}{w_{K^\star-1}^{s+1}}} \leq \frac{Tw_{K^\star-1}}{2}.$$

Finally, since $\beta = 1/(2T^2)$ and $|I| \le T$, the smoothing correction satisfies $2\beta|I| = |I|/T^2 \le 1/T$. Substituting these bounds gives, after every round at this depth,

$$\begin{aligned}\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) &\le 10^6\sqrt{\kappa}\left(1+\frac{1}{4}+\frac{1}{2}\right)Tw_{K^\star-1}+\frac{1}{T}\\ &= \frac{7}{4}\cdot 10^6\sqrt{\kappa}\,Tw_{K^\star-1}+\frac{1}{T}\\ &< 3\cdot 10^6\sqrt{\kappa}\,Tw_{K^\star-1}=\Gamma_{K^\star},\end{aligned}$$

where the strict inequality uses $Tw_{K^\star-1} \ge 1$ and $\sqrt{\kappa} \ge 1$. Thus, if the algorithm reaches depth $K^\star$, it cannot leave it. Since depth decreases one level at a time from $K_{\max}$, every visited depth satisfies $K \ge K^\star$, and every visited width is at most $w_{K^\star-1}$. No affine bound is needed at the finer widths visited before this depth.

**Step 3. Sum the interval errors and obtain the final rate.** Let $J_1, \dots, J_m$ be the visited intervals, with depths $K_1 > \cdots > K_m$ and $m \le K_{\max}$. For each group in $\mathcal{G}$, one round changes just one residual vector, by $g(x_t)(y_t - p_t)$, whose norm is at most two. Hence $\mathrm{MCerr}_{\mathcal{G}}^{(1)}(J_i) \le \Gamma_{K_i} + 2$ holds even if the horizon test fails that is when the budget is first exceeded at $t = T$. The intervals partition all rounds, including failed endpoints. The triangle inequality across intervals gives

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le \sum_{i=1}^{m}\mathrm{MCerr}_{\mathcal{G}}^{(1)}(J_i) \le 3\cdot 10^6\sqrt{\kappa}\,T\sum_{i=1}^{m}w_{K_i-1}+2m < 6\cdot 10^6\sqrt{\kappa}\,Tw_{K^\star-1}+2K_{\max}.$$

The last inequality uses the doubling of successive visited widths and their upper bound $w_{K^\star-1}$: $\sum_{i=1}^{m} w_{K_i-1} \le w_{K^\star-1}\left(1+\frac{1}{2}+\cdots+\frac{1}{2^{m-1}}\right) < 2w_{K^\star-1}$. Substituting $w_{K^\star-1} < 2v_\star$ from Step 1 gives

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le 12\cdot 10^6\sqrt{\kappa}\,Tv_\star + 2K_{\max}.$$

For $v_\star > 1$, the deterministic estimate in Step 1 implies the same inequality, without using a comparison depth. Since $v_\star \le 2w_\star + 64\sqrt{\rho}$, both cases imply

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le \underbrace{24\cdot 10^6\sqrt{\kappa}\,Tw_\star}_{\text{width term}} + \underbrace{768\cdot 10^6\sqrt{\kappa}\,T\sqrt{\rho}}_{\text{approximation term}} + \underbrace{2K_{\max}}_{\text{interval overshoot term}}.$$

We now bound the three terms on the right-hand side in the same order. Appendix F bounds $\kappa, L, K_{\max}$ by $8, 40, 6$ times $\log(en|\overline{\mathcal{G}}|(T+1)/\delta)$, respectively.

- *Width term.* Substituting the definition of $w_\star$ gives

$$\begin{aligned}\sqrt{\kappa}\,Tw_\star &= \kappa^{1/2+3s/[2(s+2)]}\left[LK_{\max}^2\right]^{1/(s+2)}(ab)^{s/(s+2)}n^{1/(s+2)}T^{(s+1)/(s+2)}\\ &\le 384\sqrt{40}\,ab\left[\log\left(\frac{en|\overline{\mathcal{G}}|(T+1)}{\delta}\right)\right]^2 n^{1/(s+2)}T^{(s+1)/(s+2)}.\end{aligned}$$

  Here the exponent of the logarithm is

$$\frac{1}{2}+\frac{3s/2+3}{s+2}=2.$$

  Since $s \ge 0$ and $a, b \ge 1$, the constant factor is at most $8^2(40\cdot 6^2)^{1/2}ab = 384\sqrt{40}\,ab$, independently of $s$.

- *Approximation term.* Using the bound on $\kappa$ and $\log(en|\overline{\mathcal{G}}|(T+1)/\delta) \ge 1$,

$$\sqrt{\kappa}\,T\sqrt{\rho} \le \sqrt{8}\left[\log\left(\frac{en|\overline{\mathcal{G}}|(T+1)}{\delta}\right)\right]^{1/2}T\sqrt{\rho}$$

- *Interval overshoot term.* Using the bound on $K_{\max}$, $\log(en|\overline{\mathcal{G}}|(T+1)/\delta) \ge 1$, and $n, T \ge 1$,

$$2K_{\max} \le 12\log\left(\frac{en|\overline{\mathcal{G}}|(T+1)}{\delta}\right) \le 12\left[\log\left(\frac{en|\overline{\mathcal{G}}|(T+1)}{\delta}\right)\right]^2 n^{1/(s+2)}T^{(s+1)/(s+2)}.$$

Substituting these three bounds into the error estimate and using

$$\log\left(\frac{en|\overline{\mathcal{G}}|(T+1)}{\delta}\right) \le 2\log\left(\frac{eTn|\overline{\mathcal{G}}|}{\delta}\right),$$

which follows from $T+1 \le 2T$ and $\log(eTn|\overline{\mathcal{G}}|/\delta) \ge 1$, gives, for a universal $C$,

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le C\left[\log\left(\frac{eTn|\overline{\mathcal{G}}|}{\delta}\right)\right]^2 \left(n^{1/(s+2)}T^{(s+1)/(s+2)} + T\sqrt{\rho}\right).$$

Because the event $\mathcal{E}_{\mathrm{global}}$ does not depend on $S, s$, this bound holds simultaneously for every admissible $S, s$, proving Remark 1. Taking $S = \mathrm{conv}\{\mu_1, \dots, \mu_T\}$ and $s = \mathrm{affdim}(S)$ gives $\rho_\infty(S) = 0$ and hence Theorem 1, even when the means and their dimension depend on the realized history. □

### C.4 Runtime

We count exact real arithmetic operations, comparisons, evaluations of logarithms and square roots, and evaluations of the given group functions, each at unit cost. A fresh uniform draw from $[0,1]$ also costs one operation; sampling a finite prediction distribution uses its cumulative probabilities. There are at most $A_{\max}$ distinct prediction values, at most $|\overline{\mathcal{G}}|A_{\max}^2$ pair aggregates, and at most one source/group closure per scale per round. Each recursive call increases scale, so all closure loops are finite. Evaluating all active losses costs $O(TnA_{\max})$ operations. Each of at most $T|\overline{\mathcal{G}}|A_{\max}K_{\max}$ closures scans at most $A_{\max}$ stored predictions at each of at most $K_{\max}$ membership levels, for $O(Tn|\overline{\mathcal{G}}|A_{\max}^2K_{\max}^2)$ operations. Recognizing an already discovered value takes at most $O(nA_{\max})$ comparisons per closure and fits this bound. Solving the linear stationarity equations costs $O(TA_{\max}^3)$, and pair updates cost $O(T|\overline{\mathcal{G}}|A_{\max}^2)$. Group evaluations, residual updates, block-weight updates and sampling fit these bounds. Together with the parameter choices, these bounds give a polynomial operation count in $T$, $n$, $|\overline{\mathcal{G}}|$, and $\log(1/\delta)$ under the stated exact-operation model.

## D Affine-dimensional lower bound

For the lower bound, we construct a sequence of contexts and specify the outcome distribution for each context. We associate each context with a point on an $s$-dimensional grid in the simplex. The learner's observation $y_t$ is the average of this grid point and an independently and uniformly chosen simplex vertex. The resulting conditional means lie in an affine space of dimension at most $s$, while the random observations fluctuate across all $n$ coordinates.

*Proof of Theorem 2.* The proof separates two cases according to the expected total distance between predictions and conditional means. If this distance is large, the group family gives a calibration lower bound. If it is small, the separation between grid means forces the learner to use many distinct prediction values. We then use the noise accumulated within these prediction bins to obtain the lower bound. We choose the grid resolution so that both cases yield the claimed rate. Write $L_T := \log(T+1)$ and $R_s(T,n) := n^{1/(s+2)}T^{(s+1)/(s+2)}$, and assume $s \ge 1$ and $T \ge \max\{n, 2^{s+2}\}$. The proof follows the following five steps.

**Step 1. Construct a low-dimensional grid with noise on all coordinates.** Let $k := \lfloor n/2 \rfloor$ and $m := 2^{\max\{1, \lceil \log_2(T/k)/(s+2) \rceil\}}$. We construct an oblivious adversary with context space $[m]^s$. The contexts $x_t = (x_{t,1}, \dots, x_{t,s})$, for $t = 1, \dots, T$, cycle through this space in a fixed order. We associate each context $x = (x_1, \dots, x_s) \in [m]^s$ with a probability vector $q(x)$. The first $s$ coordinates of these vectors form a Cartesian grid with $m$ equally spaced values in $[a, 2a]$ along each coordinate.

With $a := 1/[4(s+1)]$, define

$$q_j(x) := a + a\frac{x_j - 1}{m-1} \quad (1 \le j \le s).$$

Complete this to an $n$-coordinate probability vector by setting

$$q(x) := \Big(\underbrace{q_1(x), \dots, q_s(x)}_{s \text{ varying coordinates}}, \quad \underbrace{1 - \sum_{j=1}^{s} q_j(x)}_{\text{remaining probability mass}} \quad , \quad \underbrace{0, \dots, 0}_{n-s-1 \text{ coordinates}} \Big).$$

The final zero block is absent when $s = n - 1$. Our choice of $a$ makes the remaining coordinate at least $1 - 2sa = (s+2)/(2(s+1)) > 0$, so all coordinates are nonnegative and sum to one.

On round $t$, the adversary selects a simplex vertex $e_{J_t}$ by drawing its index $J_t$ uniformly from $[n]$. These draws are independent across rounds and of the learner's randomness. Fix an arbitrary forecaster and pre-sample its private randomness independently of $J_{1:T}$. Let $\mathcal{H}_t$ be the history consisting of that randomness and $J_1, \dots, J_{t-1}$. Then $p_t$ is $\mathcal{H}_t$-measurable. Writing $q_0 = (1/n, \dots, 1/n)$, the outcome and its conditional mean are

$$y_t := \frac{q(x_t) + e_{J_t}}{2}, \qquad \mu_t = \mathbb{E}[y_t \mid \mathcal{H}_t] = \frac{q(x_t) + \mathbb{E}[e_{J_t} \mid \mathcal{H}_t]}{2} = \frac{q(x_t) + q_0}{2}. \tag{14}$$

For $j \le s$, this gives

$$\mu_{t,j} = \frac{q_j(x_t) + 1/n}{2} = \frac{1}{2}\left(a + a\frac{x_{t,j} - 1}{m-1} + \frac{1}{n}\right).$$

The grid $q([m]^s)$ lies in an affine space of dimension $s$. Scaling by $1/2$ and translating by $q_0/2$ preserves this dimension, so the visited conditional means have affine dimension at most $s$.

Then, we construct binary group functions on the context space. For $j \le s$, we compare $p_{t,j}$, the $j$-th entry of the prediction $p_t$, with $\mu_{t,j}$. To approximate $\operatorname{sign}(p_{t,j} - \mu_{t,j})$, we use the length-$m$ Walsh system, which requires $m$ to be a power of two. Let $\psi_\ell(r) = (-1)^{\langle \ell, r \rangle_2}$, where $\ell, r \in \{0, \dots, m-1\}$ and the inner product is taken modulo two between their $\log_2 m$-bit binary expansions.

Lemma 11 of Collina et al. (2026a), with approximation error $1/2$, gives a fixed set $H_m \subseteq \{1, \dots, m-1\}$ with $|H_m| \le 4C_{\mathrm{sub}}(\log m)^3$, where $C_{\mathrm{sub}}$ is a universal constant. The lemma also provides functions $b_0, b_\ell : [0,1] \to \mathbb{R}$, $\ell \in H_m$ with $\|b_0\|_\infty \le 1$ and $\sum_{\ell \in H_m} \|b_\ell\|_\infty \le \log_2 m$, fixed before play. To approximate $\operatorname{sign}(p_{t,j} - \mu_{t,j})$, we weight each Walsh function $\psi_\ell(x_{t,j} - 1)$ by $b_\ell(p_{t,j})$ and add the constant term $b_0(p_{t,j})$. For every $j \le s$, round $t$, and possible prediction $p_t \in \Delta_{n-1}$, this approximation satisfies

$$\left| b_0(p_{t,j}) + \sum_{\ell \in H_m} b_\ell(p_{t,j}) \psi_\ell(x_{t,j} - 1) - \operatorname{sign}(p_{t,j} - \mu_{t,j}) \right| \le \tfrac{1}{2}. \tag{15}$$

For each $j \le s$ and $\ell \in H_m$, split the contexts into two groups according to whether $\psi_\ell(x_j - 1)$ equals $+1$ or $-1$. Their indicator functions are

$$g^{\pm}_{j,\ell}(x) := \frac{1 \pm \psi_\ell(x_j - 1)}{2} \qquad (j \le s,\ \ell \in H_m).$$

Let $\mathcal{G}_{T,s}$ consist of these functions and the constant function $g(x) \equiv 1$. Since

$$\psi_\ell(x_j - 1) = g^{+}_{j,\ell}(x) - g^{-}_{j,\ell}(x),$$

the approximation in (15) can be expressed as a linear combination of these group functions. The groups depend only on the context, while the coefficients $b_\ell(p_{t,j})$ depend on the prediction. The bound on $|H_m|$ and the choice of $m$ give

$$|\mathcal{G}_{T,s}| \le 1 + 2s|H_m| \le CL_T^3,$$

where $C = 1 + 64C_{\mathrm{sub}}$. This completes the construction.

**Step 2. Compare predictions with conditional means.** To lower-bound calibration error, we first compare each prediction $p_t$ with the corresponding conditional mean $\mu_t$ of the outcome. We consider the expected sum of their $\ell_1$ distances:

$$\mathbb{E}\left[\sum_{t=1}^{T} \|p_t - \mu_t\|_1\right].$$

If this quantity is large, the group functions directly give the desired calibration lower bound. If it is small, we show that the learner must use many distinct prediction values. Step 3 then uses this conclusion to obtain the calibration lower bound.

Since distinct means differ by at least

$$|\mu_{t,j} - \mu_{t',j}| = \frac{a}{2(m-1)}|x_{t,j} - x_{t',j}| \geq \frac{a}{2(m-1)} =: \Delta,$$

we compare the expected total prediction distance with the spacing of the conditional means. Writing $R := k^{1/(s+2)}T^{(s+1)/(s+2)}$, we prove the following two claims.

- **Large prediction distance gives a calibration lower bound.** We show that expected total prediction distance at least $T\Delta/8$ already yields the desired rate, and we claim that

$$\mathbb{E}\left[\sum_{t\leq T}\|p_t - \mu_t\|_1\right] \leq B\,\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)], \tag{16}$$

  where $B := 4s + 8s\log_2 m + 2 \leq 32L_T$.

  If $\mathbb{E}[\sum_{t\leq T}\|p_t - \mu_t\|_1] \geq T\Delta/8$, then (16) and the spacing bound give

$$\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)] \geq \frac{T\Delta}{256L_T} \geq \frac{R}{4096(s+1)L_T} \geq \frac{(\log 2)R}{4096L_T^2},$$

  where $L_T \geq (s+1)\log 2$.

  To prove (16), we first use the Walsh approximation to bound the prediction error in each of the first $s$ coordinates. Fix $j \leq s$. By (15),

$$\left(b_0(p_{t,j}) + \sum_{\ell\in H_m} b_\ell(p_{t,j})\psi_\ell(x_{t,j}-1)\right)(p_{t,j} - \mu_{t,j}) \geq \frac{|p_{t,j} - \mu_{t,j}|}{2}.$$

  To relate this bound to calibration error, we next replace $\mu_{t,j}$ by $y_{t,j}$ in expectation. For the fixed coordinate $j$, the conditional mean $\mu_{t,j}$ has $m$ possible values. These values divide $[0,1]$ into $m+1$ intervals. For each possible value of $\mu_{t,j}$, the sign of $p_{t,j} - \mu_{t,j}$ depends only on which interval contains $p_{t,j}$. We include each possible mean value as the lower endpoint of its interval, consistently with $\mathrm{sign}(0) = 1$.

  We therefore fix $b_0$ and $b_\ell$, $\ell \in H_m$, to be constant on these intervals, making them measurable functions of the prediction coordinate. Since $p_{t,j}$ is $\mathcal{H}_t$-measurable and $x_t$ is deterministic, the multiplier is $\mathcal{H}_t$-measurable. Then, we have

$$\begin{aligned}\mathbb{E}[b_\ell(p_{t,j})\psi_\ell(x_{t,j}-1)(p_{t,j} - y_{t,j})\,|\,\mathcal{H}_t] &= b_\ell(p_{t,j})\psi_\ell(x_{t,j}-1)\big(p_{t,j} - \mathbb{E}[y_{t,j}\mid\mathcal{H}_t]\big)\\ &= b_\ell(p_{t,j})\psi_\ell(x_{t,j}-1)(p_{t,j} - \mu_{t,j}),\end{aligned} \tag{17}$$

  where we use $\mathbb{E}[y_{t,j}\mid\mathcal{H}_t] = \mu_{t,j}$ to replace $\mu_{t,j}$ by $y_{t,j}$ without changing the expected product.

  After this roundwise replacement, group together the rounds with the same prediction vector $v$. On these rounds, $p_{t,j} = v_j$, so the coefficients can be factored out as $b_0(v_j)$ and $b_\ell(v_j)$. For each $\ell$, the identity $\psi_\ell(x_{t,j}-1) = g^+_{j,\ell}(x_t) - g^-_{j,\ell}(x_t)$ bounds its contribution as follows:

$$\begin{aligned}\left|\sum_t b_\ell(p_{t,j})\psi_\ell(x_{t,j}-1)(p_{t,j} - y_{t,j})\right| &= \left|\sum_v b_\ell(v_j)\sum_{t:p_t=v}\big(g^+_{j,\ell}(x_t) - g^-_{j,\ell}(x_t)\big)(v_j - y_{t,j})\right|\\ &\leq \|b_\ell\|_\infty \sum_{\sigma\in\{+,-\}}\sum_v\left|\sum_{t:p_t=v} g^\sigma_{j,\ell}(x_t)(v_j - y_{t,j})\right|\\ &\leq \|b_\ell\|_\infty \sum_{\sigma\in\{+,-\}}\sum_v\left\|\sum_{t:p_t=v} g^\sigma_{j,\ell}(x_t)(y_t - v)\right\|_1\\ &\leq 2\|b_\ell\|_\infty\,\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T).\end{aligned}$$

Therefore, together with (17), we obtain

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[\sum_t |p_{t,j}-\mu_{t,j}|\right] &\le \mathbb{E}\left[\sum_t b_0(p_{t,j})(p_{t,j}-y_{t,j})\right] \\ &\quad + \sum_{\ell\in H_m}\mathbb{E}\left[\sum_t b_\ell(p_{t,j})\psi_\ell(x_{t,j}-1)(p_{t,j}-y_{t,j})\right] \\ &\le \left(1+2\sum_{\ell\in H_m}\|b_\ell\|_\infty\right)\mathbb{E}\Big[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)\Big] \\ &\le (1+2\log_2 m)\mathbb{E}\Big[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)\Big].\end{aligned}$$

For all $h > s+1$, the mean coordinate is fixed at $\mu_h = 1/(2n)$. Predictable signs and the constant group control these coordinates together:

$$\begin{aligned}\mathbb{E}\left[\sum_t\sum_{h>s+1}|p_{t,h}-\mu_h|\right] &= \mathbb{E}\left[\sum_v\sum_{h>s+1}\mathrm{sign}(v_h-\mu_h)\sum_{t:p_t=v}(v_h-y_{t,h})\right] \\ &\le \mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)].\end{aligned}$$

The bound remains valid when this coordinate set is empty. For $z = p_t - \mu_t$, the identity $\sum_h z_h = 0$ implies $\|z\|_1 \le 2\sum_{j\le s}|z_j| + 2\sum_{h>s+1}|z_h|$. Combining these coordinate bounds proves (16) with $B = 4s + 8s\log_2 m + 2$. To check the constants, note that the choice of $m$ and $T \ge 2^{s+2}$ imply

$$\log_2 m \le \frac{\log_2 T}{s+2} + 1 \le \frac{2\log_2 T}{s+2}.$$

Writing $t_0 = \log_2 T \ge s+2$, we obtain

$$B \le 12s + 8t_0 + 2 \le 20t_0 - 22 < 32L_T, \qquad 1 + B \le 33L_T,$$

using $20/\log 2 < 32$ and $L_T > 1$. The expectations above were taken before grouping, not conditional on the terminal bins.

- **Small prediction distance forces many prediction values.** For each realized prediction value $v$, let $n_v := |\{t : p_t = v\}|$ denote the number of rounds on which the learner predicts $v$. Step 3 will lower-bound the expected calibration error in terms of $\mathbb{E}[Q]$, where

$$Q := \sum_v \min\{n_v, \sqrt{kn_v}\}.$$

We show that

$$\mathbb{E}\left[\sum_{t\le T}\|p_t-\mu_t\|_1\right] < T\Delta/8 \quad\Longrightarrow\quad \mathbb{E}[Q] \ge \frac{R}{2\sqrt{2}}. \tag{18}$$

To prove this, we use the number of visits to each context. There are $P := \min\{m^s, T\}$ distinct visited contexts, each occurring at most $\bar{q} := \lceil T/m^s\rceil \le 2T/P$ times. When $m^s > T$, every visited context occurs once. Let $n_v^{\mathrm{near}}$ count visits with $p_t = v$ and $\|v-\mu_t\|_1 < \Delta/4$, and let $N_{\mathrm{near}} = \sum_v n_v^{\mathrm{near}}$. Every other visit has $\|p_t-\mu_t\|_1 \ge \Delta/4$, so

$$N_{\mathrm{near}} \ge T - \frac{4}{\Delta}\sum_{t\le T}\|p_t-\mu_t\|_1.$$

Two distinct grid means cannot both be within $\Delta/4$ of $v$. Hence $n_v^{\mathrm{near}} \le \bar{q}$, although no such bound is asserted for the total count $n_v$. The function $z \mapsto \min\{z, \sqrt{kz}\}$ is nondecreasing and is at least $\min\{1, \sqrt{k/\bar{q}}\}z$ for $0 \le z \le \bar{q}$. Thus

$$Q \ge \min\{1, \sqrt{k/\bar{q}}\}\, N_{\mathrm{near}}.$$

If $\mathbb{E}[\sum_{t\le T}\|p_t-\mu_t\|_1] < T\Delta/8$, then $\mathbb{E}[N_{\mathrm{near}}] \ge T/2$, and

$$\mathbb{E}[Q] \ge \frac{T}{2}\min\{1,\sqrt{k/\bar{q}}\} \ge \frac{1}{2\sqrt{2}}\min\{T,\sqrt{kTP}\} \ge \frac{R}{2\sqrt{2}}.$$

For the last inequality, $R \le T$. If $P = T$, the displayed minimum equals $T$. Otherwise $P = m^s$, and the upward rounding of $m$ gives $\sqrt{kT}\,m^{s/2} \ge \sqrt{kT}(T/k)^{s/[2(s+2)]} = R$. This proves (18), including when only part of the grid is visited.

The next step converts this lower bound on $\mathbb{E}[Q]$ into a calibration lower bound in the small-distance case.

**Step 3. Centering exposes noise on every coordinate pair.** For each prediction value $v$, define $U_v := \sum_{t:p_t=v}(\mu_t - y_t)$. The crucial feature of (14) is the exact cancellation

$$\mu_t - y_t = \frac{q_0 - e_{J_t}}{2}.$$

Thus the noise does not require coordinates separate from the grid.

Use all $k$ disjoint pairs $(2j-1, 2j)$, $j \in [k]$. Let $K_t = j$ when $J_t$ lies in pair $j$, and $K_t = 0$ for the unpaired coordinate when $n$ is odd. Define

$$n_{v,j} := |\{t : p_t = v,\ K_t = j\}|, \qquad D := \sum_{v,j}\sqrt{n_{v,j}}.$$

The current pair is independent of the prediction, with $\Pr(K_t = j \mid \mathcal{H}_t) = 2/n$. We use this independence to show

$$\mathbb{E}[D] \ge \frac{k}{n\sqrt{2}}\mathbb{E}[Q]. \tag{19}$$

Let $D_t$ be the sum defining $D$ with counts through time $t$. Before round $t$, suppose the selected bin has $z$ previous visits, of which $c_j$ selected pair $j$. Then $\sum_j c_j \le z$, allowing for the unpaired coordinate. The fresh pair law and the identity $\sqrt{c+1}-\sqrt{c} = 1/(\sqrt{c+1}+\sqrt{c})$ give

$$\begin{aligned}\mathbb{E}[D_t - D_{t-1} \mid \mathcal{H}_t] &= \frac{2}{n}\sum_{j=1}^{k}(\sqrt{c_j+1}-\sqrt{c_j}) \\ &\ge \frac{1}{n}\sum_{j=1}^{k}\frac{1}{\sqrt{c_j+1}} \ge \frac{k}{n}\sqrt{\frac{k}{z+k}}.\end{aligned}$$

The last inequality uses convexity and monotonicity of $x \mapsto (1+x)^{-1/2}$. For a bin with $n_v$ total visits, its previous-visit counts run through $0,\dots,n_v-1$. Consequently

$$\sum_{z=0}^{n_v-1}\sqrt{\frac{k}{z+k}} \ge n_v\sqrt{\frac{k}{n_v+k}} \ge \frac{1}{\sqrt{2}}\min\{n_v,\sqrt{kn_v}\}.$$

The last bound follows by considering $n_v \le k$ and $n_v \ge k$. Summing the conditional increments over $t$ and then regrouping these chronological visits proves (19). Every round has one numerical prediction label and at most one pair label. The resulting bound is in expectation, not pathwise.

At the $h$-th visit to pair $j$, let the independent fair sign $\xi_{j,h}$ select its first or second coordinate. The $j$-th pair projection of $U_v$ is therefore minus one half of the sum of these signs on visits with prediction $v$. The uniform coordinates of $q_0$ cancel in the projection, even when the two grid coordinates have different means. Because the pairs are disjoint, the sum of the absolute pair projections is at most $\|U_v\|_1$.

To bound the noise in these prediction bins, we use Theorem 4 of Collina et al. (2026a). Let $b := \min\{C_{\mathrm{rev}}, 1/16\} > 0$, where $C_{\mathrm{rev}}$ is the universal constant in that theorem. If $Z_1,\dots,Z_N$ are independent uniform signs times $h \in (0,1]$, and each label $v_t$ depends only on past signs and auxiliary information independent of the entire sign sequence, then, for $N \ge 2$,

$$\mathbb{E}\left[\sum_v\left|\sum_{t:v_t=v}Z_t\right|\right] \ge \frac{bh}{\log(N+1)}\mathbb{E}\left[\sum_v\sqrt{|\{t:v_t=v\}|}\right]. \tag{20}$$

We will check empty and singleton subsequences directly.

For each pair, expose the pair-index sequence and the signs in all other pairs. The remaining signs stay independent, and its prediction labels are determined before each remaining sign is observed. Applying (20) at amplitude $1/2$ to this subsequence, including its empty and singleton cases, yields

$$\mathbb{E}\left[\sum_v \|U_v\|_1\right] \geq \frac{b}{2L_T}\mathbb{E}[D] \geq \frac{b}{6\sqrt{2}\,L_T}\mathbb{E}[Q] \geq \frac{b}{16L_T}\mathbb{E}[Q].$$

Here the second inequality uses (19) and $k \geq n/3$. Details of Step 3 justify the conditioning and the boundary cases.

The constant group gives the pathwise comparison

$$\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T) \geq \sum_v \|U_v\|_1 - \sum_{t\leq T} \|p_t - \mu_t\|_1.$$

Taking expectations and using (16), we obtain

$$(1+B)\,\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)] \geq \frac{b}{16L_T}\mathbb{E}[Q].$$

Since $1 + B \leq 33L_T$, this proves

$$\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)] \geq \frac{b}{528L_T^2}\mathbb{E}[Q]. \tag{21}$$

**Step 4. Combine the distance and noise bounds.** Step 2 already gives a lower bound of $(\log 2)R/(4096L_T^2)$ in the large-distance case. In the small-distance case, (18) and (21) give

$$\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)] \geq \frac{bR}{1056\sqrt{2}\,L_T^2}.$$

Both cases imply $\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)] \geq bR/(4096L_T^2)$. Since $k \geq n/3 \geq n/4$ and $4^{1/(s+2)} \leq 2$, we have $R \geq R_s(T,n)/2$. Consequently, for every $s \geq 1$ and $T \geq \max\{n, 2^{s+2}\}$,

$$\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}_{T,s}}(T)] \geq \frac{b}{8192}\frac{R_s(T,n)}{L_T^2}. \tag{22}$$

**Step 5. Include the endpoint cases and complete the theorem.** First take a constant context, the constant group, and independent uniform categorical outcomes $y_t = e_{J_t}$. Their conditional mean is always $q_0$, so its affine dimension is zero. Details of Step 5 show that, for every $T \geq n$,

$$\mathbb{E}[\mathrm{MCerr}^{(1)}_{\{\mathbf{1}\}}(T)] \geq \frac{\sqrt{nT}}{2\sqrt{10}}. \tag{23}$$

This proves the stronger categorical conclusion at $s = 0$. This endpoint uses only the uniform categorical component of the construction.

For $s \geq 1$ and $T \geq 8$, set $r := \min\{s, \lfloor \log_2 T \rfloor - 2\}$. Then $1 \leq r \leq s \leq n-1$ and $T \geq 2^{r+2}$, so the already proved bound (22) applies with $r$ in place of $s$. If $r = s$, the rates coincide. If $r < s$, then $T < 2^{r+3}$ and $n \leq T$, whence

$$\frac{R_r(T,n)}{R_s(T,n)} = \left(\frac{n}{T}\right)^{1/(r+2)-1/(s+2)} \geq T^{-1/(r+2)} > 2^{-(r+3)/(r+2)} \geq \frac{1}{4}.$$

The instance has affine dimension at most $r \leq s$ and the same universal group-size bound.

For $2 \leq n \leq T < 8$, use the categorical instance in (23). For every allowed $s$,

$$\frac{\sqrt{nT}}{R_s(T,n)} = \left(\frac{n}{T}\right)^{s/[2(s+2)]} \geq \sqrt{\frac{n}{T}} \geq \sqrt{\frac{2}{7}} > \frac{1}{2}.$$

Thus all cases satisfy

$$\mathbb{E}[\mathrm{MCerr}^{(1)}_{\mathcal{G}}(T)] \geq \frac{b}{32768}\frac{R_s(T,n)}{L_T^2}.$$

Indeed $b \leq 1/16$ and $L_T > 1$ make this constant smaller than the endpoint constants. In every case the contexts, groups, and product outcome law are fixed before choosing the forecaster. Taking $c = b/32768$ and $C = 1 + 64C_{\mathrm{sub}}$ proves all claims of Theorem 2.

**Details of Step 3: Apply the noise bound to pair subsequences.** Generate the pair indices independently, with $\Pr(K_t = j) = 2/n$ for $j \in [k]$ and $\Pr(K_t = 0) = (n-2k)/n$. Independently generate fair sign arrays $(\xi_{j,h})_{h \leq T}$ for every pair $j$. At its $h$-th visit, choose $J_t = 2j-1$ for sign $+1$ and $J_t = 2j$ for sign $-1$. When $K_t = 0$, set $J_t = n$. Given the pair indices, all used sign entries are distinct and independent, so this generates exactly the independent uniform distribution of $J_{1:T}$. Let $\tau_{j,h}$ be the time of pair $j$'s $h$-th visit and $N_j = \sum_v n_{v,j}$ its total number of visits. The exact pair projection is

$$U_{v,2j-1} - U_{v,2j} = -\frac{1}{2}\sum_{h=1}^{N_j} \mathbf{1}\{p_{\tau_{j,h}} = v\}\xi_{j,h}.$$

Every draw outside the pair contributes zero to this difference.

Label each prediction value by the round on which it first appears. Fix $j$ and condition only on an event $N_j = d$ of positive probability. The learner's seed, the full pair-index sequence, and the other pairs' sign arrays form auxiliary information independent of the entire $j$-th sign array, also under this conditional law. Before $\tau_{j,h}$, every observation is determined by that auxiliary information and $\xi_{j,1}, \ldots, \xi_{j,h-1}$. The prediction label at $\tau_{j,h}$ is therefore measurable before the current target sign. For $d \geq 2$, (20) with amplitude $1/2$ gives

$$\mathbb{E}\left[\sum_v |U_{v,2j-1} - U_{v,2j}| \,\middle|\, N_j = d\right] \geq \frac{b}{2\log(d+1)}\mathbb{E}\left[\sum_v \sqrt{n_{v,j}} \,\middle|\, N_j = d\right].$$

For $d = 0$, both sides vanish. For $d = 1$, the absolute projection sum is $1/2$, and the sum of square roots is one. Thus the same bound with denominator $L_T$ holds because $b \leq 1/16 < \log 2 \leq L_T$. For $d \geq 2$, use $\log(d+1) \leq L_T$. Averaging over $N_j$, summing over pairs, and using $\sum_j |U_{v,2j-1} - U_{v,2j}| \leq \|U_v\|_1$ yields

$$\mathbb{E}\left[\sum_v \|U_v\|_1\right] \geq \frac{b}{2L_T}\mathbb{E}[D],$$

as used in Step 3.

**Details of Step 5: Prove the categorical endpoint.** Write $C_T = \mathrm{MCerr}^{(1)}_{\{\mathbf{1}\}}(T)$, $A_T = \sum_t \|p_t - q_0\|_1$, and $\phi(v)_i = \mathrm{sign}(v_i - q_{0,i})$. The predictable signs and the constant group give

$$\begin{aligned}\mathbb{E}[A_T] &= \mathbb{E}\left[\sum_t \langle \phi(p_t), p_t - y_t\rangle\right] \\ &= \mathbb{E}\left[\sum_v \left\langle \phi(v), \sum_{t:p_t=v}(p_t - y_t)\right\rangle\right] \leq \mathbb{E}[C_T].\end{aligned}$$

Again the conditional expectation is taken roundwise before grouping. For $S_T = \sum_t (y_t - q_0)$, the triangle inequality gives

$$\|S_T\|_1 \leq C_T + A_T, \qquad \mathbb{E}\|S_T\|_1 \leq 2\mathbb{E}[C_T].$$

Fix a coordinate and write $X_t = \mathbf{1}\{J_t = i\} - 1/n$, with variance $\sigma^2 = (1/n)(1-1/n)$. Its sum $Z = \sum_t X_t$ has variance $v = T\sigma^2 \geq 1/2$. The identity $\mathbb{E}X_t^4 = \sigma^2(1-3\sigma^2)$ and independence give

$$\mathbb{E}Z^4 = T\sigma^2(1-3\sigma^2) + 6\binom{T}{2}\sigma^4 = 3v^2 + v(1-6\sigma^2) \leq 5v^2.$$

Only terms with no singly occurring index survive the fourth-moment expansion. Hölder's inequality yields

$$v \le (\mathbb{E}|Z|)^{2/3}(\mathbb{E}|Z|^4)^{1/3}, \qquad \mathbb{E}|Z| \ge \frac{v^{3/2}}{(\mathbb{E}|Z|^4)^{1/2}} \ge \sqrt{v/5} \ge \sqrt{T/(10n)}.$$

Summing this last bound over coordinates gives $2\mathbb{E}[C_T] \ge \mathbb{E}\|S_T\|_1 \ge \sqrt{nT/10}$, proving (23). □

# E General global and local packing guarantees

The same algorithm admits guarantees for means near convex sets with controlled Hellinger packing numbers, even without an affine-dimension bound.

## E.1 Packing profiles and approximating families

We use the same dyadic scales. The logarithmic parameters $L, \kappa$, the smoothing weight $\beta$, the uniform prediction $q_0 = (1/n, \dots, 1/n)$, and $J_T = K_{\max}$ are fixed explicitly in Appendix F. The structural bound in Corollary 1 uses the actual counts $M_I, B_I$. We now bound those counts using global and local Hellinger packing.

We use the global and local packing numbers from Definition 2.

**Definition 3** (Approximating family)**.** *For a comparison depth $2 \le K \le K_{\max}$, an approximating family consists of nonempty closed convex sets $S_1, \dots, S_{K-1} \subseteq \Delta_d$ such that*

$$\max_{t \le T} \operatorname{dist}_H((1-\beta)\mu_t + \beta q_0, S_j) \le w_j/16 \quad (1 \le j < K).$$

*No nesting or assumption at finer scales $j \ge K$ is required. For such a family write*

$$G_K = 1 + \sum_{j=1}^{K-1} \mathsf{G}_H(S_j; w_j),$$

$$\Lambda_K(R) = \max\left\{1, \max_{\substack{1 \le j < K \\ 0 \le h \le \lceil \log_2(2/w_j) \rceil}} \mathsf{D}_H(S_j; 2^h w_j, R)\right\}.$$

For a fixed family through depth $K$, write $\Lambda = \Lambda_K(300\sqrt{\kappa})$. The sum in $G_K$ bounds the number of stored predictions across levels. The maximum in $\Lambda_K$ includes larger resolutions because boundary comparators are counted in distance annuli around their source. These geometric quantities are used only in the analysis and are not computed by the algorithm.

**A one-dimensional example.** Let $S = \Delta_1$, and write its points as $p(\theta) = (\cos^2\theta, \sin^2\theta)$, $0 \le \theta \le \pi/2$. Then $H(p(\theta), p(\phi)) = 2\sin(|\theta - \phi|/2)$, which lies between $2|\theta - \phi|/\pi$ and $|\theta - \phi|$. Consequently the global packing number is of order $1 + v^{-1}$ for $0 < v \le 1$. Inside a Hellinger ball of radius $Rv$, any two points are at distance at most $2Rv$, so their angles lie in an interval of length at most $\pi Rv$. Separation greater than $v/2$ therefore gives at most $1 + 2\pi R$ points. Thus global packing grows at finer resolution, while local packing stays bounded across scales for fixed $R$.

## E.2 General theorem and profile corollaries

**Theorem 3** (Global and local packing bounds)**.** *On one event of probability at least $1 - \delta$, simultaneously for every comparison depth and approximating family in Definition 3, with $w = w_{K-1}$,*

$$\mathrm{MCerr}^{(1)}_{\mathcal{G}}(T) \le 24 \cdot 10^6 \sqrt{\kappa} \left[ Tw + \min\left\{ G_K\sqrt{nLT}, \sqrt{nLG_KT} + \frac{nLK_{\max}G_K\Lambda_K(300\sqrt{\kappa})}{w} \right\} \right] + 2K_{\max} + 5.$$

*The deterministic bound $\mathrm{MCerr}^{(1)}_{\mathcal{G}}(T) \le 2T$ also holds. All packing quantities are used only in the analysis.*

For the following corollaries, assume that the sets $S_j$ satisfy Definition 3 at every positive scale used by the algorithm. All packing numbers are measured in Hellinger distance. The notation $\widetilde{O}$ hides at most a factor of $C[\log(en|\overline{\mathcal{G}}|(T+1)/\delta)]^2$, where $C$ is universal.

**Corollary 2** (Global packing profile)**.** *If $\mathsf{G}_H(S_j; w_j) \le A w_j^{-s}$ for all $j$, where $A \ge 1$ and $s \ge 0$, then*

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le \widetilde{O}\left(A^{1/(s+1)} n^{1/[2(s+1)]} T^{(2s+1)/(2s+2)}\right).$$

**Corollary 3** (Global and local packing profiles)**.** *Suppose the preceding global condition holds and every local profile used in $\Lambda_K(300\sqrt{\kappa})$ is at most $Bn^{q/2}$, where $B \ge 1$ and $q \ge 0$. Then*

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le \widetilde{O}\left((AB)^{1/(s+2)} n^{(1+q/2)/(s+2)} T^{(s+1)/(s+2)}\right).$$

**Corollary 4** (Polynomial local packing growth)**.** *Suppose $\mathsf{D}_H(S_j; v, R) \le H_0 R^s$ for every required $v = 2^h w_j$ and every $R \ge 1$, where $H_0 \ge 1$ and $s \ge 0$. Then*

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le \widetilde{O}\left(H_0^{2/(s+2)} n^{1/(s+2)} T^{(s+1)/(s+2)}\right).$$

### E.3 General packing-count lemma

The following lemma converts global and local packing bounds into discovery counts.

**Lemma 12** (Global and chronological annular packing)**.** *On the event of Lemma 5, for every approximating family through depth $K$ and every interval $I$ at this depth defined in Section 4.2,*

$$M_I \le G_K, \qquad B_I \le \min\{G_K,\ J_T \Lambda_K(300\sqrt{\kappa})\}.$$

*Proof.* Work on $\mathcal{E}_{\mathrm{global}}$ and fix an approximating family through depth $K$ and an interval $I$ at this depth. Throughout this proof, $P_j$ abbreviates $P_j(\max I + 1)$ for $j \ge 1$. We follow the proof of Lemma 3, using the sets $S_j$ in place of $S_\beta$ and their global and local packing numbers in place of the affine packing bounds.

First, for each $1 \le j < K$, every value stored in $P_j$ lies within distance $w_j/8$ of $S_j$. Indeed, it is the mean $z_E$ of a completed level-$\ell$ block with $\ell \ge j$. Each smoothed conditional mean is within $w_j/16$ of $S_j$, so convexity of $S_j$ and joint convexity of $H^2$ give the same bound for their weighted average $\bar{\mu}'_E$. Applying (10) at level $\ell$ yields

$$\mathrm{dist}_H(z_E, S_j) \le H(z_E, \bar{\mu}'_E) + \mathrm{dist}_H(\bar{\mu}'_E, S_j) \le \frac{w_\ell + w_j}{16} \le \frac{w_j}{8}.$$

This argument uses only $S_j$, not nesting of the approximating sets, and includes values inherited from earlier depths or inserted on the final round of $I$.

**Step 1. Count prediction values $M_I$.** For each $a \in P_j$, choose $\tilde{a} \in S_j$ with $H(a, \tilde{a}) \le w_j/8$. The insertion rule preserves separation after this replacement:

$$H(\tilde{a}, \tilde{b}) \ge H(a, b) - H(a, \tilde{a}) - H(b, \tilde{b}) > \frac{3w_j}{4} > \frac{w_j}{2} \qquad (a \ne b).$$

Hence $|P_j| \le \mathsf{G}_H(S_j; w_j)$, and summing over levels gives

$$M_I = \left| P_0 \cup \bigcup_{j=1}^{K-1} P_j \right| \le 1 + \sum_{j=1}^{K-1} |P_j| \le 1 + \sum_{j=1}^{K-1} \mathsf{G}_H(S_j; w_j) = G_K.$$

**Step 2. Count boundary comparators in chronological order.** Fix a group, source $a$, and positive level $j < K$, and list the completed boundary blocks' comparators $q_1, q_2, \ldots \in P_j$ in closure order. Each earlier comparator is available throughout every later block, so it cannot be chosen again for a boundary block. Thus the comparators are distinct, and their number is at most $|P_j| \le G_K$.

To obtain the local packing bound, reuse the chronological comparison from Step 2 of the proof of Lemma 3. That comparison uses only (3), $H(q_m, z_m) \le w_j$, and $N_m \ge m_j = 1024nL/w_j^2$,

where $z_m, N_m$ are the mean and weight of the block associated with $q_m$. It does not require an affine approximating set and gives

$$H(q_m, q_i) \geq \frac{H(q_m, a)}{A} - w_j, \qquad i < m, \qquad A = 16\sqrt{\kappa}.$$

This strengthens the separation for comparators far from $a$. Choose representatives $\tilde{q}_i \in S_j$ within distance $w_j/8$ of $q_i$, and use the same near and far regions as in that proof.

- **Near region:** $H(q_i, a) \leq 4Aw_j$**.** Step 1 gives separation greater than $w_j/2$ between representatives, and
$$H(\tilde{q}_i, a) \leq \frac{w_j}{8} + 4Aw_j \leq 5Aw_j.$$
- **Far regions:** $4A2^h w_j < H(q_i, a) \leq 8A2^h w_j$**,** $h = 0, 1, \ldots$**.** For two comparators in the same region, order them as $i < m$. The chronological comparison and the replacement error give
$$H(\tilde{q}_m, \tilde{q}_i) > (4 \cdot 2^h - 1)w_j - \frac{w_j}{4} > 2 \cdot 2^h w_j,$$
$$H(\tilde{q}_i, a) \leq \frac{w_j}{8} + 8A2^h w_j \leq 9A2^h w_j.$$

In each region the representatives are more than $v/2$ apart and lie within distance $300\sqrt{\kappa}\, v$ of $a$, with $v = w_j$ in the near region and $v = 2^h w_j$ in the $h$-th far region, since $9A < 300\sqrt{\kappa}$. The definition of $\Lambda_K$ therefore bounds each region by $\Lambda_K(300\sqrt{\kappa})$. If $h \geq j$, the far-region condition would require

$$H(q_i, a) > 4A2^h w_j = 8A2^{h-j} \geq 8A > \sqrt{2},$$

which is impossible. Thus only $h = 0, \ldots, j-1$ can contribute far regions, all among the resolutions defining $\Lambda_K$. Including the near region gives at most $j + 1 \leq K \leq K_{\max} = J_T$ regions. Combining this count with the distinct-comparator bound and taking the maximum over groups, sources, and positive levels yields

$$B_I \leq \min\{G_K,\ J_T \Lambda_K(300\sqrt{\kappa})\}.$$

Both bounds are at least one, so they also cover the minimum value of 1 in the definition of $B_I$. □

### E.4 Proof of Theorem 3

*Proof.* Work on $\mathcal{E}_{\text{global}}$. We first derive two interval bounds, then choose a comparison width, show that the algorithm cannot move to a coarser scale, and sum the interval errors.

**Step 1. Bound the error on each interval.** For an interval $I$ at depth $K \geq 2$, write $w = w_{K-1}$. Start from Lemma 11 and substitute the counts from Lemma 10.

*(i) Using only the prediction count.* For a dyadic set of positive $x$, splitting at $\sqrt{a/b}$ and summing each geometric tail gives $\sum_x \min(a/x, bx) \leq 4\sqrt{ab}$. Using $Q_j(I) \leq \min(8M_I^2, |I|/m_j)$ and $m_j = 1024nL/w_j^2$ therefore bounds the lower-piece term by

$$48\sqrt{8 \cdot 1024}\sqrt{\kappa}\, M_I \sqrt{nL|I|} < 4400\sqrt{\kappa}\, M_I \sqrt{nL|I|}.$$

The unfinished-zero term is at most $32\sqrt{nLM_I|I|}$, since $m_0 = 256nL$. Using $M_I, \kappa \geq 1$ combines these two terms with coefficient $4400 + 32 = 4432$, while leaving the sampling term separate:

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \leq 6\sqrt{\kappa}\,|I|w + 4432\sqrt{\kappa}\, M_I \sqrt{nL|I|} + 4\sqrt{nLM_I T} + 2\beta|I|.$$

*(ii) Combining with the boundary-block count.* The proof of Corollary 1 in Appendix B.7 gives the stronger estimate

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \leq 6\sqrt{\kappa}\,|I|w + 98816\sqrt{\kappa}\,\frac{nLM_I B_I}{w} + 4\sqrt{nLM_I T} + 2\beta|I|.$$

Both estimates hold on the same event for every permitted endpoint, so taking their minimum also gives the stronger two-branch form

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \leq 10^6\sqrt{\kappa}\left[|I|w + \min\left\{M_I\sqrt{nL|I|}, \frac{nLM_I B_I}{w}\right\}\right] + 4\sqrt{nLM_I T} + 2\beta|I|.$$

For the subsequent controller analysis, the same estimates also imply the horizon-based form

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \le 10^6\sqrt{\kappa}\left[Tw + \min\left\{M_I\sqrt{nLT},\ \frac{nLM_IB_I}{w} + \sqrt{nLM_IT}\right\}\right] + 2\beta|I|.$$

Indeed, $|I| \le T$ and $M_I, \kappa \ge 1$ bound the first estimate's lower-piece and sampling terms by $4436\sqrt{\kappa}\,M_I\sqrt{nLT}$. For the second estimate, $98816 < 10^6$ and $4 \le 10^6\sqrt{\kappa}$ give the displayed second branch.

**Step 2. Choose a comparison width.** For a fixed comparison family through depth $K$, let $F$ be either $G_K\sqrt{nLT}$ or $\sqrt{nLG_KT} + nLJ_TG_K\Lambda_K(300\sqrt{\kappa})/w$. By Lemma 12, $M_I \le G_K$ and $B_I \le J_T\Lambda_K(300\sqrt{\kappa})$ at that depth. Substituting these counts into the horizon-based bound from Step 1 gives $\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \le 10^6\sqrt{\kappa}(Tw + F) + 2\beta|I|$ for either choice of $F$. At every coarser depth the profile counts and inverse-width term cannot increase. Let $x = \max(w, F/T)$. If $x > 1$, the deterministic bound $2T$ proves the assertion. Otherwise choose the smallest dyadic resolution $v \in \{w_0, \dots, w_{K_{\max}-1}\}$ with $v \ge x$, so $v < 2x$.

**Step 3. The algorithm never moves to a coarser scale.** At this depth every prefix obeys

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(I) \le 10^6\sqrt{\kappa}(Tv + F) + 1/T \le 2\cdot 10^6\sqrt{\kappa}\,Tv + 1/T < 3\cdot 10^6\sqrt{\kappa}\,Tv.$$

The strict inequality follows from $Tv \ge Tw_{K_{\max}-1} > 1/2$. Thus the algorithm cannot cross this depth, even though it never computes $F$. Only the sets at scales below this comparison depth were used.

**Step 4. Sum the interval errors.** At an earlier failed endpoint the residual is at most its threshold plus two, since one sampled residual per group changes by norm at most two per round. The endpoint belongs to the old interval and no new interval is created at $T$. Visited widths double, so their sum is less than twice the final width, which is at most $v$. The triangle inequality across intervals gives

$$\mathrm{MCerr}_{\mathcal{G}}^{(1)}(T) \le 6\cdot 10^6\sqrt{\kappa}\,Tv + 2K_{\max} \le 24\cdot 10^6\sqrt{\kappa}(Tw + F) + 2J_T + 1.$$

The argument holds for both choices on the same event, proving the minimum. Substituting $J_T = K_{\max}$ and using $2J_T + 1 \le 2K_{\max} + 5$ gives the stated form of the bound. □

### E.5 Proofs of the packing corollaries

*Proof of Corollaries 2–4.* For the global case, $G_K \le AJ_Tw^{-s}$. Balance at $w_* = [AJ_T\sqrt{nL/T}]^{1/(s+1)}$. For the global–local case, the inverse term is at most $ABn^{1+q/2}LJ_T^2w^{-(s+1)}$, so balance at $w_* = [ABn^{1+q/2}LJ_T^2/T]^{1/(s+2)}$. Its geometric mean with $Tw$ dominates the sampling term because $Bn^{q/2}J_T \ge 1$. For polynomial local growth, diameter $\sqrt{2}$ gives $\mathsf{G}_H \le H_0(\sqrt{2}/w_j)^s$, and $\Lambda_K \le H_0(300\sqrt{\kappa})^s$. The inverse coefficient becomes $H_0^2nLJ_T^2(300\sqrt{2\kappa})^s$, and the same balance applies. In all cases $w_* \ge T^{-1/2}$. If $w_* \le 1$, round upward to a dyadic resolution less than $2w_*$, which cannot enlarge an inverse-width term. Otherwise use $2T \le 2Tw_*$. Writing $Z = \log(en|\overline{\mathcal{G}}|(T+1)/\delta)$, the definitions give $\kappa \le 8Z$, $L \le 40Z$, and $J_T \le 6Z$. The three logarithmic exponents are respectively $1/2 + 3/[2(s+1)]$, $1/2 + 3/(s+2)$, and $(s+4)/(s+2)$, all at most two. The constant base in the third bound has exponent $s/(s+2) < 1$. These substitutions give exactly the three displayed rates, and absorb $2J_T + 1$ with a universal constant. They include $s = 0$ and $T = 1$. □

## F Choice of Universal Constants

Recall $K_{\max} = 2 + \lceil \log_2 T \rceil$, $w_j = 2^{1-j}$, $q_0 = (1/n, \dots, 1/n)$, and $A_{\max} = 1 + |\overline{\mathcal{G}}|T$. The algorithm's smoothing weight and coordinate floor are $\beta = 1/(2T^2)$ and $\epsilon = \beta/n$. Set

$$\bar{D} = \log(1/\epsilon), \qquad \kappa = \bar{D} + 2, \qquad \eta = \frac{\bar{D}}{8\kappa}, \qquad J_T = K_{\max},$$

and

$$L = 2 + \log\left(\frac{8n|\overline{\mathcal{G}}|A_{\max}^2(T+1)^2(K_{\max}-1)}{\delta}\right), \qquad m_j = \frac{1024nL}{w_j^2}.$$

All these quantities are fixed before play. For $Z = \log(en|\overline{\mathcal{G}}|(T+1)/\delta)$, their definitions give $\kappa \le 8Z$, $L \le 40Z$, and $J_T \le 6Z$. Set $C_{\mathrm{test}} = 3 \cdot 10^6$, giving the controller threshold $C_{\mathrm{test}}\sqrt{\kappa}\,T w_{K-1}$.

The coordinate floor gives $0 \le \ell(a, y') \le \log(1/\epsilon) = \bar{D}$, as required by the normalized comparison updates.