

Optimal Lower Bounds for Online Multicalibration

Natalie Collina

Jiuyao Lu

Georgy Noarov

Aaron Roth

University of Pennsylvania

January 9, 2026

Abstract

We prove tight lower bounds for online multicalibration, establishing an information-theoretic separation from marginal calibration.

In the general setting where group functions can depend on both context and the learner’s predictions, we prove an $\Omega(T^{2/3})$ lower bound on expected multicalibration error using just three disjoint binary groups. This matches the upper bounds of [Noarov et al. \[2025\]](#) up to logarithmic factors and exceeds the $O(T^{2/3-\varepsilon})$ upper bound for marginal calibration [\[Dagan et al., 2025\]](#), thereby separating the two problems.

We then turn to lower bounds for the more difficult case of group functions that may depend on context but not on the learner’s predictions. In this case, we establish an $\tilde{\Omega}(T^{2/3})$ lower bound for online multicalibration via a $\Theta(T)$ -sized group family constructed using orthogonal function systems, again matching upper bounds up to logarithmic factors.

Contents

1	Introduction	1
1.1	Our Results	2
1.2	Proof Overviews	4
1.2.1	Lower Bound for the General Case	5
1.2.2	Lower Bound for Prediction Independent Groups	7
2	Model and Definitions	9
3	Optimal Lower Bound for the General Case	10
3.1	The Hard Instance	11
3.1.1	Defining $\mathcal{D}_{T,m}$	11
3.1.2	Defining G	12
3.2	Probabilistic Tools	12
3.3	Proof of Theorem 1	14
3.3.1	Big Deviations are Punished by g_1 and g_2	15
3.3.2	Many η -Honest Rounds are Punished by g_3	18
3.4	Context-wise tradeoff between big deviations and noise	19
3.4.1	Lower Bound for the “Honest” Group	22
3.4.2	Putting it All Together	22
4	Optimal Lower Bound for Prediction-Independent Groups	24
4.1	The Hard Distribution: Time-Augmented Contexts.	24
4.2	The Hard Group Family.	25
4.2.1	Hadamard and Walsh systems.	25
4.2.2	The Complete Group Family	27
4.3	Global Walsh Groups Enforce ℓ_1 -Truthfulness	29
4.3.1	Walsh Expansion of Discrete Threshold Signs on the Grid	30
4.3.2	Global Walsh Groups Enforce ℓ_1 -Truthfulness	31
4.4	Multicalibration Requires Diverse Predictions	33
4.4.1	ℓ_1 Time-Quantization and Diverse Predictions	33
4.5	Controlling Bias Using Hadamard Groups	36
4.6	Controlling the Noise Contribution under Adaptive Bucketing	37
4.6.1	Proof Roadmap	38
4.6.2	Proof of Theorem 3: Step 1	39
4.6.3	Proof of Theorem 3: Step 2	40
4.6.4	Finishing the proof of Theorem 3	42
4.6.5	Applying Theorem 3 to the Noise Contribution	42
4.7	Proof of Theorem 2	43
5	Discussion	46
A	Reducing Constant Sized Families of Binary Prediction Independent Groups to Marginal Calibration	49
B	An Oracle Lower Bound for Better Black-Box Reductions	52
B.1	Context-blind oracles and proper reductions	52
B.2	Hard instance and a logarithmic-size group family	54
B.3	Main theorem	54
B.4	Proof of Theorem 5	55

1 Introduction

Online calibration A sequence of predictions p^1, \dots, p^T is *calibrated* to a sequence of outcomes y^1, \dots, y^T if, informally, the average of the predictions equals the average of the outcomes, even *conditional* on the value of the prediction [Dawid, 1982]. To measure the deviation from perfect calibration, one can define the cumulative empirical bias conditional on a prediction $v \in \mathbb{R}$ as $B_T(v) = \sum_{t:p^t=v} (p^t - y^t)$. The classical mis-calibration measure known as *expected calibration error (ECE)* sums the magnitude of the empirical bias conditional on each prediction:

$$\text{Err}_T = \sum_{v \in \{p^1, \dots, p^T\}} |B_T(v)|.$$

In a seminal result, Foster and Vohra [1998] showed that there exists a randomized algorithm able to generate predictions that are guaranteed to have expected calibration error scaling as $o(T)$ for arbitrary/adversarially selected sequences y^1, \dots, y^T . The optimal *rate* at which calibration error can be guaranteed has been a long-standing open question, which has seen recent partial progress. A long-standing upper bound established that it was possible to obtain calibration error scaling as $O(T^{2/3})$ [Foster and Vohra, 1998, Hart, 2025, Abernethy et al., 2011]. For many years no lower bound better than $\Omega(T^{0.5})$ was known, until Qiao and Valiant [2021] proved a lower bound of $\Omega(T^{0.528})$. The current state of the art, due to Dagan et al. [2025], establishes that the optimal rate for calibration is between $\Omega(T^{0.54389})$ and $O(T^{2/3-\varepsilon})$ for some (extremely small) constant $\varepsilon > 0$. Dagan et al. [2025]’s result was a breakthrough for giving the first upper bound improvement showing that the long-standing $T^{2/3}$ rate was not optimal for marginal calibration.

Online multicalibration Calibration is on its own a weak guarantee in that it marginalizes over the entire sequence, which substantially limits its applicability in contextual prediction settings. But it is possible to give stronger guarantees, asking for calibration not just marginally, but simultaneously on many different subsequences or weightings of the data that can be defined both by external context and the predictions themselves [Dawid, 1985, Lehrer, 2003, Sandroni et al., 2003].

A modern CS formulation of this idea is called *multicalibration*, introduced by Hébert-Johnson et al. [2018]. Multicalibration reweights the residuals of the predictions by “group functions”, which are simply mappings $g : X \times \mathbb{R} \rightarrow [0, 1]$ from any pair $(x, v) = (\text{context}, \text{learner’s prediction})$ to a bounded weight $g(x, v)$. When $g(x, v)$ is independent of v , we will refer to such a group as *prediction-independent*. The group- and prediction-conditional cumulative empirical bias is defined as $B_T(v, g) = \sum_{t:p^t=v} g(x^t, p^t)(p^t - y^t)$, and the group-conditional calibration error is then given by $\text{Err}_T(g) = \sum_{v \in \{p^1, \dots, p^T\}} |B_T(v, g)|$. The multicalibration error with respect to a collection of group functions G is defined as

$$\text{MCerr}_T(G) = \max_{g \in G} \text{Err}_T(g).$$

Multicalibration and related guarantees have found many applications in recent years, from learning in a loss-function agnostic manner [Gopalan et al., 2022a, 2023a] to strengthening complexity theoretic constructions [Casacuberta et al., 2024, Dwork and Tankala, 2025] to low complexity algorithms for distributed information aggregation [Collina et al., 2025, 2026]. Moreover, similar techniques to those that have been used to derive algorithms guaranteeing marginal calibration in sequential adversarial settings have also been adapted to multicalibration, including methods based on multi-objective optimization and Blackwell approachability [Gupta et al., 2022, Lee et al., 2022,

Noarov et al., 2025, Haghtalab et al., 2023], swap regret minimization [Globus-Harris et al., 2023, Gopalan et al., 2023b, Garg et al., 2024], and defensive forecasting [Perdomo and Recht, 2025].

What are the optimal multicalibration rates? Just as for marginal calibration, the minimax online multicalibration rate has remained a difficult open challenge. Recently, Noarov et al. [2025] (and later Ghuge et al. [2025]) established that online multicalibration can be obtained at the rate $\tilde{O}(T^{2/3}\sqrt{\log |G|})$; very recently, Hu et al. [2025] gave corresponding *oracle-efficient* rates (not just for means, but for any elicitable property; c.f. Noarov and Roth [2023]). In fact, given their benign $O(\sqrt{\log G})$ dependence on $|G|$, these algorithms guarantee $\tilde{O}(T^{2/3})$ rates even for $\text{poly}(T)$ -sized group families G .

However, to date no lower bound for online multicalibration has been obtained, beyond the $\Omega(T^{0.54389})$ lower bound inherited from the easier problem of marginal calibration [Dagan et al., 2025]. Nor have any $O(T^{2/3-\epsilon})$ -rate multicalibration algorithms been derived for any $\epsilon > 0$.

It is natural to wonder whether the answer to the online multicalibration rate problem is simply the (as yet unknown) minimax marginal calibration rate. Indeed, consider any *constant-sized* collection G of *prediction-independent* groups¹ (i.e., groups that only depend on context). Then it is possible to obtain multicalibration at the rate of marginal calibration: one can just instantiate, in parallel, $2^{|G|}$ copies of a minimax-optimal marginal calibration algorithm for all regions in the Venn diagram partition corresponding to G (i.e., for all possible group intersection patterns).

However, this rate-preserving reduction breaks for group collections G that are more complex than just described. First, if the groups in G depend on the learner’s predictions, this invalidates the reduction mechanism: on any given round, the set of active groups (and thus the set of active copies of the marginal calibration algorithm) will not be determined until the learner makes the prediction. Second, if the size of G is not constant but instead grows even logarithmically with T , then combining all of the copies’ guarantees will incur overhead that will destroy the optimal marginal calibration rate.

Therefore, we have two challenging regimes in which to pin down the complexity of online multicalibration and to establish whether it is still exactly as hard as marginal calibration or strictly harder: (1) Prediction-dependent group collections; and (2) prediction-independent group collections whose size grows with T . To summarize, in this paper our goal will be to answer the fundamental questions:

What are the minimax rates for sequential multicalibration?

Is sequential multicalibration a strictly harder problem than sequential marginal calibration?

1.1 Our Results

We answer the above questions in a strong sense: We prove $\tilde{\Omega}(T^{2/3})$ online multicalibration lower bounds that match (up to logarithmic factors) the existing $\tilde{O}(T^{2/3})$ upper bounds — both in the prediction-independent case, and in the prediction-dependent case with growing-in- T group family size. Therefore, our results *strictly separate the complexity of marginal calibration from that of multicalibration*, while also separating two regimes of complexity for multicalibration —

¹For convenience further assume the groups in G are binary-valued, so that the Venn diagram corresponding to G is well-defined for the purposes of this reduction. We note that our lower bounds hold even for binary groups, thus showing that binary groups are as hard as real-valued ones in the minimax sense.

Setting	Groups	Upper bounds	Lower bounds
Marginal calibration	none	$O(T^{2/3-\varepsilon})$ (for some $\varepsilon > 0$) Dagan et al. (2025)	$\Omega(T^{0.54389})$ Dagan et al. (2025)
Multicalibration (general)	prediction-dependent $g(x, v)$	$\tilde{O}(T^{2/3}\sqrt{\log G })$ Noarov et al. (2025)	$\Omega(T^{2/3})$ (even $ G = 3$ disjoint binary groups)
Multicalibration (restricted)	prediction-independent $g(x)$, $ G = O(1)$	Rate preserving reduction to marginal calibration	(no separation possible from marginal calibration)
Multicalibration (restricted)	prediction-independent $g(x)$, $ G = \Theta(T)$	$\tilde{O}(T^{2/3}\sqrt{\log G })$ Noarov et al. (2025)	$\tilde{\Omega}(T^{2/3})$

Figure 1: **Summary of regimes and rates.** We study online adversarial multicalibration. For general prediction-dependent groups $g(x, v)$ we prove an optimal $\Omega(T^{2/3})$ lower bound, separating multicalibration from marginal calibration. For prediction-independent groups $g(x)$, constant-sized families reduce to marginal calibration up to a $2^{|G|}$ factor, precluding a separation from marginal calibration. For a group family of size $|G| = \Theta(T)$ we again prove an optimal $\tilde{\Omega}(T^{2/3})$ lower bound.

prediction-dependent and *prediction-independent* groups — that existing upper bounds have treated identically.

The work of [Hébert-Johnson et al. \[2018\]](#) defined multicalibration in terms of what we will call binary prediction-independent groups, that depend on the context but not the prediction, and map to a binary range: $g : X \times \mathbb{R} \rightarrow \{0, 1\}$ such that for all $v, v' \in \mathbb{R}$ and for all $x \in X$, $g(x, v) = g(x, v')$. Subsequent work generalized the notion of groups to allow them to be weighting functions with range $[0, 1]$ rather than binary valued, and to explicitly depend on predictions p^t [[Gopalan et al., 2022b](#), [Kim et al., 2022](#), [Deng et al., 2023](#)] — see also [Kakade and Foster \[2008\]](#) and [Sandroni et al. \[2003\]](#) for earlier work using similar grouping functions. [Sandroni et al. \[2003\]](#) called prediction-dependent binary groups “forecast based checking rules”. The algorithms that were developed for online multicalibration [[Gupta et al., 2022](#), [Noarov et al., 2025](#), [Haghtalab et al., 2023](#)] are all able to support this general notion of group functions at the same rates as they support the special case of binary prediction-independent group functions. As foreshadowed by the above discussion, our lower bounds identify a distinction between the general case and prediction independent groupings.

1. In Section 3 we show an optimal $\Omega(T^{2/3})$ lower bound for multicalibration in the general case, when groups can be defined in terms of predictions. This matches (up to logarithmic factors) the rate obtained by existing efficient algorithms [[Noarov et al., 2025](#), [Ghuge et al., 2025](#), [Hu et al., 2025](#)], and so establishes the optimal statistical rate for multicalibration in the general case. Because of the $O(T^{2/3-\varepsilon})$ upper bound for marginal calibration established by [Dagan et al. \[2025\]](#) it also formally separates the statistical complexity of multicalibration from marginal calibration. Our lower bound instance is realized by just 3 binary valued groups that are also disjoint (such that exactly one is active on any given round). Therefore, this lower bound is driven neither by the complexity nor by the intersectionality of the group family, but rather only by their prediction-dependent nature.

2. We observe in Appendix A that no similar separation is possible for constant-sized families of binary groups if one further restricts them to be prediction independent. Indeed, there is an extremely simple reduction from multicalibration to marginal calibration for prediction independent groups that, given binary groups G , instantiates a marginal calibration algorithm for each of the $2^{|G|}$ possible *intersection patterns* of groups. At each round t exactly one group intersection pattern is realized and (since the groups are prediction independent) we can route the round to the appropriate marginal calibration algorithm. This gives multicalibration at the best rate obtainable for marginal calibration, up to a $2^{|G|}$ factor (slightly more nuanced bounds are possible, and we give these in Appendix A). If $|G|$ is constant valued (independent of T), then this exponential-in- $|G|$ blowup is also only constant valued, and does not affect the asymptotic rate. Hence any separation between marginal calibration and multicalibration for the special case of prediction independent groups must depend on group families with cardinality $|G|$ growing as a function of T . We note that *upper bounds* [Gupta et al., 2022, Noarov et al., 2025, Ghuge et al., 2025, Hu et al., 2025] depend only logarithmically on $|G|$, so even families of size $|G| = \text{poly}(T)$ incur only logarithmic overhead.
3. In Section 4, we give an optimal $\tilde{\Omega}(T^{2/3})$ lower bound for multicalibration even restricting to prediction independent binary groups, using a group family of size $|G| = O(T)$. This once again matches existing upper bounds up to logarithmic factors [Noarov et al., 2025, Ghuge et al., 2025, Hu et al., 2025] and separates multicalibration (with group families with cardinality depending polynomially on T) from marginal calibration [Dagan et al., 2025]. This is the technically most challenging result of the paper.
4. Finally, in Appendix B, we formalize a natural notion of a “proper” oracle reduction from multicalibration to marginal calibration which captures standard reduction techniques in learning theory like aggregation with no regret learning algorithms and sleeping experts. We provide an oracle lower bound for group families of size $|G| = \Theta(\log T)$, showing that any ‘proper’ black-box reduction to marginal calibration obtaining non-trivial multicalibration rates must use exponentially many oracles (in the cardinality of the group family), showing a sense in which our reduction from Appendix A is tight. This serves as a barrier to extending the constant-sized group family upper bounds from Appendix A to group families of cardinality scaling logarithmically with T .

We note that all of our lower bound constructions use only *binary* groups, which also establishes that these are already as hard as arbitrary weighted groups (which the upper bounds support).

1.2 Proof Overviews

We now sketch our lower bound constructions and their analyses. Since it turns out that rates of $\tilde{\Theta}(T^{2/3})$ are the “right answer” for multicalibration, it is helpful to understand how these rates arise in upper bounds. This is most easily understood through the “minimax” lens of Hart [2025] in which the order-of-play of the learner and the adversary are reversed in the analysis using the minimax theorem. In the reversed order-of-play, the adversary first commits to a (possibly adaptive) strategy mapping histories to distributions over outcomes, and the learner has knowledge of this strategy before it must make predictions. One option for the learner in this reversed order of play is the “honest” strategy that at every round predicts $p^t = \mathbb{E}[y^t]$ which is feasible, given that in this order of play, the learner knows the adversary’s strategy. This is not a good strategy on its own, as the

calibration error metric sums the magnitude of the empirical bias across all prediction values the learner uses, and it might be that $\mathbb{E}[y^t]$ takes on a distinct value for each t ; in this case there would be no cancellations and the learner’s calibration error would scale linearly with T . But the learner could *round* their prediction $p^t \approx \mathbb{E}[y^t]$ to the nearest multiple of $1/m$, which would introduce bias at most $1/m$ at each round and cumulatively at most T/m across all T rounds. Standard (anti)-concentration arguments establish that if there is a value v that the learner predicts k times, then the empirical sum of the labels on the rounds in which v is predicted will differ from its expectation by roughly $\approx \sqrt{k}$. The “rounded” honest learner uses at most m different prediction values, and the worst-case for the learner is if they are all used equally frequently: $k \approx T/m$. In this case the summed noise magnitude of the learner’s predictions scales as $m\sqrt{k} \approx \sqrt{mT}$. Picking m to trade this noise term off against the bias term of T/m results in a $\tilde{O}(T^{2/3})$ upper bound. This style of argument applies both to marginal calibration and multicalibration, because the “rounded honest” strategy obtains bounds of this form simultaneously on any subsequence of rounds. However, it does not give a lower bound because there may be a strategy for the learner that obtains better calibration error than “honesty” by cleverly setting up cancellations — indeed, this is exactly what [Dagan et al. \[2025\]](#) show in the case of marginal calibration. At a very high level, our goal is to rule out that “dishonest” strategies can be beneficial. We note in passing a conceptual similarity to recent work on designing truthful calibration measures [[Haghtalab et al., 2024](#), [Qiao and Zhao, 2025](#), [Hartline et al., 2025](#)], although our settings are incomparable.

Both lower bounds we prove in Sections 3 and 4 share a common pattern. The lower bound instances are both *oblivious*/non-adaptive sequences of context/label pairs (x^t, y^t) such that:

1. The labels y^t are independent random variables with $\mathbb{E}[y^t] = x^t$, and
2. The contexts themselves x^t are uniformly spread out in a grid in $[1/4, 3/4]$.

Instances like this make it possible for the learner to make “honest predictions” of $p^t = \mathbb{E}[y^t] = x^t$, because the label mean is communicated to the learner through the context. However, as discussed, if the learner were to engage in this “honest” prediction strategy, their predictions (although unbiased) would incur high error because of noise. Our group constructions are designed to punish dishonest strategies, and thus to force the learner into the high-error “honest” regime. For prediction dependent groups, there is a conceptually straightforward way to do this—although there are a number of technical obstacles to carrying the idea through formally. For prediction independent groups this is more complex.

1.2.1 Lower Bound for the General Case

We use a Bernoulli environment in which contexts x^t cycle over a fixed grid in $[1/4, 3/4]$, and labels are drawn as

$$y^t \sim \text{Bernoulli}(x^t)$$

independently across time. We choose a grid size $m \approx T^{1/3}$ and a small margin parameter $\eta \approx \sqrt{m/T}$. The construction uses only three disjoint binary prediction-dependent groups that partition the prediction space according to whether the learner made a prediction that was approximately “honest”, or was dishonest either by predicting substantially above the label mean (an “overshoot”)

or by predicting substantially below the label mean (an “undershoot”):

$$\begin{aligned} g_1(x, v) &= \mathbf{1}[v \geq x + \eta] \quad (\text{large overshoots}), \\ g_2(x, v) &= \mathbf{1}[v \leq x - \eta] \quad (\text{large undershoots}), \\ g_3(x, v) &= \mathbf{1}[|v - x| < \eta] \quad (\text{approximately honest predictions}). \end{aligned}$$

The proof shows that any algorithm must incur multicalibration error $\Omega(T^{2/3})$ on at least one of these groups, by splitting into two complementary cases.

1. **Partition rounds into big deviations and “honest” rounds.** For a fixed algorithm and realization, define

$$\begin{aligned} \text{big-deviation rounds } B &:= \{t : |p^t - x^t| \geq \eta\}, \\ \text{honest rounds } H &:= \{t : |p^t - x^t| < \eta\}, \end{aligned}$$

with counts $B_T = |B|$ and $H_T = |H| = T - B_T$.

On B , exactly one of g_1 or g_2 is active. On H , only g_3 is active. We chose η small enough relative to the grid spacing so that the intervals $(x - \eta, x + \eta)$ around distinct grid points are disjoint; this will later let us localize the g_3 error by context.

2. **Many big deviations force large error on g_1 or g_2 .** The first part of our analysis shows that if the algorithm predicts dishonestly often, it pays linearly in multicalibration error.

Let $r^t := p^t - x^t$. On big-deviation rounds, either $r^t \geq \eta$ or $r^t \leq -\eta$. For each fixed prediction v , the expected contribution of the overshoot rounds ($r^t \geq \eta$) to the calibration bias of g_1 is

$$\mathbb{E}[B_T(v, g_1)] = \mathbb{E}\left[\sum_{t:p^t=v} \mathbf{1}[p^t \geq x^t + \eta](p^t - y^t)\right] = \mathbb{E}\left[\sum_{t:p^t=v} \mathbf{1}[p^t \geq x^t + \eta]r^t\right],$$

since $\mathbb{E}[y^t \mid x^t, p^t] = x^t$. Whenever $p^t \geq x^t + \eta$, we have $r^t \geq \eta$, so we get expected positive bias $\geq \eta$ on those rounds; a symmetric argument holds for g_2 . Summing over v , this implies

$$\mathbb{E}[\text{MCerr}_T] \gtrsim \eta \mathbb{E}[B_T].$$

So, if B_T is large, we are already done: either g_1 or g_2 must have large calibration error.

3. **Few big deviations force many “honest” rounds per context.** The complementary case is when the algorithm mostly stays close to honest: If $\mathbb{E}[B_T]$ is not large, then $\mathbb{E}[H_T]$ is large. Contexts cycle on a regular grid, and so each grid point x appears about T/m times.

We refine the partition by context to avoid cancellations: For each grid point x , let H_x be the honest rounds with that context, and $n_x := |H_x|$. Similarly define B_x and b_x for big deviations at context x , with $T_x = n_x + b_x$ the total number of times x appears. If B_T is small, the structure of the instance forces many contexts x to have a substantial number of honest rounds n_x ; this is where we will extract noise using a martingale argument.

4. **Honest rounds accumulate uncontrollable noise on g_3 .** On honest rounds for a fixed context x , we decompose the calibration error for g_3 into:

- **Noise:** $N_x := \sum_{t \in H_x} (x^t - y^t)$, a sum of centered Bernoulli deviations around x^t ; and
- **Bias:** $R_x := \sum_{t \in H_x} (p^t - x^t)$, which is small in magnitude because $|p^t - x^t| < \eta$ on H_x .

We show a structural lemma (Lemma 4) that the calibration error on g_3 satisfies

$$\sum_v |B_T(v, g_3)| \geq \sum_x |N_x| - \sum_x |R_x|.$$

By construction, we know that the bias is small on honest rounds: $\sum_x |R_x| \leq \eta H_T$ which will be negligible at our choice of η .

The core probabilistic step is a martingale-transform lower bound applied context-wise: on the rounds where any fixed context x occurs, the noise terms $Z_t := x^t - y^t$ form a martingale difference sequence with variance bounded away from zero. Using a martingale moment argument, we show (Proposition 1) that whenever a nontrivial fraction of the occurrences of x are honest rounds (i.e., many indices in H_x), the noise magnitude $\mathbb{E}[|N_x|] \gtrsim \sqrt{n_x}$ is large.

Summing over contexts and using that the T_x 's are all $\Theta(T/m)$, we obtain a tradeoff: either there are many big deviations (large B_T), or

$$\mathbb{E}\left[\sum_x |N_x|\right] \gtrsim \sqrt{mT},$$

which in turn forces the calibration error on g_3 to be of that order up to the small bias term. Choosing parameters to optimize the tradeoff yields our optimal $\Omega(T^{2/3})$ lower bound.

1.2.2 Lower Bound for Prediction Independent Groups

At a high level, our approach to proving lower bounds while restricting to *prediction-independent* groups mirrors our approach for the general case: (1) we use an oblivious stochastic instance in which label means are revealed by the context, so that “honest” predictions are feasible but suffer high calibration error because of noise, and (2) we construct groups to punish “dishonest” strategies that try to cancel noise by grouping predictions; and (3) we decompose the multicalibration error into a *bias* component (controlled by group constraints) and an unavoidable *noise* component, which we show must be large.

Unlike the prediction dependent case, we cannot directly detect and constrain deviations from honesty, since our groups must be defined only through context (not prediction). Instead, we define a family of groups constructed from orthonormal bases and show that obtaining low multicalibration error on these groups obligates the learner to make predictions that are on average close to honest in an ℓ_1 sense.

Concretely, we consider contexts $c^t = (x^t, t)$ where x^t encodes the label mean and t encodes the time index. Label means x^t cycle through an m -point grid in $[1/4, 3/4]$ with $m = \Theta(T^{1/3})$, and outcomes are generated as

$$y^t = x^t + \frac{\xi^t}{4}, \quad \xi^t \in \{\pm 1\} \text{ i.i.d.}$$

We define a prediction-independent group family G consisting of: (i) the constant group g_{all} (enforcing marginal calibration), (ii) $O(m)$ “global” Walsh half-groups $g_\ell^{\text{Wal}, \pm}$ on the m -point grid, whose differences yield signed Walsh functionals $w_\ell = g_\ell^{\text{Wal}, +} - g_\ell^{\text{Wal}, -}$, which form an orthonormal

basis, and (iii) $O(T)$ blockwise Hadamard half-groups $g_{a,j}^\pm$ supported on disjoint time blocks, whose differences yield signed Hadamard functionals $h_{a,j} = g_{a,j}^+ - g_{a,j}^-$ which form an orthonormal basis on each of the time blocks. We show that:

1. Global Walsh groups enforce ℓ_1 -truthfulness. Let

$$A := \sum_{t \leq T} |p^t - x^t|$$

be the total ℓ_1 deviation from honest predictions. Using a Walsh expansion of the threshold-sign pattern $\text{sign}(p^t - x^t)$ on the mean grid, we show that if the forecaster has small multi-calibration error on the global Walsh family, then A must be small in expectation:

$$\mathbb{E}[A] \leq O(\log(m)) \cdot \mathbb{E}[\text{MCerr}_T'(G)].$$

Intuitively: systematic deviations from x^t would either be detected by marginal calibration (g_{all}) or be witnessed by some Walsh functional.

2. ℓ_1 -truthfulness forces many moderately-used prediction values. Let $n_v := |\{t \leq T : p^t = v\}|$ and define

$$N := \sum_{v \in \mathcal{V}_T} \sqrt{n_v}.$$

An honest forecaster on this instance would spread its forecasts equally across all relevant values v . We show that small ℓ_1 -distance to honesty A prevents the forecaster from concentrating mass on only a few prediction values, and similarly forces spread out predictions in the sense that:

$$N \gtrsim \frac{T}{\sqrt{A + T/m}}.$$

Since x^t traverses a grid of size $m = \Theta(T^{1/3})$ and $T/m = \Theta(T^{2/3})$, this implies that when A is small we must have $N \gtrsim \sqrt{mT} \approx T^{2/3}$.

3. Buckets decompose into bias and noise. We decompose

$$p^t - y^t = (p^t - x^t) + (x^t - y^t) =: \delta_t + Z_t.$$

The first term δ_t is a bias term determined by the algorithm, and the second term Z_t is the noise coming from the randomized labels. We partition time into disjoint blocks J_a . Using blockwise Hadamard signs $\psi_{a,j}$, we form signed bucket sums for each block of time J_a and each prediction value v :

$$D_v^{(a,j)} := \sum_{t \in J_a : p^t = v} \psi_{a,j}(\cdot) \delta_t, \quad N_v^{(a,j)} := \sum_{t \in J_a : p^t = v} \psi_{a,j}(\cdot) Z_t,$$

so that the calibration error of the signed Hadamard functional $h_{a,j}$ decomposes as

$$\sum_v |D_v^{(a,j)} + N_v^{(a,j)}| \gtrsim (\text{noise on that block}) - (\text{bias on that block}).$$

4. **A martingale argument lower-bounds the noise.** The signed noise terms $Z_t = x^t - y^t$ are i.i.d. Rademachers (scaled by 1/4). We analyze an arbitrary adaptive bucketing strategy that, at each time, chooses a bucket based on the past noise realizations. Via a potential-function argument for the bucket sums and a decomposition of the noise random walk into excursions away from zero, we prove an “adaptive noise bucketing” theorem: up to logarithmic factors,

$$\mathbb{E} \left[\sum_v |N_v^{(a,j)}| \right] \gtrsim \sum_v \sqrt{n_{v,a}},$$

where $n_{v,a}$ is the number of times prediction value v appears in block J_a . In other words, no matter how the algorithm routes the noise, the total signed noise magnitude at block (a,j) must be large whenever the bucket counts on that block are large.

5. **Hadamard groups upper bound the bias.** The bias sequence (δ_t) is arbitrary, but we can view it blockwise in a Hadamard basis. Orthogonality of the block Hadamard system implies a Parseval identity: when we average the squared bias coefficients $(D_v^{(a,j)})^2$ over all Hadamard functions on a block, we recover exactly the block’s total squared bias $E_a = \sum_{t \in J_a} \delta_t^2$, which is exactly the (blockwise) squared error of the predictions to the honest predictions. From this we derive that, on average over j , the ℓ_1 -mass $\sum_v |D_v^{(a,j)}|$ is at most $\tilde{O}(\sqrt{N_a E_a})$, where $N_a = \sum_v \sqrt{n_{v,a}}$. Combining this with the global lower bound on N that we derive under low ℓ_1 -distance to honesty there must be at least one Hadamard functional $h_{a,j}$ for which:

- the noise term $\sum_v |N_v^{(a,j)}|$ is large (by the martingale bucketing theorem and large N_a), while
- the bias term $\sum_v |D_v^{(a,j)}|$ is comparatively small (by the Hadamard averaging and the global ℓ_1 -truthfulness).

Intuitively, the global constraint on distance to honesty renders the bias vector compressible—it cannot correlate strongly with many orthogonal Hadamard directions simultaneously. In contrast, the random noise is incompressible and accumulates significant magnitude in every direction; thus, by averaging over the basis, we ensure the existence of a group where the incompressible noise overwhelms the bias. Combining: either some global Walsh group already has large calibration error (if “distance to honesty” is large), or there exists a Hadamard functional $h_{a,j}$ whose noise dominates its bias, forcing large $\text{Err}_{T'}(h_{a,j})$. Using that each $h_{a,j}$ is the difference of two groups in G then yields

$$\mathbb{E}[\text{MCerr}_{T'}(G)] \geq \tilde{\Omega}(T^{2/3}),$$

matching upper bounds up to polylogarithmic factors for $|G| = \Theta(T)$.

2 Model and Definitions

Fix a time horizon $T \in \mathbb{N}$. On each round $t = 1, \dots, T$:

1. A context x^t in a context space X is revealed.
2. The prediction algorithm outputs a distribution P^t on $[0, 1]$.

3. An outcome $y^t \in [0, 1]$ is selected (by the adversary/environment).
4. A prediction $p^t \in [0, 1]$ is drawn from P^t .

We allow the algorithm to be adaptive: P^t can be any (possibly randomized) function of $(x^1, y^1, p^1, \dots, x^{t-1}, y^{t-1}, p^{t-1}, x^t)$. In all of our lower bounds, the adversary will be oblivious — the context/label sequence is selected independently of the interaction with the prediction algorithm.

Definition 1 (Group functions). A *group function* is a map $g : X \times [0, 1] \rightarrow [0, 1]$. Given a finite set G of group functions, we will measure calibration error separately on each group. Throughout, we use G to denote a finite family of group functions and write $|G|$ for its cardinality.

Definition 2 (Binary and prediction-independent groups). A group function $g : X \times [0, 1] \rightarrow [0, 1]$ is *binary-valued* if $g(x, v) \in \{0, 1\}$ for all $(x, v) \in X \times [0, 1]$. We say g is *prediction-independent* if there exists a function $h : X \rightarrow [0, 1]$ such that $g(x, v) = h(x)$ for all (x, v) . In this case we may identify g with h and write $g(x)$ instead of $g(x, v)$.

We sometimes refer to general group functions $g : X \times [0, 1] \rightarrow [0, 1]$ that may depend on the prediction value v as *prediction-dependent* groups. All of the lower bounds we prove in this paper use only binary groups.

Definition 3 (Empirical Bias and multicalibration error). Given a sequence $(x^t, p^t, y^t)_{t=1}^T$ and a group g , the *empirical bias* at prediction value $v \in [0, 1]$ is

$$B_T(v, g) = \sum_{t=1}^T \mathbf{1}[p^t = v] g(x^t, p^t) (p^t - y^t).$$

Let $V_T := \{p^t : t = 1, \dots, T\}$ be the (finite) set of prediction values actually used. The expected calibration error for group g is

$$\text{Err}_T(g) := \sum_{v \in V_T} |B_T(v, g)|.$$

Given a finite family G of groups, the *expected multicalibration error* at time T is

$$\text{MCerr}_T(G) := \max_{g \in G} \text{Err}_T(g) = \max_{g \in G} \sum_{v \in V_T} |B_T(v, g)|.$$

When the group family is clear from context we abbreviate the multicalibration error as MCerr_T .

Randomness arises both from the algorithm (if it randomizes) and from the environment. All expectations $\mathbb{E}[\cdot]$ are taken with respect to this joint randomness.

3 Optimal Lower Bound for the General Case

In this section, we give a lower bound instance consisting of three binary prediction-dependent groups, showing that any algorithm must obtain multicalibration error over these groups scaling as $\Omega(T^{2/3})$. This matches the upper bound of [Noarov et al. \[2025\]](#) up to log factors.

Proof overview. The hard instance reveals contexts x^t that cycle through a grid of $m = \Theta(T^{1/3})$ values, with labels drawn as $y^t \sim \text{Bernoulli}(x^t)$. The key insight is that prediction-dependent groups can directly detect when the learner deviates from “honest” predictions $p^t = x^t$:

- Groups g_1 and g_2 activate when predictions overshoot or undershoot the context by more than η . Any such “big deviation” incurs expected bias of at least η , so many big deviations yield large error on g_1 or g_2 .
- Group g_3 activates on “ η -honest” rounds where $|p^t - x^t| < \eta$. On these rounds, predictions are approximately honest, so calibration error is driven by the inherent noise $x^t - y^t$. A martingale argument shows this noise accumulates to $\Omega(\sqrt{mT}) = \Omega(T^{2/3})$.

Either the algorithm makes many big deviations (punished by g_1, g_2) or mostly honest predictions (punished by g_3), yielding $\Omega(T^{2/3})$ error in both cases.

Theorem 1 (Prediction-dependent lower bound). *Let $(\mathcal{D}_{T,m}, G)$ be the hard instance defined in this section. There exists a constant $c > 0$ and $T_0 \in \mathbb{N}$ such that for all $T \geq T_0$, and for any (possibly randomized) prediction algorithm A :*

$$\mathbb{E}_{\mathcal{D}_{T,m}}[\text{MCerr}_T(G)] \geq cT^{2/3}$$

3.1 The Hard Instance

3.1.1 Defining $\mathcal{D}_{T,m}$

First, we will define the hard distribution over contexts and labels, $\mathcal{D}_{T,m}$. For some $T \geq 1$ and $m \geq 8$, define grid points

$$z_j := \frac{j}{m}, \quad j = 1, \dots, m-1.$$

We restrict attention to the “interior” grid points

$$J := \left\{ j \in \{1, \dots, m-1\} : z_j \in \left[\frac{1}{4}, \frac{3}{4}\right] \right\},$$

and set

$$X_0 := \{x_j : j \in J\}, \quad x_j := z_j.$$

For $m \geq 4$ we have the uniform bound

$$|J| = |\{j \in \{1, \dots, m-1\} : j/m \in [1/4, 3/4]\}| \geq \left\lfloor \frac{3m}{4} \right\rfloor - \left\lceil \frac{m}{4} \right\rceil + 1 \geq \frac{m-1}{2}.$$

In particular, for all $m \geq 4$, $|J| \geq \frac{3}{8} \cdot m$.

Let $m_0 := |J|$. We will only use the contexts in X_0 .

Definition 4 (Hard distribution $\mathcal{D}_{T,m}$). Fix T, m and X_0 as above. Define $(x^t, y^t)_{t=1}^T$ as follows:

- Contexts: x^t cycles through X_0 in round-robin order. Formally, fix any enumeration $(x^{(1)}, \dots, x^{(m_0)})$ of X_0 , and set

$$x^t := x^{(k)}, \quad \text{where } k \equiv t \pmod{m_0}, \quad k \in \{1, \dots, m_0\}.$$

Thus each $x^{(k)}$ appears either $\lfloor T/m_0 \rfloor$ or $\lceil T/m_0 \rceil$ times.

- Labels: given the context (x^t) , draw y^1, \dots, y^T independently with

$$y^t \sim \text{Bernoulli}(x^t)$$

We denote by $\mathcal{D}_{T,m}$ the joint distribution of $(x^t, y^t)_{t=1}^T$ constructed above. For our purposes, the order of the contexts will not matter: the important properties of this distribution are that each context is approximately equally frequent, and that it encodes the label mean at each round.

We observe a useful property about this distribution. For each t , define $Z_t := x^t - y^t$, the residuals of “honest predictions”. Then $(Z_t)_{t=1}^T$ are independent, mean zero, and have nontrivial variance:

$$\mathbb{E}[Z_t | x^t] = 0, \quad \text{Var}(Z_t | x^t) = x^t(1 - x^t) \in [\frac{3}{16}, \frac{1}{4}]$$

because $x^t \in [1/4, 3/4]$ for all t .

3.1.2 Defining G

We now define the family of disjoint groups $G = \{g_1, g_2, g_3\}$ used for the lower bound. The groups are designed to create a dilemma for the forecaster: deviating from honest predictions is detected by g_1 and g_2 , while staying honest exposes the forecaster to noise accumulation on g_3 .

Let $\eta > 0$ be a threshold parameter (to be chosen later):

$$\begin{aligned} g_1(x, v) &:= \mathbf{1}[v \geq x + \eta], \\ g_2(x, v) &:= \mathbf{1}[v \leq x - \eta], \\ g_3(x, v) &:= \mathbf{1}[|v - x| < \eta]. \end{aligned}$$

Thus:

- g_1 is active on rounds where the prediction overshoots the context by at least η ;
- g_2 is active on rounds where the prediction undershoots by at least η ;
- g_3 is active on rounds where the prediction is η -close to the context.

η will be set equal to $\delta \sqrt{\frac{m}{T}}$, for a carefully selected constant δ .

3.2 Probabilistic Tools

The main technical tool we need is a lower bound on the deviation of a “filtered” martingale. When analyzing the η -honest rounds (group g_3), we must show that the noise terms $Z_t = x^t - y^t$ accumulate to large magnitude even when summed only over an adaptively-chosen subset of rounds. The key challenge is that the forecaster’s predictions—and hence which rounds are η -honest—can depend on past noise realizations, potentially allowing cancellations.

Proposition 1 below shows this cannot help much: as long as at least a constant fraction of rounds are included (in expectation), the filtered sum still has $\Omega(\sqrt{L})$ expected magnitude.

We now record some basic properties of a martingale difference sequence (Z_t) that arise from the Bernoulli environment we use in our lower bound. Let $(x^t)_{t=1}^T \subset [1/4, 3/4]$, and given (x^t) let y^1, \dots, y^T be independent with $y^t \sim \text{Bernoulli}(x^t)$ for each t . $Z_t := x^t - y^t$. Consequently,

$$x^t - y^t \in \left[-\frac{3}{4}, -\frac{1}{4} \right] \cup \left[\frac{1}{4}, \frac{3}{4} \right] \quad \text{and hence} \quad \frac{1}{4} \leq |Z_t| \leq \frac{3}{4}$$

for every t .

The contexts (x^t) are deterministic (fixed independently of the history, as the adversary is non-adaptive/oblivious), and the labels y^1, \dots, y^T are independent with $y^t \sim \text{Bernoulli}(x^t)$. Let

$$\mathcal{F}_t := \sigma(y^1, \dots, y^t), \quad t = 0, 1, \dots, T,$$

with the convention that \mathcal{F}_0 is the trivial σ -algebra. Then $Z_t = x^t - y^t$ is \mathcal{F}_t -measurable and

$$\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = x^t - \mathbb{E}[y^t | \mathcal{F}_{t-1}] = x^t - \mathbb{E}[y^t] = x^t - x^t = 0,$$

since y^t is independent of (y^1, \dots, y^{t-1}) and has mean x^t . Thus (Z_t, \mathcal{F}_t) is a martingale difference sequence. Moreover, for a fixed prediction algorithm, each prediction p^t is a measurable function of $(x^1, \dots, x^t, y^1, \dots, y^{t-1})$. Because the contexts (x^t) are deterministic under $\mathcal{D}_{T,m}$, this implies p^t is \mathcal{F}_{t-1} -measurable for every t .

Multicalibration error adds up the magnitude of *empirical bias* the algorithm has obtained over various subsequences of the data, and those subsequences can be defined by the predictions of the algorithm itself, which in turn can depend on the history of the sequence in arbitrary ways. To reason about this we will use the following deviation bound for a martingale transform $N = \sum_{t=1}^L I_t Z_t$, where (Z_t) is a martingale difference sequence and the predictable indicators (I_t) select a dense subset of times (in expectation at least αL).

Proposition 1 (Dense martingale transform deviation). *Fix constants $0 < \sigma \leq 1$ and $\alpha \in (0, 1]$. Let $(Z_t)_{t=1}^L$ be a sequence of real-valued random variables adapted to a filtration $(\mathcal{F}_t)_{t=0}^L$ such that*

$$\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0, \quad \mathbb{E}[Z_t^2 | \mathcal{F}_{t-1}] \geq \sigma^2, \quad |Z_t| \leq 1 \quad \text{almost surely for all } t = 1, \dots, L. \quad (1)$$

Let $(I_t)_{t=1}^L$ be any predictable $\{0, 1\}$ -valued sequence, i.e. I_t is \mathcal{F}_{t-1} -measurable for each t , and define

$$N := \sum_{t=1}^L I_t Z_t, \quad n := \sum_{t=1}^L I_t.$$

If $\mathbb{E}[n] \geq \alpha L$, then there exists a constant $c_{\sigma,\alpha} > 0$ (depending only on σ and α) such that

$$\mathbb{E}|N| \geq c_{\sigma,\alpha} \sqrt{L}.$$

Proof. Define the martingale $(M_t)_{t=0}^L$ by $M_0 = 0$ and $M_t := \sum_{s=1}^t I_s Z_s$, $t = 1, \dots, L$.

Since I_s is \mathcal{F}_{s-1} -measurable and (1) gives $\mathbb{E}[Z_s | \mathcal{F}_{s-1}] = 0$, we have

$$\mathbb{E}[M_t | \mathcal{F}_{t-1}] = M_{t-1} + \mathbb{E}[I_t Z_t | \mathcal{F}_{t-1}] = M_{t-1} + I_t \mathbb{E}[Z_t | \mathcal{F}_{t-1}] = M_{t-1},$$

so (M_t, \mathcal{F}_t) is a martingale. Its increments are $D_t = I_t Z_t$, and its predictable quadratic variation is

$$\langle M \rangle_L := \sum_{t=1}^L \mathbb{E}[D_t^2 | \mathcal{F}_{t-1}] = \sum_{t=1}^L I_t \mathbb{E}[Z_t^2 | \mathcal{F}_{t-1}].$$

By (1), whenever $I_t = 1$ we have $\mathbb{E}[Z_t^2 | \mathcal{F}_{t-1}] \geq \sigma^2$, while if $I_t = 0$ the corresponding term is zero. Therefore, pathwise

$$\langle M \rangle_L \geq \sigma^2 \sum_{t=1}^L I_t = \sigma^2 n.$$

Taking expectations yields the identity $\mathbb{E}[M_L^2] = \mathbb{E}[\langle M \rangle_L]$, since the cross terms in $M_L^2 = (\sum_{t=1}^L D_t)^2$ vanish by $\mathbb{E}[D_t | \mathcal{F}_{t-1}] = 0$. Thus using the assumption $\mathbb{E}[n] \geq \alpha L$,

$$\mathbb{E}[M_L^2] = \mathbb{E}[\langle M \rangle_L] \geq \sigma^2 \mathbb{E}[n] \geq \sigma^2 \alpha L. \quad (2)$$

On the other hand, since $|Z_t| \leq 1$ we have $|D_t| \leq |I_t||Z_t| \leq 1$ almost surely. Also $D_t^2 \leq I_t$, so $\langle M \rangle_L \leq \sum_{t=1}^L I_t = n \leq L$ almost surely. Since $|M_L| \leq \sup_{0 \leq t \leq L} |M_t|$, we have

$$\mathbb{E}[M_L^4] \leq \mathbb{E}\left[\sup_{0 \leq t \leq L} |M_t|^4\right].$$

We will now make use of the Burkholder–Rosenthal inequality.

Lemma 1 (Burkholder–Rosenthal Inequality [Burkholder, 1973], Theorem 21.1). *Let $(M_t)_{t=0}^T$ be a martingale with increments $D_t = M_t - M_{t-1}$ and predictable quadratic variation $\langle M \rangle_T = \sum_{t=1}^T \mathbb{E}[D_t^2 | \mathcal{F}_{t-1}]$. For any $p \geq 2$, there exists a constant C_p depending only on p such that:*

$$\mathbb{E}\left[\sup_{0 \leq t \leq T} |M_t|^p\right] \leq C_p \left(\mathbb{E}\left[\langle M \rangle_T^{p/2}\right] + \mathbb{E}\left[\sup_{1 \leq t \leq T} |D_t|^p\right] \right).$$

Applying the Burkholder–Rosenthal inequality to the above with $p = 4$ gives, assuming $L \geq 1$, that

$$\mathbb{E}[M_L^4] \leq C_4 \left(\mathbb{E}[\langle M \rangle_L^2] + \mathbb{E}\left[\sup_{1 \leq t \leq L} |D_t|^4\right] \right) \leq C_4 (L^2 + 1) \leq 2C_4 L^2. \quad (3)$$

To finish the proof, define the random variable $X := M_L^2$. From (2) we have $\mathbb{E}[X] = \mathbb{E}[M_L^2] \geq \sigma^2 \alpha L$, and from (3) we have $\mathbb{E}[X^2] = \mathbb{E}[M_L^4] \leq 2C_4 L^2$. Now, consider the event $\{X \geq \frac{1}{2} \mathbb{E}[X]\}$. On this event, we have:

$$|M_L| = \sqrt{X} \geq \sqrt{\frac{1}{2} \mathbb{E}[X]} \geq \sqrt{\frac{1}{2} \sigma^2 \alpha L} = \sigma \sqrt{\frac{\alpha}{2}} \sqrt{L}.$$

Recall the Paley-Zygmund inequality: for any r.v. $\tilde{X} \geq 0$ with $0 < \mathbb{E}[\tilde{X}^2] < \infty$, it holds for any $\theta \in (0, 1)$ that $\mathbb{P}(\tilde{X} \geq \theta \mathbb{E}[\tilde{X}]) \geq (1 - \theta)^2 \frac{\mathbb{E}[\tilde{X}]^2}{\mathbb{E}[\tilde{X}^2]}$. Applying this to X with $\theta = 1/2$, we get

$$\mathbb{P}\left(X \geq \frac{1}{2} \mathbb{E}[X]\right) \geq (1 - \frac{1}{2})^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \geq \frac{1}{4} \frac{(\sigma^2 \alpha L)^2}{2C_4 L^2} =: p_0 > 0,$$

where p_0 depends only on σ and α . Hence, the proposition holds for $c_{\sigma, \alpha} := \sigma p_0 \sqrt{\alpha/2}$, since:

$$\mathbb{E}|N| = \mathbb{E}|M_L| \geq \sigma \sqrt{\frac{\alpha}{2}} \sqrt{L} \cdot \mathbb{P}\left(X \geq \frac{1}{2} \mathbb{E}[X]\right) \geq \sigma \sqrt{\frac{\alpha}{2}} \sqrt{L} \cdot p_0. \quad \square$$

3.3 Proof of Theorem 1

We are now ready to prove our lower bound for prediction-dependent groups, using the instance $(\mathcal{D}_{T,m}, G)$ defined above.

Given an algorithm and a realization of $(x^t, p^t)_{t=1}^T$, we will partition the rounds as follows.

Definition 5 (Big deviations and η -honest rounds). Define the index sets

$$\begin{aligned} B &:= \{t : |p^t - x^t| \geq \eta\}, \\ H &:= \{t : |p^t - x^t| < \eta\}, \end{aligned}$$

and denote their sizes by

$$B_T := |B|, \quad H_T := |H|.$$

We refer to $t \in B$ as *big-deviation rounds* and $t \in H$ as *η -honest rounds*.

Note that $B_T + H_T = T$ holds pathwise.

The proof of Theorem 1 reduces to two main components which we will combine at the end:

1. Big-deviation rounds incur large error under g_1 or g_2 (Section 3.3.1).
2. Many η -honest rounds incur large error under g_3 (Section 3.3.2).

3.3.1 Big Deviations are Punished by g_1 and g_2

In this section we prove that big deviations (rounds t with $|p^t - x^t| \geq \eta$) necessarily incur large expected error on g_1 or g_2 .

We first note that since our lower bound instance is fixed and non-adaptive, it suffices to consider deterministic predictors.

Lemma 2 (Reduction to deterministic predictors). *Fix T, m and the distribution $\mathcal{D}_{T,m}$. Let A be any (possibly randomized) prediction algorithm. Then there exists a deterministic algorithm A' such that*

$$\mathbb{E}_{\mathcal{D}_{T,m}}[\text{MCerr}_T(A')] \leq \mathbb{E}_{\mathcal{D}_{T,m}, A}[\text{MCerr}_T],$$

where the first expectation is over $\mathcal{D}_{T,m}$ only (since A' is deterministic), and the second expectation is over both the randomness of A and $\mathcal{D}_{T,m}$.

Proof. We model all of the algorithm's internal randomization (including any sampling of p^t from an internal distribution P^t) by an independent random seed R . For each outcome of R , the resulting algorithm A_R produces predictions p^t as deterministic functions of the history $(x^1, y^1, \dots, x^{t-1}, y^{t-1}, x^t)$. The seed R is independent of $(x^t, y^t)_{t=1}^T$ (as the lower bound instance is fixed and non-adaptive), so we can view (x^t, y^t, R) as the environment's sample space. Then

$$\mathbb{E}_{\mathcal{D}_{T,m}, A}[\text{MCerr}_T] = \mathbb{E}_R [\mathbb{E}_{\mathcal{D}_{T,m}}[\text{MCerr}_T(A_R)]]$$

is an average of the quantities $\mathbb{E}_{\mathcal{D}_{T,m}}[\text{MCerr}_T(A_R)]$ over all outcomes R . Therefore there exists at least one outcome R' such that

$$\mathbb{E}_{\mathcal{D}_{T,m}}[\text{MCerr}_T(A_{R'})] \leq \mathbb{E}_{\mathcal{D}_{T,m}, A}[\text{MCerr}_T].$$

Taking $A' := A_{R'}$ yields the claim. \square

Thus it suffices to fix an arbitrary deterministic algorithm and analyze its error.

Fix a deterministic prediction algorithm and $\mathcal{D}_{T,m}$. For each t , p^t is a deterministic function of $(x^1, y^1, \dots, x^{t-1}, y^{t-1}, x^t)$. Let B_T be the (random) number of big-deviation rounds:

$$B_T := |\{t : |p^t - x^t| \geq \eta\}| = |B|.$$

First we show that the multicalibration error of any algorithm grows linearly with the number of big deviations, witnessed by either g_1 or g_2 .

Lemma 3. *For any deterministic prediction algorithm and any $\eta > 0$, under $\mathcal{D}_{T,m}$ we have*

$$\mathbb{E}[\text{MCerr}_T] \geq \frac{\eta}{2} \mathbb{E}[B_T].$$

Proof. Write $r^t := p^t - x^t$ to denote the round t deviation from “honest” prediction. Define the indicators

$$I_t^+ := \mathbf{1}[p^t \geq x^t + \eta], \quad I_t^- := \mathbf{1}[p^t \leq x^t - \eta].$$

to indicate whether round t represented a big deviation in either the positive or negative direction. Then $|p^t - x^t| \geq \eta$ implies $I_t^+ + I_t^- = 1$. Hence

$$B_T = \sum_{t=1}^T (I_t^+ + I_t^-).$$

Consider group g_1 . For a fixed $v \in [0, 1]$,

$$B_T(v, g_1) = \sum_{t=1}^T \mathbf{1}[p^t = v] g_1(x^t, p^t) (p^t - y^t) = \sum_{t=1}^T \mathbf{1}[p^t = v] I_t^+ (p^t - y^t).$$

For each t , let $\mathcal{F}_t := \sigma(x^1, y^1, \dots, x^{t-1}, y^{t-1}, x^t, p^t)$ be the history available just before y^t is revealed. Then $\mathbf{1}[p^t = v]$ and I_t^+ are \mathcal{F}_t -measurable, and under $\mathcal{D}_{T,m}$ the label y^t satisfies

$$\mathbb{E}[y^t \mid \mathcal{F}_t] = \mathbb{E}[y^t \mid x^t] = x^t,$$

because y^t is drawn independently of the past given x^t . Thus

$$\mathbb{E}[p^t - y^t \mid \mathcal{F}_t] = p^t - x^t = r^t.$$

Using the tower property and linearity of expectation,

$$\begin{aligned} \mathbb{E}[B_T(v, g_1)] &= \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}[p^t = v] I_t^+ (p^t - y^t)\right] \\ &= \sum_{t=1}^T \mathbb{E}[\mathbf{1}[p^t = v] I_t^+ \mathbb{E}[p^t - y^t \mid \mathcal{F}_t]] \\ &= \sum_{t=1}^T \mathbb{E}[\mathbf{1}[p^t = v] I_t^+ r^t]. \end{aligned}$$

Whenever $I_t^+ = 1$, we have $r^t = p^t - x^t \geq \eta$, so

$$\mathbb{E}[B_T(v, g_1)] \geq \eta \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}[p^t = v] I_t^+\right].$$

In particular, $\mathbb{E}[B_T(v, g_1)] \geq 0$, so Jensen's inequality gives

$$\mathbb{E}[|B_T(v, g_1)|] \geq |\mathbb{E}[B_T(v, g_1)]| = \mathbb{E}[B_T(v, g_1)].$$

Summing over v and using $\sum_{v \in V_T} \mathbf{1}[p^t = v] = 1$,

$$\begin{aligned} \mathbb{E}\left[\sum_{v \in V_T} |B_T(v, g_1)|\right] &\geq \sum_{v \in V_T} \mathbb{E}[B_T(v, g_1)] \\ &\geq \eta \mathbb{E}\left[\sum_{t=1}^T I_t^+\right]. \end{aligned}$$

Equivalently, for g_2 , we have that

$$\mathbb{E}[B_T(v, g_2)] = \sum_{t=1}^T \mathbb{E}[\mathbf{1}[p^t = v] I_t^- r^t].$$

Whenever $I_t^- = 1$, we have $r^t \leq -\eta$, so

$$\mathbb{E}[B_T(v, g_2)] \leq -\eta \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}[p^t = v] I_t^-\right].$$

In particular, $\mathbb{E}[B_T(v, g_2)] \leq 0$, so

$$\mathbb{E}[|B_T(v, g_2)|] \geq |\mathbb{E}[B_T(v, g_2)]| = -\mathbb{E}[B_T(v, g_2)].$$

Summing over v and using $\sum_{v \in V_T} \mathbf{1}[p^t = v] = 1$,

$$\begin{aligned} \mathbb{E}\left[\sum_{v \in V_T} |B_T(v, g_2)|\right] &\geq -\sum_{v \in V_T} \mathbb{E}[B_T(v, g_2)] \\ &\geq \eta \mathbb{E}\left[\sum_{t=1}^T I_t^-\right]. \end{aligned}$$

Combine the two:

$$\mathbb{E}\left[\sum_{v \in V_T} |B_T(v, g_1)| + \sum_{v \in V_T} |B_T(v, g_2)|\right] \geq \eta \mathbb{E}\left[\sum_{t=1}^T (I_t^+ + I_t^-)\right] = \eta \mathbb{E}[B_T].$$

Since

$$\text{MCerr}_T \geq \max\left\{\sum_v |B_T(v, g_1)|, \sum_v |B_T(v, g_2)|\right\} \geq \frac{1}{2} \left(\sum_v |B_T(v, g_1)| + \sum_v |B_T(v, g_2)|\right),$$

we obtain

$$\mathbb{E}[\text{MCerr}_T] \geq \frac{1}{2} \mathbb{E}\left[\sum_v |B_T(v, g_1)| + \sum_v |B_T(v, g_2)|\right] \geq \frac{\eta}{2} \mathbb{E}[B_T],$$

as claimed. \square

3.3.2 Many η -Honest Rounds are Punished by g_3

We now show that if the algorithm does not make many big deviations (i.e., if the number B_T of big-deviation rounds is not too large), then it must accumulate large error on group g_3 , which records the calibration error in the η -honest rounds.

Again fix a deterministic prediction algorithm and $\mathcal{D}_{T,m}$. Recall that $H = \{t : |p^t - x^t| < \eta\}$ and $B = \{t : |p^t - x^t| \geq \eta\}$, with $H_T = |H|$ and $B_T = |B| = T - H_T$.

For each context $x \in X_0$, we denote the rounds in which the context was x and the prediction was η -honest (respectively a big deviation) as:

$$H_x := \{t \in H : x^t = x\}, \quad n_x := |H_x|,$$

and

$$B_x := \{t : x^t = x, |p^t - x^t| \geq \eta\}, \quad b_x := |B_x|.$$

Thus $T_x := n_x + b_x$ is the (deterministic) number of times the context x appears in the sequence $(x^t)_{t=1}^T$. Per the definition of $\mathcal{D}_{T,m}$, each T_x is either $\lfloor T/m_0 \rfloor$ or $\lceil T/m_0 \rceil$, so there exist constants $c_{\text{occ}}, C_{\text{occ}} > 0$ such that

$$c_{\text{occ}} \frac{T}{m_0} \leq T_x \leq C_{\text{occ}} \frac{T}{m_0} \quad \text{for all } x \in X_0. \quad (4)$$

We also have

$$B_T = \sum_{x \in X_0} b_x, \quad H_T = \sum_{x \in X_0} n_x, \quad \sum_{x \in X_0} T_x = T.$$

On η -honest rounds for context x , we define the noise and drift contributions

$$N_x := \sum_{t \in H_x} Z_t = \sum_{t \in H_x} (x^t - y^t), \quad R_x := \sum_{t \in H_x} (p^t - x^t).$$

Note that $|p^t - x^t| < \eta$ on H_x , so $|R_x| \leq \eta n_x$.

For group g_3 , the contribution of context x to the bias is

$$S_x := \sum_{t \in H_x} (p^t - y^t) = N_x + R_x.$$

We next show that the calibration error over group g_3 is always at least the summed magnitude of the “honest” noise terms (summed over contexts x) minus the summed magnitude of the “honest” drift terms.

Lemma 4. *For any realization, if $\eta \leq 1/(2m)$:*

$$\sum_{v \in V_T} |B_T(v, g_3)| \geq \sum_{x \in X_0} |N_x| - \sum_{x \in X_0} |R_x|.$$

The condition $\eta \leq 1/(2m)$ ensures that the η -neighborhoods around different grid points $x \in X_0$ are disjoint, so each prediction value v can be η -close to at most one context. This allows us to decompose the g_3 error cleanly by context.

Proof. For each context $x \in X_0$ and prediction value v , let

$$H_{x,v} := \{t \in H_x : p^t = v\}, \quad S_{x,v} := \sum_{t \in H_{x,v}} (p^t - y^t).$$

Then $H_x = \bigsqcup_v H_{x,v}$ and

$$S_x = \sum_{t \in H_x} (p^t - y^t) = \sum_v S_{x,v} = N_x + R_x.$$

By the definition of g_3 ,

$$B_T(v, g_3) = \sum_{t=1}^T \mathbf{1}[p^t = v] g_3(x^t, p^t) (p^t - y^t) = \sum_{x \in X_0} S_{x,v}.$$

If $\eta \leq 1/(2m)$ then for each prediction value v there is at most one context $x \in X_0$ such that $|v - x| < \eta$. Hence, for each v , at most one of the sets $H_{x,v}$ is nonempty, and therefore at most one $S_{x,v}$ is nonzero. Thus $B_T(v, g_3) = S_{x,v}$ for that x , and we obtain

$$\sum_{v \in V_T} |B_T(v, g_3)| = \sum_{v \in V_T} \sum_{x \in X_0} |S_{x,v}| = \sum_{x \in X_0} \sum_{v \in V_T} |S_{x,v}|.$$

For each fixed x , by the triangle inequality,

$$\sum_{v \in V_T} |S_{x,v}| \geq \left| \sum_{v \in V_T} S_{x,v} \right| = |S_x| = |N_x + R_x| \geq |N_x| - |R_x|.$$

Summing over $x \in X_0$ yields the claimed bound. \square

Since $|R_x| \leq \eta n_x$ and $\sum_x n_x = H \leq T$, we have

$$\sum_{x \in X_0} |R_x| \leq \eta \sum_{x \in X_0} n_x = \eta H \leq \eta T. \tag{5}$$

Taking expectations in Lemma 4 and using this inequality gives

$$\mathbb{E} \left[\sum_{v \in V_T} |B_T(v, g_3)| \right] \geq \mathbb{E} \left[\sum_{x \in X_0} |N_x| \right] - \eta T. \tag{6}$$

As such, it remains to obtain a lower bound on $\mathbb{E} \sum_x |N_x|$.

3.4 Context-wise tradeoff between big deviations and noise

We now establish the key tradeoff: for each context x , either the algorithm makes many big deviations (contributing to g_1/g_2 error) or it makes mostly η -honest predictions (contributing noise to g_3 error). By aggregating over contexts, we will show that at least one of these error sources must be large.

For each $x \in X_0$, the sequence of labels $\{y^t : x^t = x\}$ consists of T_x independent draws from $\text{Bernoulli}(x)$, and therefore $Z_t = x^t - y^t$ takes values in $[-3/4, -1/4] \cup [1/4, 3/4]$ with

$$\mathbb{E}[Z_t | x^t = x] = 0, \quad \mathbb{E}[Z_t^2 | x^t = x] = x(1-x) \in [\frac{3}{16}, \frac{1}{4}].$$

Fix a context $x \in X_0$ and focus only on the subsequence of rounds with $x^t = x$. We will reindex these rounds in their own “local time” and define filtrations that, at each such step, contain exactly the information revealed after the prediction on that round but before its label is drawn.

For each t , let

$$\mathcal{H}_t := \sigma(x^1, \dots, x^t, y^1, \dots, y^{t-1}, p^1, \dots, p^t)$$

be the σ -field generated by the history just after the prediction p^t is chosen and just before y^t is revealed. Under $\mathcal{D}_{T,m}$ the label y^t is independent of \mathcal{H}_t with mean x^t , so for $Z_t = x^t - y^t$ we have

$$\mathbb{E}[Z_t | \mathcal{H}_t] = 0, \quad \mathbb{E}[Z_t^2 | \mathcal{H}_t] = x^t(1 - x^t) \in [\frac{3}{16}, \frac{1}{4}]. \quad (7)$$

For each $x \in X_0$, let $t_1 < \dots < t_{T_x}$ be the times with $x^{t_i} = x$. We define a local filtration

$$\mathcal{F}_0^{(x)} := \mathcal{H}_{t_1}, \quad \mathcal{F}_i^{(x)} := \mathcal{H}_{t_{i+1}} \text{ for } i = 1, \dots, T_x - 1,$$

and set $\mathcal{F}_{T_x}^{(x)} := \sigma(\mathcal{H}_{t_{T_x}}, y^{t_{T_x}})$. Clearly $(\mathcal{F}_i^{(x)})_{i=0}^{T_x}$ is a filtration and, for each $i = 1, \dots, T_x$, the random variable

$$Z_i^{(x)} := x - y^{t_i}$$

is $\mathcal{F}_i^{(x)}$ -measurable. Moreover, by (7) and the fact that $\mathcal{F}_{i-1}^{(x)}$ contains \mathcal{H}_{t_i} , we have

$$\mathbb{E}[Z_i^{(x)} | \mathcal{F}_{i-1}^{(x)}] = 0, \quad \mathbb{E}[(Z_i^{(x)})^2 | \mathcal{F}_{i-1}^{(x)}] \in [\frac{3}{16}, \frac{1}{4}], \quad |Z_i^{(x)}| \leq 1.$$

Finally, define

$$I_i^{(x)} := \mathbf{1}[t_i \in H_x], \quad i = 1, \dots, T_x.$$

Membership $t_i \in H_x$ depends only on the context x^{t_i} and the prediction p^{t_i} , both of which are \mathcal{H}_{t_i} -measurable; hence $I_i^{(x)}$ is $\mathcal{F}_{i-1}^{(x)}$ -measurable. Thus the sequence $(Z_i^{(x)}, \mathcal{F}_i^{(x)})$ together with the predictable indicators $I_i^{(x)}$ satisfies the assumptions (1) of Proposition 1 with $\sigma^2 = 3/16$ and $L = T_x$. Moreover,

$$N_x = \sum_{t \in H_x} Z_t = \sum_{i=1}^{T_x} I_i^{(x)} Z_i^{(x)}, \quad n_x = \sum_{i=1}^{T_x} I_i^{(x)}.$$

We define

$$\bar{B}_x := \mathbb{E}[b_x].$$

We know that $\sum_x b_x = B_T$ pathwise, hence $\sum_x \bar{B}_x = \mathbb{E}[B_T]$.

Intuitively, if the total expected number of big-deviation rounds $\mathbb{E}[B_T]$ is small, then big deviations cannot be spread across too many contexts—they must concentrate on a small subset. The remaining “dense” contexts have mostly η -honest predictions, which means the noise terms N_x on these contexts are sums over a dense (constant-fraction) subset of the T_x rounds. This is precisely the setting where Proposition 1 applies, giving $\mathbb{E}|N_x| = \Omega(\sqrt{T_x})$ for each dense context.

Lemma 5. *Suppose $\mathbb{E}[B_T] \leq \frac{T}{4}$. Then there exists a subset $D \subseteq X_0$ of contexts with*

$$\sum_{x \in D} T_x \geq \frac{T}{2}, \quad (8)$$

such that for every $x \in D$ we have

$$\mathbb{E}[n_x] \geq \frac{T_x}{2}. \quad (9)$$

Proof. Define the set of “sparse” contexts

$$S := \left\{ x \in X_0 : \bar{B}_x > \frac{1}{2} T_x \right\}, \quad D := X_0 \setminus S.$$

Then

$$\sum_{x \in S} \bar{B}_x > \frac{1}{2} \sum_{x \in S} T_x.$$

On the other hand, $\sum_x \bar{B}_x = \mathbb{E}[B_T] \leq \frac{T}{4} = \frac{1}{4} \sum_x T_x$, so

$$\frac{1}{2} \sum_{x \in S} T_x < \sum_{x \in S} \bar{B}_x \leq \sum_x \bar{B}_x \leq \frac{1}{4} \sum_x T_x.$$

Multiplying by 2 yields

$$\sum_{x \in S} T_x < \frac{1}{2} \sum_x T_x = \frac{T}{2},$$

and therefore (8) holds for $D = X_0 \setminus S$.

For any $x \in D$ we have $\bar{B}_x \leq \frac{T_x}{2}$, and since $n_x = T_x - b_x$ pathwise,

$$\mathbb{E}[n_x] = T_x - \bar{B}_x \geq \frac{T_x}{2},$$

establishing (9). \square

We can now apply Proposition 1 to each dense context.

Lemma 6. *There exists a constant $c > 0$ (independent of T and m) such that the following holds.*

Assume $\mathbb{E}[B_T] \leq \frac{T}{4}$. Then

$$\mathbb{E} \left[\sum_{x \in X_0} |N_x| \right] \geq c \sqrt{m_0 T}.$$

Proof. Let $D \subseteq X_0$ be the set of dense contexts from Lemma 5. For each $x \in D$, the process $\{Z_i^{(x)}\}$, the indicators $\{I_i^{(x)}\}$, and the horizon $L = T_x$ satisfy the assumptions of Proposition 1 with $\sigma^2 = 3/16$ and $\alpha = \frac{1}{2}$ (so that $\mathbb{E}[n_x] \geq \frac{T_x}{2}$). Applying the proposition with $L = T_x$ and this value of α yields $\mathbb{E}|N_x| \geq c_{\sigma,\alpha} \sqrt{T_x}$, for a constant $c_{\sigma,\alpha} > 0$ depending only on the variance lower bound and α . Hence there exists a constant $c' > 0$ such that

$$\mathbb{E}|N_x| = \mathbb{E} \left| \sum_{i=1}^{T_x} I_i^{(x)} Z_i^{(x)} \right| \geq c' \sqrt{T_x} \quad \text{for all } x \in D.$$

We therefore have

$$\mathbb{E} \left[\sum_{x \in X_0} |N_x| \right] \geq \sum_{x \in D} \mathbb{E}|N_x| \geq c \sum_{x \in D} \sqrt{T_x}.$$

Using (4) and (8), we obtain

$$\sum_{x \in D} \sqrt{T_x} \geq \sqrt{T_x^{\min}} \sum_{x \in D} 1 \geq \sqrt{T_x^{\min}} \frac{\sum_{x \in D} T_x}{T_x^{\max}} \geq \sqrt{c_{\text{occ}} \frac{T}{m_0}} \frac{T/2}{C_{\text{occ}} T/m_0} = \frac{\sqrt{c_{\text{occ}}}}{2C_{\text{occ}}} \sqrt{m_0 T}.$$

(Here T_x^{\min} and T_x^{\max} denote the minimum and maximum of the T_x , controlled by (4).) Thus

$$\mathbb{E} \left[\sum_{x \in X_0} |N_x| \right] \geq c' \frac{\sqrt{c_{\text{occ}}}}{2C_{\text{occ}}} \sqrt{m_0 T} =: c \sqrt{m_0 T},$$

as claimed. \square

3.4.1 Lower Bound for the “Honest” Group

We can now combine the previous bounds to obtain a lower bound on the g_3 contribution.

Lemma 7. *Assume $T \geq m_0$ and $\eta \leq 1/(2m)$. Then there exists a constant $c > 0$ (independent of T and m) such that, for any deterministic prediction algorithm under $\mathcal{D}_{T,m}$, either*

$$\mathbb{E}[B_T] \geq \frac{T}{4} \quad \text{and hence} \quad \mathbb{E}[\text{MCerr}_T] \geq \frac{\eta}{8} T, \quad (10)$$

or else $\mathbb{E}[B_T] < \frac{T}{4}$ and

$$\mathbb{E}\left[\sum_{v \in V_T} |B_T(v, g_3)|\right] \geq c\sqrt{m_0 T} - \eta T. \quad (11)$$

Proof. If $\mathbb{E}[B_T] \geq \frac{T}{4}$, then Lemma 3 yields

$$\mathbb{E}[\text{MCerr}_T] \geq \frac{\eta}{2} \mathbb{E}[B_T] \geq \frac{\eta}{8} T,$$

which is (10).

Otherwise, if $\mathbb{E}[B_T] < \frac{T}{4}$, then Lemma 6 gives

$$\mathbb{E}\left[\sum_{x \in X_0} |N_x|\right] \geq c\sqrt{m_0 T},$$

and plugging this into (6) yields

$$\mathbb{E}\left[\sum_{v \in V_T} |B_T(v, g_3)|\right] \geq c\sqrt{m_0 T} - \eta T.$$

This gives us (11). \square

3.4.2 Putting it All Together

We now combine all the pieces to prove Theorem 1. The argument proceeds by case analysis:

- If the algorithm makes many big deviations ($\mathbb{E}[B_T] \geq T/4$), then Lemma 3 gives $\mathbb{E}[\text{MCerr}_T] \geq \frac{\eta}{8} T = \Omega(T^{2/3})$.
- If the algorithm makes few big deviations ($\mathbb{E}[B_T] < T/4$), then most rounds are η -honest. Lemma 7 shows the noise contribution on group g_3 is $\Omega(\sqrt{mT}) = \Omega(T^{2/3})$.

The threshold $\eta = \Theta(\sqrt{m/T}) = \Theta(T^{-1/3})$ is chosen to balance these two cases.

Fix $T \geq 1$ and let $m := \lfloor T^{1/3} \rfloor$. For T sufficiently large, we have $m \geq 8$ and $m \leq T^{1/3}$.

Let $c_0 > 0$ be the density constant from the hard distribution construction (so $m_0 \geq c_0 m$), and let $c > 0$ be the constant from Lemma 7.

Set

$$\eta := \delta \sqrt{\frac{m}{T}}.$$

Where δ is a constant > 0 such that:

$$\delta \leq \frac{c\sqrt{c_0}}{4} \quad \text{and} \quad \delta \leq \frac{1}{2}. \quad (12)$$

Using $m_0 \geq c_0 m$, we have for all sufficiently large T ,

$$\eta T = \delta \sqrt{mT} \leq \delta \frac{1}{\sqrt{c_0}} \sqrt{m_0 T} \leq \frac{c}{4} \sqrt{m_0 T} \leq \frac{c}{2} \sqrt{m_0 T},$$

so the drift term ηT is always at most $\frac{c}{2} \sqrt{m_0 T}$. Moreover, since $m \leq T^{1/3}$ we have $\eta = \delta \sqrt{m/T} \leq \delta T^{-1/3}$, and the bound $\delta \leq 1/2$ in (12) yields

$$\eta \leq \frac{1}{2} T^{-1/3} \leq \frac{1}{2m}$$

for all sufficiently large T , so the condition $\eta \leq 1/(2m)$ needed in Lemmas 4 and 7 also holds.

Let A be any prediction algorithm. By Lemma 2, we may assume without loss of generality that A is deterministic. We distinguish two cases according to the size of $\mathbb{E}[B_T]$.

Case 1: $\mathbb{E}[B_T] \geq \frac{T}{4}$. In this case, Lemma 3 yields

$$\mathbb{E}[\text{MCerr}_T] \geq \frac{\eta}{2} \mathbb{E}[B_T] \geq \frac{\eta}{8} T = \frac{\delta}{8} \sqrt{mT}.$$

Since $m = \Theta(T^{1/3})$, this gives

$$\mathbb{E}[\text{MCerr}_T] = \Omega(T^{2/3}).$$

Case 2: $\mathbb{E}[B_T] < \frac{T}{4}$. In this case, Lemma 7 (specifically (11)) gives

$$\mathbb{E}\left[\sum_{v \in V_T} |B_T(v, g_3)|\right] \geq c\sqrt{m_0 T} - \eta T \geq \frac{c}{2} \sqrt{m_0 T},$$

where the last inequality uses $\eta T \leq (c/2)\sqrt{m_0 T}$ from our choice of δ in (12). Thus the drift term subtracts at most half of the noise lower bound, and the right-hand side remains of order $\sqrt{m_0 T}$. Since $m_0 \geq c_0 m$ for a constant $c_0 > 0$, we have

$$\sqrt{m_0 T} \geq \sqrt{c_0} \sqrt{mT} = \Theta(\sqrt{mT}),$$

and therefore

$$\mathbb{E}\left[\sum_{v \in V_T} |B_T(v, g_3)|\right] = \Theta(\sqrt{mT}) = \Theta(T^{2/3}).$$

Since MCerr_T is at least the g_3 contribution, this again implies

$$\mathbb{E}[\text{MCerr}_T] = \Theta(T^{2/3}).$$

Combining the two cases, we have established

$$\mathbb{E}[\text{MCerr}_T] \geq c' T^{2/3}$$

for some constant $c' > 0$ and all sufficiently large T (depending only on the fixed constants c_0, c_1, δ).

Thus, even for a family of only three simple groups, no online prediction algorithm can guarantee expected multicalibration error $o(T^{2/3})$ against adversarially chosen contexts and outcomes, matching (up to logarithmic factors) known online multicalibration upper bounds [Noarov et al., 2025], and separating the statistical complexity of multicalibration from marginal calibration for which $O(T^{2/3-\varepsilon})$ upper bounds are known [Dagan et al., 2025].

4 Optimal Lower Bound for Prediction-Independent Groups

This section gives a lower bound for online multicalibration using only *prediction-independent* group functions $g : \mathcal{C} \rightarrow [0, 1]$, where \mathcal{C} is the context space.

Specifically, this section is devoted to proving the following theorem:

Theorem 2 (Prediction-independent lower bound). *There exist universal constants $c, C > 0$ and $T_0 \in \mathbb{N}$ such that for all $T \geq T_0$, under the distribution and prediction-independent group family G defined in this section, every (possibly randomized) online forecaster satisfies*

$$\mathbb{E}[\text{MCerr}_{T'}(G)] \geq c \cdot \frac{T^{2/3}}{\log^C(T+1)}.$$

Proof roadmap. The key challenge in proving lower bounds for prediction-independent groups is that we cannot directly detect when the learner deviates from “honest” predictions $p^t = \mathbb{E}[y^t] = x^t$, since our groups cannot depend on the prediction value. Instead, we proceed in three steps:

1. **Global orthogonal groups enforce approximate honesty** (Section 4.3): We define a family of Walsh-based groups on the mean grid. Using the orthogonality and “prefix-sum” properties of the Walsh basis, we show that any forecaster with small multicalibration error on these groups must have small total ℓ_1 deviation from honest predictions: $\sum_t |p^t - x^t| = \tilde{O}(\text{MCerr})$.
2. **Approximate honesty forces diverse predictions** (Section 4.3): Small ℓ_1 deviation from honesty prevents the forecaster from concentrating predictions on a few values, forcing $N := \sum_v \sqrt{n_v}$ to be large, where n_v counts predictions equal to v .
3. **Blockwise Hadamard groups extract unavoidable noise** (Sections 4.5–4.7): We partition time into blocks and define Hadamard-based groups on each block. The calibration error decomposes into *bias* (controlled by approximate honesty) and *noise* (inherently random). Using a martingale argument, we show the noise contribution scales as $\tilde{\Omega}(N)$, which combined with step 2 yields the $\tilde{\Omega}(T^{2/3})$ lower bound.

Recall that for a group function $g : \mathcal{C} \rightarrow [0, 1]$, its calibration error is:

$$\text{Err}_T(g) := \sum_{v \in \mathcal{V}_T} \left| \sum_{t=1}^T \mathbf{1}[p^t = v] g(c^t) (p^t - y^t) \right|,$$

where $\mathcal{V}_T := \{p^1, \dots, p^T\}$. In the analysis we will also apply this definition to signed weight functions $w : \mathcal{C} \rightarrow [-1, 1]$ using the same formula.

4.1 The Hard Distribution: Time-Augmented Contexts.

Our hard distribution is similar to the one we used in Section 3: the contexts are fixed deterministically, and encode label means uniformly spread throughout $[1/4, 3/4]$. We modify the instance in two ways: the labels are no longer Bernoulli, but result from adding Rademacher noise to the label mean (which simplifies some of our arguments), and the contexts are augmented with a time index, which we will use when defining our hard group family.

Fix a horizon $T \geq 2$. Let $m := \max\{2, 2^{\lfloor \log_2(T^{1/3}) \rfloor}\}$ and define grid means

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

so $x_i \in [1/4, 3/4]$ and $|x_{i+1} - x_i| = \Theta(1/m)$. Define the deterministic round-robin mean sequence

$$x^t := x_{1+((t-1) \bmod m)} \quad (t = 1, \dots, T).$$

Augment the context with time:

$$c^t := (x^t, t) \in [0, 1] \times \{1, \dots, T\}.$$

Outcomes are generated by adding independent Rademacher noise to the means. Let $(\xi^t)_{t=1}^T$ be independent signs with $\mathbb{P}(\xi^t = 1) = \mathbb{P}(\xi^t = -1) = 1/2$, and set

$$y^t := x^t + \frac{\xi^t}{4}, \quad t = 1, \dots, T. \tag{13}$$

Then $y^t \in [0, 1]$ for all t , $\mathbb{E}[y^t | c^t] = x^t$, and $\text{Var}(y^t | c^t) = 1/16$.

4.2 The Hard Group Family.

4.2.1 Hadamard and Walsh systems.

We will construct our prediction-independent groups from $\{\pm 1\}$ -valued orthogonal systems. The key idea is that orthogonal systems let us “test” the forecaster’s behavior in many independent directions simultaneously. If the forecaster has small calibration error on all groups in an orthogonal family, we can use Parseval’s identity to conclude that the forecaster’s aggregate deviation from honest predictions must be small.

These orthogonal systems will be used in two different roles: (i) *globally* on the mean grid (where we will instantiate a specific system, the Walsh system), and (ii) *locally* within time blocks that we will specify later (where any Hadamard system suffices). We now introduce these systems.

Definition 6 (Hadamard system). Let n be a power of two. A collection of functions

$$\{\psi_j\}_{j=0}^{n-1}, \quad \psi_j : \{0, \dots, n-1\} \rightarrow \{\pm 1\},$$

is called a *Hadamard system of length n* if for all $j, j' \in \{0, \dots, n-1\}$ we have

$$\sum_{s=0}^{n-1} \psi_j(s) \psi_{j'}(s) = \begin{cases} n, & j = j', \\ 0, & j \neq j'. \end{cases} \tag{14}$$

We will refer to (14) as property (H1), and to the constraint $\psi_j(s) \in \{\pm 1\}$ for all j, s as property (H2).

We will use the following consequence of (H1).

Proposition 2 (Parseval's identity for Hadamard systems). *For any vector $A \in \mathbb{R}^n$ with entries $A(s)$ for $s = 0, \dots, n - 1$, we have, for any Hadamard system of length n ,*

$$\sum_{j=0}^{n-1} \left(\sum_{s=0}^{n-1} A(s) \psi_j(s) \right)^2 = n \sum_{s=0}^{n-1} A(s)^2.$$

Equivalently, if we write $\langle f, g \rangle := \sum_{s=0}^{n-1} f(s)g(s)$ and $\|A\|_2^2 := \sum_{s=0}^{n-1} A(s)^2$, then

$$\sum_{j=0}^{n-1} \langle A, \psi_j \rangle^2 = n \|A\|_2^2.$$

A canonical example of a Hadamard system is the *Walsh system*.

Definition 7 (Walsh system). Let n be a power of two. For any $j \in \{0, \dots, n-1\}$, $s \in \{0, \dots, n-1\}$, write j and s in binary and let $\langle j, s \rangle_2$ denote their mod-2 inner product. Define

$$\psi_j^{\text{Wal}} : \{0, \dots, n-1\} \rightarrow \{\pm 1\}, \quad \psi_j^{\text{Wal}}(s) := (-1)^{\langle j, s \rangle_2}.$$

The resulting collection of functions $\{\psi_j^{\text{Wal}}\}_{j=0}^{n-1}$ is called a *Walsh system of length n* .

The Walsh system is a Hadamard system as it satisfies properties (H1) and (H2). In addition, it satisfies a prefix-sum bound as stated in Lemma 8, which we will refer to as property (H3). This additional property will be crucial for showing that calibration error on the Walsh groups controls the ℓ_1 distance to honest predictions.

Lemma 8 (Walsh prefix-sum bound). *For every $j \in \{1, \dots, n-1\}$, denote the number of trailing zeroes in its binary expansion as:*

$$\text{tz}(j) := \max\{d \geq 0 : 2^d \text{ divides } j\}.$$

Then we have:

$$\max_{r \in \{0, \dots, n\}} \left| \sum_{s < r} \psi_j^{\text{Wal}}(s) \right| \leq 2^{\text{tz}(j)}.$$

Proof. Fix $j \in \{1, \dots, n-1\}$. Write the binary expansions

$$j = \sum_{b \geq 0} j_b 2^b, \quad s = \sum_{b \geq 0} s_b 2^b,$$

where $j_b, s_b \in \{0, 1\}$ denote the b -th (least-significant) bits.

By definition of tz , we have:

$$j_0 = j_1 = \dots = j_{\text{tz}(j)-1} = 0 \quad \text{and} \quad j_{\text{tz}(j)} = 1.$$

As a result,

$$\langle j, s \rangle_2 = \sum_{b \geq 0} j_b s_b = \sum_{b \geq \text{tz}(j)} j_b s_b \pmod{2}.$$

Therefore, $\psi_j^{\text{Wal}}(s)$ doesn't depend on the first $\text{tz}(j)$ bits of s . That is to say, ψ_j^{Wal} is constant on each dyadic block of length $2^{\text{tz}(j)}$:

$$\{k \cdot 2^{\text{tz}(j)}, k \cdot 2^{\text{tz}(j)} + 1, \dots, (k+1) \cdot 2^{\text{tz}(j)} - 1\}.$$

Next we group two consecutive blocks into a superblock of length $2^{\text{tz}(j)+1}$. Consider the two blocks:

$$I_{k,0} := \{2k \cdot 2^{\text{tz}(j)}, \dots, (2k+1) \cdot 2^{\text{tz}(j)} - 1\}, \quad I_{k,1} := \{(2k+1) \cdot 2^{\text{tz}(j)}, \dots, (2k+2) \cdot 2^{\text{tz}(j)} - 1\}.$$

For any $s \in I_{k,0}$, the $\text{tz}(j)$ -th bit satisfies $s_{\text{tz}(j)} = 0$, while for any $s' \in I_{k,1}$ we have $s'_{\text{tz}(j)} = 1$. In addition, $s_b = s'_b$ for all $b > \text{tz}(j)$, i.e., the higher bits are the same across the two halves of a superblock.

Since $j_{\text{tz}(j)} = 1$, this implies

$$\langle j, s' \rangle_2 - \langle j, s \rangle_2 = \sum_{b \geq \text{tz}(j)} j_b s'_b - \sum_{b \geq \text{tz}(j)} j_b s_b = j_{\text{tz}(j)} (s'_{\text{tz}(j)} - s_{\text{tz}(j)}) = 1 \pmod{2},$$

and hence

$$\psi_j^{\text{Wal}}(s') = -\psi_j^{\text{Wal}}(s).$$

Therefore the sum of $\psi_j^{\text{Wal}}(s)$ over a full superblock cancels:

$$\sum_{s \in I_{k,0} \cup I_{k,1}} \psi_j^{\text{Wal}}(s) = 0.$$

Every prefix sum can be decomposed into a disjoint union of complete superblocks (each contributing 0) plus a remainder that contributes at most $2^{\text{tz}(j)}$. Hence

$$\left| \sum_{s < r} \psi_j^{\text{Wal}}(s) \right| \leq 2^{\text{tz}(j)} \quad \text{for all } r \in \{0, \dots, n\}.$$

□

4.2.2 The Complete Group Family

We now define the prediction-independent group family used in our lower bound. The family consists of three types of groups:

- A **constant group** g_{all} that enforces marginal calibration.
- **Global Walsh groups** defined on the mean grid, which together enforce that the forecaster's predictions stay close to the honest predictions in an ℓ_1 sense.
- **Block Hadamard groups** that partition time into blocks and test the forecaster's behavior within each block. These groups will extract the unavoidable noise contribution to calibration error.

Let

$$K := \lceil \log^{10}(T + 1) \rceil,$$

and let L be the largest power of two such that $KL \leq T$ (so $L \geq T/(2K)$). Write $T' := KL$ and partition the first T' times into K contiguous blocks

$$J_a := \{(a - 1)L + 1, \dots, aL\}, \quad a = 1, \dots, K.$$

From here on out we will prove lower bounds as a function of T' for convenience. Note that $T' \in [T/2, T]$, so this will not affect the asymptotic rate.

Definition 8 (Grid index map). Define $\text{idx} : [0, 1] \rightarrow \{1, \dots, m\}$ by $\text{idx}(x_i) = i$ for each $i = 1, \dots, m$, and define $\text{idx}(x) = 1$ for all other $x \in [0, 1]$.

Constant groups We include the constant group

$$g_{\text{all}}(x, t) := 1.$$

Global Walsh groups. For each $\ell \in \{1, \dots, m - 1\}$ we define the signed Walsh feature

$$w_\ell(x, t) := \psi_\ell^{\text{Wal}}(\text{idx}(x) - 1) = (-1)^{\langle \ell, \text{idx}(x) - 1 \rangle_2} \in \{\pm 1\},$$

and convert it to two binary Walsh half-groups by

$$g_\ell^{\text{Wal},+}(x, t) := \frac{1 + w_\ell(x, t)}{2} \in \{0, 1\}, \quad (15)$$

$$g_\ell^{\text{Wal},-}(x, t) := \frac{1 - w_\ell(x, t)}{2} \in \{0, 1\}. \quad (16)$$

(We omit $\ell = 0$ since $w_0 \equiv 1$ is already covered by g_{all} ; we will sometimes identify g_{all} with w_0 for convenience.)

Block Hadamard groups. For each block J_a , index local times by $s \in \{0, \dots, L - 1\}$ via $t = (a - 1)L + 1 + s$. Fix, for each $a \in [K]$, an arbitrary Hadamard system $\{\psi_{a,j}\}_{j=0}^{L-1}$ of length L . For concreteness one may take the length- L Walsh system, but all later blockwise arguments will only use properties (H1) and (H2).

For each $a \in [K]$ and $j \in \{0, \dots, L - 1\}$, we include two *binary* Hadamard half-groups

$$g_{a,j}^+(x, t) := \mathbf{1}[t \in J_a] \cdot \frac{1 + \psi_{a,j}(t - (a - 1)L - 1)}{2} \in \{0, 1\}, \quad (17)$$

$$g_{a,j}^-(x, t) := \mathbf{1}[t \in J_a] \cdot \frac{1 - \psi_{a,j}(t - (a - 1)L - 1)}{2} \in \{0, 1\}. \quad (18)$$

Note that $g_{a,j}^+ + g_{a,j}^- = \mathbf{1}[t \in J_a]$, hence

$$T_{g_{a,j}^+} + T_{g_{a,j}^-} = \sum_{t \leq T'} \mathbf{1}[t \in J_a] = L, \quad \text{so} \quad \max\{T_{g_{a,j}^+}, T_{g_{a,j}^-}\} \geq L/2.$$

The full group family. Let

$$G := \{g_{\text{all}}\} \cup \{g_\ell^{\text{Wal},+}, g_\ell^{\text{Wal},-} : \ell = 1, \dots, m-1\} \cup \{g_{a,j}^+, g_{a,j}^- : a \in [K], j \in \{0, \dots, L-1\}\}.$$

The family consists of $1 + 2(m-1) + 2KL = O(m+T') = O(T)$ binary prediction-independent groups.

Simulating signed weights by differences. Our groups take values in $\{0, 1\}$, but it will be useful to consider calibration errors with respect to weighting functions that can take negative values. Towards this end, define the auxiliary signed function which takes the difference between two groups in G :

$$h_{a,j} := g_{a,j}^+ - g_{a,j}^- = \mathbf{1}[t \in J_a] \cdot \psi_{a,j}(t - (a-1)L - 1) \in \{0, \pm 1\}. \quad (19)$$

Similarly, by definition of the signed Walsh feature and Walsh half-groups, we have

$$w_\ell = g_\ell^{\text{Wal},+} - g_\ell^{\text{Wal},-} \in \{\pm 1\}. \quad (20)$$

Lemma 9 (Difference-of-two reduction). *Fix any two groups $g^+, g^- : \mathcal{C} \rightarrow [0, 1]$. For every realization,*

$$\text{Err}_{T'}(g^+ - g^-) \leq \text{Err}_{T'}(g^+) + \text{Err}_{T'}(g^-),$$

and consequently

$$\max\{\text{Err}_{T'}(g^+), \text{Err}_{T'}(g^-)\} \geq \frac{1}{2} \text{Err}_{T'}(g^+ - g^-).$$

In particular, with $h_{a,j}$ as in (19) and w_ℓ as in (20),

$$\text{MCerr}_{T'}(G) \geq \frac{1}{2} \max_{a,j} \text{Err}_{T'}(h_{a,j}), \quad \text{MCerr}_{T'}(G) \geq \frac{1}{2} \max_\ell \text{Err}_{T'}(w_\ell),$$

Proof. Fix a prediction value v . Let

$$B(v, g) := \sum_{t \leq T'} \mathbf{1}[p^t = v] g(c^t) (p^t - y^t).$$

Then $B(v, g^+ - g^-) = B(v, g^+) - B(v, g^-)$, so $|B(v, g^+ - g^-)| \leq |B(v, g^+)| + |B(v, g^-)|$. Summing over v gives $\text{Err}_{T'}(g^+ - g^-) \leq \text{Err}_{T'}(g^+) + \text{Err}_{T'}(g^-)$, and hence the max is at least half. For the last inequality, apply this to each pair $(g_{a,j}^+, g_{a,j}^-)$ and take a maximum over (a, j) , and apply this to each pair $(g_\ell^{\text{Wal},+}, g_\ell^{\text{Wal},-})$ and take a maximum over ℓ . \square

4.3 Global Walsh Groups Enforce ℓ_1 -Truthfulness

We now show that multicalibration with respect to the *global Walsh groups* forces the forecaster to be close to the honest predictor in total ℓ_1 loss.

The high-level idea is as follows. For any prediction value v , the sign pattern $\text{sign}(v - x_i)$ across grid points i can be expanded in the Walsh basis. The Walsh prefix-sum bound (Lemma 8) ensures that this expansion has small ℓ_1 coefficient mass—at most $O(\log m)$. Therefore, the total ℓ_1 deviation $\sum_t |p^t - x^t|$ can be written as a weighted combination of calibration biases on the Walsh groups, with total weight $O(\log m)$. If all these biases are small (i.e., if multicalibration error is small), then the ℓ_1 deviation must also be small.

4.3.1 Walsh Expansion of Discrete Threshold Signs on the Grid

Fix a prediction value $v \in [0, 1]$. Because the context means always lie on the grid $\{x_1, \dots, x_m\}$, the sign pattern $\text{sign}(v - x_i)$ is determined solely by the number of grid points $\leq v$. Define

$$r(v) := |\{i \in \{1, \dots, m\} : x_i \leq v\}| \in \{0, 1, \dots, m\}.$$

Define the discrete sign function on indices $u \in \{0, \dots, m-1\}$ by

$$f_r(u) := \begin{cases} +1 & \text{if } u \leq r-1, \\ -1 & \text{if } u \geq r, \end{cases} \quad r \in \{0, 1, \dots, m\}.$$

Then for every grid point x_i we have $f_{r(v)}(i-1) = +1$ if $x_i \leq v$ and $f_{r(v)}(i-1) = -1$ if $x_i > v$. In particular, for every time $t \leq T'$ (with $x^t \in \{x_1, \dots, x_m\}$),

$$|v - x^t| = f_{r(v)}(\text{idx}(x^t) - 1) \cdot (v - x^t).$$

We now expand the threshold sign pattern f_r in the Walsh basis on $\{0, \dots, m-1\}$ and bound the ℓ_1 mass of its coefficients. Recall that $m = \max\{2, 2^{\lfloor \log_2(T^{1/3}) \rfloor}\}$ is a power of two.

Lemma 10 (Walsh expansion of discrete threshold signs). *Fix m a power of two. Let $\{\psi_\ell^{\text{Wal}}\}_{\ell=0}^{m-1}$ be the length- m Walsh system. Then for every $r \in \{0, \dots, m\}$, there exist coefficients $\{\alpha_\ell(r)\}_{\ell=0}^{m-1}$ such that*

$$f_r(u) = \sum_{\ell=0}^{m-1} \alpha_\ell(r) \psi_\ell^{\text{Wal}}(u) \quad \text{for every } u \in \{0, \dots, m-1\}.$$

Moreover, the coefficients satisfy that

$$\sum_{\ell=0}^{m-1} \max_{r \in \{0, \dots, m\}} |\alpha_\ell(r)| \leq 1 + \log_2 m.$$

Proof. Define the Walsh coefficients by:

$$\alpha_\ell(r) := \frac{1}{m} \sum_{u=0}^{m-1} f_r(u) \psi_\ell^{\text{Wal}}(u), \quad \ell = 0, \dots, m-1.$$

Then the expansion $f_r(u) = \sum_\ell \alpha_\ell(r) \psi_\ell^{\text{Wal}}(u)$ follows from (H1).

It remains to bound the ℓ_1 mass. For $\ell = 0$ we have $\psi_0^{\text{Wal}} \equiv 1$, so

$$|\alpha_0(r)| = \left| \frac{1}{m} \sum_{u=0}^{m-1} f_r(u) \right| = \left| \frac{2r - m}{m} \right| \leq 1.$$

Now fix $\ell \in \{1, \dots, m-1\}$. Using the identity $f_r(u) = 2\mathbf{1}[u \leq r-1] - 1$ and $\sum_{u=0}^{m-1} \psi_\ell^{\text{Wal}}(u) = 0$ (which holds for $\ell \neq 0$), we obtain:

$$\alpha_\ell(r) = \frac{1}{m} \sum_{u=0}^{m-1} f_r(u) \psi_\ell^{\text{Wal}}(u) = \frac{2}{m} \sum_{u=0}^{r-1} \psi_\ell^{\text{Wal}}(u).$$

Therefore, by the Walsh prefix-sum bound, i.e., Lemma 8, we have:

$$\max_{r \in \{0, \dots, m\}} |\alpha_\ell(r)| \leq \frac{2}{m} \cdot 2^{\text{tz}(\ell)}.$$

Group indices by $d := \text{tz}(\ell) \in \{0, 1, \dots, \log_2 m - 1\}$. There are exactly $m/2^{d+1}$ values of $\ell \in \{1, \dots, m-1\}$ with $\text{tz}(\ell) = d$. Hence,

$$\sum_{\ell=1}^{m-1} \max_{r \in \{0, \dots, m\}} |\alpha_\ell(r)| \leq \sum_{\ell=1}^{m-1} \frac{2}{m} \cdot 2^{\text{tz}(\ell)} \leq \sum_{d=0}^{\log_2 m - 1} \frac{m}{2^{d+1}} \cdot \frac{2}{m} \cdot 2^d = \sum_{d=0}^{\log_2 m - 1} 1 = \log_2 m.$$

Combining with $|\alpha_0(r)| \leq 1$ yields

$$\sum_{\ell=0}^{m-1} \max_{r \in \{0, \dots, m\}} |\alpha_\ell(r)| \leq 1 + \log_2 m. \quad \square$$

4.3.2 Global Walsh Groups Enforce ℓ_1 -Truthfulness

Now we use the global Walsh groups to enforce a form of ℓ_1 truthfulness.

Lemma 11 (ℓ_1 -truthfulness from Walsh groups). *Define the total ℓ_1 deviation from honesty on the first T' rounds by:*

$$A := \sum_{t=1}^{T'} |p^t - x^t|.$$

Then there exists a universal constant $C_{\ell_1} > 0$ such that, under the environment (13), every forecaster satisfies:

$$\mathbb{E}[A] \leq C_{\ell_1} \log(m+1) \cdot \mathbb{E}[\text{MCerr}_{T'}(G)].$$

Proof. Fix a forecaster. For each realized prediction value $v \in \mathcal{V}_{T'}$, define the prediction-dependent sign weight on contexts

$$s_v(x, t) := f_{r(v)}(\text{idx}(x) - 1) \in \{\pm 1\}.$$

For grid points $x = x_i$, we have $s_v(x_i, t) = +1$ if $x_i \leq v$ and $s_v(x_i, t) = -1$ otherwise, and hence for all $t \leq T'$ (with $x^t \in \{x_1, \dots, x_m\}$),

$$|v - x^t| = s_v(x^t, t) (v - x^t).$$

Let $S_v := \{t \leq T' : p^t = v\}$. Then

$$A = \sum_{t \leq T'} |p^t - x^t| = \sum_{v \in \mathcal{V}_{T'}} \sum_{t \in S_v} s_v(x^t, t) (v - x^t).$$

Let \mathcal{F}_t be the sigma-field generated by the transcript up to and including the realized prediction p^t , but excluding y^t . Since the environment is oblivious, y^t is independent of \mathcal{F}_t and satisfies

$\mathbb{E}[y^t \mid \mathcal{F}_t] = x^t$. Thus $\mathbb{E}[p^t - y^t \mid \mathcal{F}_t] = p^t - x^t$. Moreover $s_{p^t}(x^t, t)$ is \mathcal{F}_t -measurable, so

$$\begin{aligned}\mathbb{E}[A] &= \mathbb{E} \left[\sum_{t \leq T'} s_{p^t}(x^t, t)(p^t - x^t) \right] \\ &= \mathbb{E} \left[\sum_{t \leq T'} s_{p^t}(x^t, t) \mathbb{E}[p^t - y^t \mid \mathcal{F}_t] \right] \\ &= \mathbb{E} \left[\sum_{t \leq T'} s_{p^t}(x^t, t)(p^t - y^t) \right] \\ &= \mathbb{E} \left[\sum_{v \in \mathcal{V}_{T'}} \sum_{t \in S_v} s_v(x^t, t)(v - y^t) \right].\end{aligned}$$

By Lemma 10, for each fixed v and each grid point x we can expand $s_v(x, t)$ as:

$$s_v(x, t) = f_{r(v)}(\text{idx}(x) - 1) = \sum_{\ell=0}^{m-1} \alpha_\ell(r(v)) \psi_\ell^{\text{Wal}}(\text{idx}(x) - 1).$$

Substituting into the sum gives:

$$\begin{aligned}\mathbb{E}[A] &= \mathbb{E} \left[\sum_{v \in \mathcal{V}_{T'}} \sum_{t \in S_v} \sum_{\ell=0}^{m-1} \alpha_\ell(r(v)) \psi_\ell^{\text{Wal}}(\text{idx}(x^t) - 1)(v - y^t) \right] \\ &= \mathbb{E} \left[\sum_{\ell=0}^{m-1} \sum_{v \in \mathcal{V}_{T'}} \alpha_\ell(r(v)) \sum_{t \in S_v} \psi_\ell^{\text{Wal}}(\text{idx}(x^t) - 1)(v - y^t) \right] \\ &= \mathbb{E} \left[\sum_{\ell=0}^{m-1} \sum_{v \in \mathcal{V}_{T'}} \alpha_\ell(r(v)) \sum_{t \in S_v} w_\ell(x^t, t)(v - y^t) \right] \\ &= \mathbb{E} \left[\sum_{\ell=0}^{m-1} \sum_{v \in \mathcal{V}_{T'}} \alpha_\ell(r(v)) B_{T'}(v, w_\ell) \right]\end{aligned}$$

By the triangle inequality, the above sum is bounded by:

$$\begin{aligned}\mathbb{E} \left[\sum_{\ell=0}^{m-1} \sum_{v \in \mathcal{V}_{T'}} |\alpha_\ell(r(v))| |B_{T'}(v, w_\ell)| \right] &\leq \mathbb{E} \left[\sum_{\ell=0}^{m-1} \max_{v \in \mathcal{V}_{T'}} |\alpha_\ell(r(v))| \sum_{v \in \mathcal{V}_{T'}} |B_{T'}(v, w_\ell)| \right] \\ &= \mathbb{E} \left[\sum_{\ell=0}^{m-1} \max_{r \in \{0, \dots, m\}} |\alpha_\ell(r)| \text{Err}_{T'}(w_\ell) \right]\end{aligned}$$

where $\text{Err}_{T'}(w_\ell)$ denotes the calibration-error functional extended to signed weights $w_\ell \in \{\pm 1\}$.

Finally, since $w_0 \equiv 1$ is covered by g_{all} and for each $\ell \geq 1$ we have $w_\ell = g_\ell^{\text{Wal},+} - g_\ell^{\text{Wal},-}$, the difference-of-two reduction (Lemma 9) implies that

$$\begin{aligned}\text{Err}_{T'}(w_0) &= \text{Err}_{T'}(g_{\text{all}}) \leq \text{MCerr}_{T'}(G), \\ \text{Err}_{T'}(w_\ell) &\leq \text{Err}_{T'}(g_\ell^{\text{Wal},+}) + \text{Err}_{T'}(g_\ell^{\text{Wal},-}) \leq 2 \text{MCerr}_{T'}(G) \quad \text{for each } \ell \geq 1.\end{aligned}$$

Moreover, Lemma 10 gives the uniform bound

$$\sum_{\ell=0}^{m-1} \max_{r \in \{0, \dots, m\}} |\alpha_\ell(r)| \leq 1 + \log_2 m.$$

Combining these bounds yields

$$\mathbb{E}[A] \leq C_{\ell_1} \log(m+1) \cdot \mathbb{E}[\text{MCerr}_{T'}(G)]$$

for a universal constant C_{ℓ_1} . □

4.4 Multicalibration Requires Diverse Predictions

The “honest” prediction strategy of predicting $\mathbb{E}[y^t] = x^t$ at every round obtains high marginal calibration error because our instance has x^t take on many different values, and so the honest predictor accumulates noise-driven empirical bias in each of these many “prediction bins.”

A natural question is whether a forecaster can do better by “consolidating” predictions—using fewer distinct prediction values to reduce the number of bins. The key insight of this section is that such consolidation is incompatible with the ℓ_1 -truthfulness constraint established above. A forecaster that stays close to honest predictions (in ℓ_1) cannot concentrate its predictions on a small number of values; it must spread them out roughly as the honest forecaster would.

Formally, we show that an algorithm with small ℓ_1 deviation from honesty must have $N := \sum_v \sqrt{n_v}$ large, where n_v counts predictions equal to v . This quantity N will later serve as a lower bound on the noise contribution to calibration error.

Let $n_v := |\{t \leq T' : p^t = v\}|$ denote the number of rounds on which an algorithm makes a prediction with value v , and let

$$N := \sum_{v \in \mathcal{V}_{T'}} \sqrt{n_v}.$$

4.4.1 ℓ_1 Time-Quantization and Diverse Predictions

We now prove an ℓ_1 time-quantization inequality relating A to the bucket counts $(n_v)_v$.

Lemma 12 (ℓ_1 time-quantization). *Let $(p^t)_{t \leq T'}$ be any prediction sequence and recall that $(x^t)_{t \leq T'}$ cycles through the m grid means. Let $n_v := |\{t \leq T' : p^t = v\}|$, $N := \sum_{v \in \mathcal{V}_{T'}} \sqrt{n_v}$, and $A := \sum_{t \leq T'} |p^t - x^t|$. Then for all $T' \geq 2$ we have*

$$A \geq \frac{1}{16T'} \sum_{v \in \mathcal{V}_{T'}} n_v^2 - \frac{T'}{m} - 1.$$

Proof. We use a sorting/proxy argument for ℓ_1 loss. Let $x(1) \leq \dots \leq x(T')$ be a nondecreasing rearrangement of the multiset $\{x^t\}_{t \leq T'}$, and let $p(1), \dots, p(T')$ be the corresponding permuted predictions. Then

$$A = \sum_{t \leq T'} |p^t - x^t| = \sum_{i=1}^{T'} |p(i) - x(i)|.$$

Define the linear proxy sequence

$$z_i := \frac{1}{4} + \frac{i-1}{2(T'-1)}, \quad i = 1, \dots, T',$$

and let $d := z_{i+1} - z_i = \frac{1}{2(T'-1)}$.

Lemma 13 (Order statistics of the round-robin grid). *Let $(x^t)_{t \leq T'}$ be the round-robin sequence over the grid $\{x_1, \dots, x_m\}$ and let $x(1) \leq \dots \leq x(T')$ be a nondecreasing rearrangement of the multiset $\{x^t\}_{t \leq T'}$. Let $(z_i)_{i \leq T'}$ be defined above. Then*

$$|x(i) - z_i| \leq \frac{1}{m} \quad \text{for all } i = 1, \dots, T'.$$

Proof. More concretely, the empirical CDF of the multiset $\{x^t\}_{t \leq T'}$ places mass only on the m grid points x_1, \dots, x_m , each of which appears either $\lfloor T'/m \rfloor$ or $\lceil T'/m \rceil$ times; thus the sorted sequence $(x(i))_{i \leq T'}$ consists of m contiguous flat blocks. Over any such block, the proxy values (z_i) vary by at most

$$\frac{\lceil T'/m \rceil - 1}{2(T'-1)} \leq \frac{1}{2m},$$

while adjacent grid points satisfy $x_{j+1} - x_j = \frac{1}{2(m-1)} \leq \frac{1}{m}$. It follows that every $x(i)$ in the block lies within at most $1/m$ of the corresponding z_i , as claimed. \square

Using $|a - b| \geq |a - c| - |c - b|$ with $(a, b, c) = (p(i), x(i), z_i)$,

$$|p(i) - x(i)| \geq |p(i) - z_i| - \frac{1}{m}.$$

Summing over i gives

$$A \geq \sum_{i=1}^{T'} |p(i) - z_i| - \frac{T'}{m}.$$

Now partition indices by prediction value: for each $v \in \mathcal{V}_{T'}$, let $S_v := \{i : p(i) = v\}$ so that $|S_v| = n_v$ and $\{1, \dots, T'\} = \bigsqcup_{v \in \mathcal{V}_{T'}} S_v$. Then

$$\sum_{i=1}^{T'} |p(i) - z_i| = \sum_{v \in \mathcal{V}_{T'}} \sum_{i \in S_v} |v - z_i| \geq \sum_{v \in \mathcal{V}_{T'}} \min_{u \in \mathbb{R}} \sum_{i \in S_v} |u - z_i|.$$

Fix v and write $n := n_v$. Let $i_1 < \dots < i_n$ be the elements of S_v in increasing order, and define $a_j := z_{i_j}$ for $j = 1, \dots, n$. Since (z_i) is an arithmetic progression with step d and $i_{j+1} \geq i_j + 1$, we have $a_{j+1} - a_j \geq d$.

Claim. Let $a_1 \leq \dots \leq a_n$ satisfy $a_{j+1} - a_j \geq d$ for all j . Then

$$\min_{u \in \mathbb{R}} \sum_{j=1}^n |u - a_j| \geq d \left\lfloor \frac{n^2}{4} \right\rfloor \geq \frac{d}{4}(n^2 - 1).$$

Proof. Write $n = 2q$ or $n = 2q + 1$. If $n = 2q$ is even, any minimizer lies in $[a_q, a_{q+1}]$ and the minimum equals $\sum_{j=1}^q (a_{q+j} - a_j)$. By telescoping and the gap condition, $a_{q+j} - a_j \geq qd$ for each j , hence the minimum is at least q^2d .

If $n = 2q + 1$ is odd, the unique minimizer is $u^\star = a_{q+1}$ and the minimum equals $\sum_{j=1}^q (a_{q+1+j} - a_j)$. Again by telescoping, $a_{q+1+j} - a_j \geq (q + 1)d$, so the minimum is at least $q(q + 1)d$. \square

Applying the claim to each bucket v and summing over v yields

$$\sum_{i=1}^{T'} |p(i) - z_i| \geq \sum_{v \in \mathcal{V}_{T'}} \frac{d}{4} (n_v^2 - 1) = \frac{d}{4} \sum_{v \in \mathcal{V}_{T'}} n_v^2 - \frac{d}{4} |\mathcal{V}_{T'}|.$$

Since $|\mathcal{V}_{T'}| \leq T'$ and $d = \frac{1}{2(T'-1)} \leq \frac{1}{T'}$ for $T' \geq 2$, we have $\frac{d}{4} |\mathcal{V}_{T'}| \leq \frac{1}{4}$. Therefore

$$\sum_{i=1}^{T'} |p(i) - z_i| \geq \frac{d}{4} \sum_{v \in \mathcal{V}_{T'}} n_v^2 - \frac{1}{4}.$$

Plugging this into the bound on A above and using that $d = \frac{1}{2(T'-1)} \geq \frac{1}{4T'}$ for $T' \geq 2$ gives

$$A \geq \frac{1}{16T'} \sum_{v \in \mathcal{V}_{T'}} n_v^2 - \frac{T'}{m} - 1,$$

as claimed. \square

Lemma 12 gives us an upper bound on $\sum_v n_v^2$. Recall that by definition $\sum_v n_v = T'$. The following corollary uses this fact to convert our upper bound into a lower bound on $\sum_v \sqrt{n_v}$: intuitively, controlling the second moment $\sum_v n_v^2$ prevents too much of the total mass T' from concentrating on a few heavily used prediction values, which forces many moderately populated buckets and hence makes $\sum_v \sqrt{n_v}$ large.

Corollary 1 (Diverse predictions from ℓ_1 truthfulness). *Let $n_v := |\{t \leq T' : p^t = v\}|$, $N := \sum_{v \in \mathcal{V}_{T'}} \sqrt{n_v}$, and $A := \sum_{t \leq T'} |p^t - x^t|$. Then*

$$N \geq \frac{T'}{4\sqrt{A + \frac{T'}{m}} + 1}, \quad \text{and thus } \mathbb{E}[N] \geq \frac{T'}{4\sqrt{\mathbb{E}[A] + \frac{T'}{m}} + 1}.$$

Proof. From Lemma 12, rearranging gives

$$\sum_v n_v^2 \leq 16T' \left(A + \frac{T'}{m} + 1 \right).$$

Define $a_v := \sqrt{n_v/T'} \geq 0$. Then $\|a\|_2^2 = \sum_v a_v^2 = \sum_v n_v/T' = 1$, and

$$\|a\|_4^4 = \sum_v a_v^4 = \sum_v \frac{n_v^2}{(T')^2} \leq 16 \left(\frac{A}{T'} + \frac{1}{m} + \frac{1}{T'} \right).$$

Lemma 14 (Norm interpolation). *Let $a \in \mathbb{R}_{\geq 0}^d$ satisfy $\|a\|_2 = 1$. Then $\|a\|_1 \geq \|a\|_4^{-2}$.*

Proof. By Hölder's inequality, for $\alpha \in [0, 1]$ satisfying $\frac{1}{2} = \frac{\alpha}{1} + \frac{1-\alpha}{4}$, we have $\|a\|_2 \leq \|a\|_1^\alpha \|a\|_4^{1-\alpha}$. Solving gives $\alpha = 1/3$, hence $1 = \|a\|_2 \leq \|a\|_1^{1/3} \|a\|_4^{2/3}$. Raising both sides to the power 3 yields $1 \leq \|a\|_1 \|a\|_4^2$, i.e. $\|a\|_1 \geq \|a\|_4^{-2}$. \square

By Lemma 14, we conclude the pathwise inequality:

$$N = \sum_v \sqrt{n_v} = \sqrt{T'} \sum_v a_v = \sqrt{T'} \|a\|_1 \geq \sqrt{T'} \cdot \|a\|_4^{-2} \geq \frac{T'}{4\sqrt{A + \frac{T'}{m} + 1}}.$$

Now note that the map $u \mapsto 1/\sqrt{u}$ is convex on $(0, \infty)$, so $u \mapsto 1/\sqrt{u + \frac{T'}{m} + 1}$ is convex as well. Taking expectations and applying Jensen's inequality gives

$$\mathbb{E}[N] \geq T' \cdot \mathbb{E} \left[\frac{1}{4\sqrt{A + \frac{T'}{m} + 1}} \right] \geq \frac{T'}{4\sqrt{\mathbb{E}[A] + \frac{T'}{m} + 1}}. \quad \square$$

4.5 Controlling Bias Using Hadamard Groups

We have now established that multicalibration with respect to the global Walsh groups forces $N := \sum_v \sqrt{n_v}$ to be large. If our algorithm made “honest” predictions $\mathbb{E}[y^t] = x^t$ at each round, our predictions would be (by definition) unbiased, and calibration error would arise only from “noise”—the fact that averages of the realizations y^t will be anti-concentrated around their expectations, resulting in calibration error scaling as N .

An arbitrary algorithm need not make honest predictions, however. Our strategy is to decompose the prediction error at each round as:

$$p^t - y^t = \underbrace{(p^t - x^t)}_{\text{bias } \delta_t} + \underbrace{(x^t - y^t)}_{\text{noise } Z_t}$$

The *bias* $\delta_t := p^t - x^t$ reflects the algorithm’s deviation from the honest prediction; the *noise* $Z_t := x^t - y^t$ is the random deviation of the label from its mean. We will control these separately.

In this section, we show how to use multicalibration with respect to the Hadamard groups to control the bias term. The key observation is that Parseval’s identity (Proposition 2) lets us relate the *average* squared bias coefficient across Hadamard directions to the total bias energy. Since multicalibration controls the error in each Hadamard direction, on average the bias contribution must be small.

In what follows, the only properties of the block Hadamard system $\{\psi_{a,j}\}$ that are used are its orthogonality on each block (H1) and the property $\psi_{a,j}(s) \in \{\pm 1\}$ for all a, j, s (H2).

Let $\delta_t := p^t - x^t$ denote the *bias* of the prediction at round t , i.e. its difference relative to the honest prediction $\mathbb{E}[y^t] = x^t$. For each block J_a , define its counts $n_{v,a} := |\{t \in J_a : p^t = v\}|$ and

$$N_a := \sum_v \sqrt{n_{v,a}}, \quad E_a := \sum_{t \in J_a} \delta_t^2, \quad q_a := |\{p^t : t \in J_a\}|.$$

Our goal in this section is to relate the blockwise bias energies E_a to calibration error via the Hadamard groups, so that any substantial bias (in aggregate) forces large multicalibration error.

First, we relate the sum of the blockwise quantities N_a to N :

Lemma 15 (Block mass inequalities). *Pathwise, we have $\sum_{a=1}^K N_a \geq N$ and $\sum_{a=1}^K N_a \leq \sqrt{K}N$.*

Proof. For each v , we have $n_v = \sum_a n_{v,a}$, so $\sqrt{n_v} = \sqrt{\sum_a n_{v,a}} \leq \sum_a \sqrt{n_{v,a}}$, and summing this inequality over v gives $N \leq \sum_a N_a$.

For the other direction, by Cauchy–Schwarz, $\sum_a \sqrt{n_{v,a}} \leq \sqrt{K} \sqrt{\sum_a n_{v,a}} = \sqrt{K} \sqrt{n_v}$, and summing over v yields $\sum_a N_a \leq \sqrt{K}N$. \square

Define the *signed Hadamard bias* for $h_{a,j}$:

$$D_v^{(a,j)} := \sum_{t \in J_a : p^t = v} \psi_{a,j}(t - (a-1)L - 1) \delta_t.$$

Intuitively, for each block J_a and bucket value v , the quantity $D_v^{(a,j)}$ is the Hadamard coefficient of the bias sequence $(\delta_t)_{t \in J_a}$ restricted to those rounds with $p^t = v$.

The next lemma records two basic consequences of this viewpoint. The first one is a direct consequence of Parseval’s identity: on average over Hadamard functions j , the squared bias coefficients $(D_v^{(a,j)})^2$ recover the total bias energy E_a on block J_a . The second inequality bounds the average ℓ_1 mass $\sum_v |D_v^{(a,j)}|$ in terms of $\sqrt{q_a E_a}$ (and hence $\sqrt{N_a E_a}$), which will let us convert bias energy E_a into large calibration error for some Hadamard group.

Lemma 16 (Hadamard bias averaging). *For each block a ,*

$$\frac{1}{L} \sum_{j=0}^{L-1} \sum_v (D_v^{(a,j)})^2 = E_a, \quad \text{and} \quad \frac{1}{L} \sum_{j=0}^{L-1} \sum_v |D_v^{(a,j)}| \leq \sqrt{q_a E_a} \leq \sqrt{N_a E_a}.$$

Proof. For each v , define $A_v \in \mathbb{R}^L$ by $A_v[s] := \delta_{(a-1)L+1+s}$ if $p^{(a-1)L+1+s} = v$, and 0 otherwise. Then $D_v^{(a,j)} = \langle \psi_{a,j}, A_v \rangle$. By Proposition 2 (Parseval), we obtain $\sum_j \langle \psi_{a,j}, A_v \rangle^2 = L \|A_v\|_2^2$. Summing this over v yields the first claimed identity, since $\sum_v \|A_v\|_2^2 = \sum_{t \in J_a} \delta_t^2 = E_a$.

Next, fix j . By Cauchy–Schwarz over the q_a nonzero buckets, $\sum_v |D_v^{(a,j)}| \leq \sqrt{q_a} \sqrt{\sum_v (D_v^{(a,j)})^2}$. Averaging over j and applying Jensen’s inequality to $x \mapsto \sqrt{x}$, we obtain

$$\frac{1}{L} \sum_j \sum_v |D_v^{(a,j)}| \leq \sqrt{q_a} \cdot \sqrt{\frac{1}{L} \sum_j \sum_v (D_v^{(a,j)})^2} = \sqrt{q_a E_a}.$$

Finally, $q_a \leq N_a$, as each distinct prediction value used in the block contributes at least 1 to N_a . \square

4.6 Controlling the Noise Contribution under Adaptive Bucketing

We have seen that the Hadamard groups control the bias terms; here we control the noise terms.

The core difficulty is that the forecaster may perform *adaptive noise bucketing*: it observes past noise realizations Z_1, \dots, Z_{t-1} before choosing which “bucket” (prediction value) to place the next noise increment Z_t in. In principle, this adaptivity could allow the forecaster to create cancellations—for example, by attempting to arrange for cancellations of positive noise increments with negative cumulative sums, and vice versa.

The main technical result of this section (Theorem 3) shows that no such strategy can substantially reduce the total bucketed noise magnitude. Up to a logarithmic factor, the noise contribution

$\sum_v |B_v|$ must be at least as large as $\sum_v \sqrt{n_v}$, where n_v is the number of noise increments routed to bucket v . This matches (up to logs) what would happen under non-adaptive routing.

The proof proceeds by analyzing the “returns to zero” of each bucket’s random walk: each time a bucket’s cumulative sum returns to zero, it starts a fresh excursion. Using classical random walk estimates, we show that the expected square-root length of each excursion is $O(\log L)$, which limits how much the forecaster can save by adaptive routing.

Adaptive bucketing We now isolate the probabilistic statement we need as an abstract “noise routing” problem on a block. Let $(Z_t)_{t=1}^L$ be i.i.d. with $\mathbb{E}[Z_t] = 0$ and $Z_t \in \{\pm h\}$ for some $h \in (0, 1]$. Let \mathcal{H} be any σ -field independent of $(Z_t)_{t=1}^L$, and suppose a (possibly randomized) strategy chooses a bucket label v_t on each round t before observing Z_t , i.e. v_t is measurable with respect to

$$\mathcal{F}_{t-1} := \mathcal{H} \vee \sigma(Z_1, \dots, Z_{t-1}), \quad t = 1, \dots, L.$$

Define the bucket counts and bucket noise sums

$$n_v := |\{t \in \{1, \dots, L\} : v_t = v\}|, \quad B_v := \sum_{t: v_t=v} Z_t.$$

Now, intuitively, observe that if the buckets $(v_t)_{t=1}^L$ were fixed independently of $(Z_t)_{t=1}^L$, then each B_v would be a length- n_v simple random walk with step size h , so typically $|B_v| = \Theta(h\sqrt{n_v})$ and $\sum_v |B_v| = \Theta(h \sum_v \sqrt{n_v})$. The following key theorem that we prove below shows that even under adaptive bucketing this baseline can be reduced by at most an $O(\log L)$ factor.

Theorem 3 (Adaptive Noise Bucketing). *There exists a universal constant $C_{\text{rev}} > 0$ such that for every $L \geq 2$, every $h \in (0, 1]$, and every adaptive bucketing strategy as above (such that every v_t is \mathcal{F}_{t-1} -measurable), it holds that*

$$\mathbb{E} \left[\sum_v |B_v| \right] \geq \frac{C_{\text{rev}} h}{\log(L+1)} \mathbb{E} \left[\sum_v \sqrt{n_v} \right].$$

In Corollary 2, we will then apply this theorem with $h = 1/4$ inside each block J_a , to the signed noise increments obtained from $(x^t - y^t)$ after multiplying by the fixed block sign $\psi_{a,j}$, and with bucket labels given by the realized predictions.

4.6.1 Proof Roadmap

We now outline the proof of Theorem 3. The proof proceeds in two steps.

Step 1: Reduce to bounding the expected number of returns. For each bucket v , define its cumulative sum up to each time t as

$$B_v(t) := \sum_{s \leq t: v_s=v} Z_s, \quad t = 0, 1, \dots, L,$$

and write $B_v := B_v(L)$. Fix $\varepsilon := h/4$, so for any $x \in h\mathbb{Z}$ we have $|x| \leq \varepsilon \iff x = 0$. Define

$$L_\varepsilon := \sum_{t=1}^L \mathbf{1}[|B_{v_t}(t-1)| \leq \varepsilon] = \sum_{t=1}^L \mathbf{1}[B_{v_t}(t-1) = 0].$$

In Lemma 17 we show that on rounds with $B_{v_t}(t-1) = 0$, the conditional expected increase in $\sum_v |B_v|$ is h , while on other rounds it is nonnegative, and so

$$\mathbb{E}\left[\sum_v |B_v|\right] \geq h\mathbb{E}[L_\varepsilon].$$

Step 2: Excursions imply L_ε must be large if $\sum_v \sqrt{n_v}$ is large. For a bucket v , the evolution of its noise *at the times it is selected* is a simple random walk on $h\mathbb{Z}$ starting at 0. Each time it returns to 0 it starts a new *excursion*. Let R_v be the number of excursions for bucket v , so that $L_\varepsilon = \sum_v R_v$. Writing the excursion lengths for bucket v as $\ell_1^v, \dots, \ell_{R_v}^v$, so that $n_v = \sum_{j=1}^{R_v} \ell_j^v$, we have $\sqrt{n_v} \leq \sum_{j=1}^{R_v} \sqrt{\ell_j^v}$. To bound this quantity for each v , we prove a truncated return-time bound for a simple random walk: for an excursion length ℓ (truncated at horizon L) one has:

$$\mathbb{E}[\sqrt{\ell}] \leq c \log(L+1) \text{ for a universal constant } c > 0.$$

Applying this excursion-by-excursion, summing over buckets, and recalling Step 1 then yields

$$\mathbb{E}\left[\sum_v \sqrt{n_v}\right] \leq c \log(L+1) \mathbb{E}[L_\varepsilon] \implies \mathbb{E}\left[\sum_v |B_v|\right] \geq \frac{h}{c \log(L+1)} \mathbb{E}\left[\sum_v \sqrt{n_v}\right].$$

4.6.2 Proof of Theorem 3: Step 1

Lemma 17 (Lower-bounding noise by returns count). *Under the setup defined above, we have:*

$$\mathbb{E}\left[\sum_v |B_v|\right] \geq h\mathbb{E}[L_\varepsilon].$$

Proof. Let $\Phi_t := \sum_v |B_v(t)|$ for $0 \leq t \leq L$, with $\Phi_0 = 0$ and $\sum_v |B_v| = \Phi_L$. Since only bucket v_t changes at time t , the one-round increment satisfies:

$$\Phi_t - \Phi_{t-1} = |B_{v_t}(t-1) + Z_t| - |B_{v_t}(t-1)|.$$

Condition on \mathcal{F}_{t-1} and abbreviate $b := B_{v_t}(t-1)$. If $|b| \leq \varepsilon$, then $b = 0$ (as $\varepsilon = h/4$ and $b \in h\mathbb{Z}$), so $\Phi_t - \Phi_{t-1} = |Z_t| = h$ deterministically. Otherwise, by convexity of $x \mapsto |x|$ and $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$,

$$\mathbb{E}[\Phi_t - \Phi_{t-1} | \mathcal{F}_{t-1}] = \mathbb{E}[|b + Z_t| | \mathcal{F}_{t-1}] - |b| \geq 0.$$

Therefore, for every t ,

$$\mathbb{E}[\Phi_t - \Phi_{t-1} | \mathcal{F}_{t-1}] \geq h \cdot 1[|B_{v_t}(t-1)| \leq \varepsilon].$$

Taking expectations, summing over t , and telescoping yields $\mathbb{E}\left[\sum_v |B_v|\right] = \mathbb{E}[\Phi_L] \geq h\mathbb{E}[L_\varepsilon]$. \square

4.6.3 Proof of Theorem 3: Step 2

Fix a bucket v with $n_v \geq 1$. Enumerate its update times as $1 \leq t_{v,1} < \dots < t_{v,n_v} \leq L$, setting $t_{v,0} := 0$ for convenience. Define the local-time partial sums $S_0^v := 0$ and $S_k^v := \sum_{i=1}^k Z_{t_{v,i}}$ for $k = 1, \dots, n_v$. Thus $B_v = S_{n_v}^v$.

Define the local “near-zero-before-update” indicators

$$A_k^v := \mathbf{1}\{|S_{k-1}^v| \leq \varepsilon\} = \mathbf{1}[S_{k-1}^v = 0], \quad k = 1, \dots, n_v,$$

where the equality uses $S_{k-1}^v \in h\mathbb{Z}$ and the choice $\varepsilon = h/4$.

Define the corresponding count R_v :

$$R_v := \sum_{k=1}^{n_v} A_k^v, \quad \text{so that } L_\varepsilon = \sum_v R_v.$$

Define renewal indices by $\kappa_1^v := 1$ and, for $j \geq 1$,

$$\kappa_{j+1}^v := \min\{k > \kappa_j^v : A_k^v = 1\},$$

with the convention $\kappa_{R_v+1}^v := n_v + 1$. Define excursion lengths

$$\ell_j^v := \kappa_{j+1}^v - \kappa_j^v, \quad j = 1, \dots, R_v.$$

Lemma 18 (Subadditivity decomposition). *For every bucket v , $\sqrt{n_v} \leq \sum_{j=1}^{R_v} \sqrt{\ell_j^v}$. Consequently,*

$$\sum_v \sqrt{n_v} \leq \sum_v \sum_{j=1}^{R_v} \sqrt{\ell_j^v}.$$

Proof. By construction, we have $n_v = \sum_{j=1}^{R_v} \ell_j^v$. Now applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, iterated over the R_v terms, and summing over v yields the inequality. \square

The following result bounds the expected truncated root-length of a random walk excursion.

Proposition 3 (Return-time bound for increments). *Let $h \in (0, 1]$ and let $(X_k)_{k \geq 1}$ be i.i.d. with $\mathbb{P}(X_k = h) = \mathbb{P}(X_k = -h) = 1/2$. Let $S_n := \sum_{k=1}^n X_k$ with $S_0 := 0$ and define the first return time*

$$\tau_0 := \inf\{n \geq 1 : S_n = 0\}.$$

Then there exists a universal constant $C_{\text{ret}} > 0$ such that for all integers $L \geq 2$,

$$\mathbb{E}[\sqrt{\min(\tau_0, L)}] \leq C_{\text{ret}} \log(L+1).$$

Moreover, the same bound holds conditionally on any past σ -field \mathcal{G} that is independent of the future increments $(X_k)_{k \geq 1}$ (so that, given \mathcal{G} , the increments remain i.i.d. with the same $\{\pm h\}$ law).

Proof. Consider the rescaled random walk: let $\tilde{X}_k := X_k/h \in \{\pm 1\}$ and $\tilde{S}_n := S_n/h$. Then, $\tau_0 = \inf\{n \geq 1 : \tilde{S}_n = 0\}$, which does not depend on h , so it suffices to prove the claim for $h = 1$.

Recall the standard fact [Feller, 1968, Lemma 2 and Eq. 3.7, p. 78] that for the simple symmetric walk with steps ± 1 , its first return time τ_0 is even a.s., and satisfies for $n \geq 1$ that:

$$\mathbb{P}(\tau_0 = 2n) = \frac{1}{2n-1} \binom{2n}{n} 4^{-n}.$$

Using $\binom{2n}{n} \leq 4^n/\sqrt{\pi n}$, we obtain $\mathbb{P}(\tau_0 = 2n) \leq c_0 n^{-3/2}$ for a universal $c_0 > 0$, so for all $m \geq 1$,

$$\mathbb{P}(\tau_0 > m) = \sum_{k>\lceil m/2 \rceil} \mathbb{P}(\tau_0 = 2k) \leq c_0 \sum_{k>\lceil m/2 \rceil} k^{-3/2} \leq \frac{c_2}{\sqrt{m}}$$

for a universal $c_2 > 0$. We now have for $L \geq 2$ that

$$\mathbb{E}[\sqrt{\min(\tau_0, L)}] = \sum_{m=0}^{L-1} (\sqrt{m+1} - \sqrt{m}) \mathbb{P}(\min(\tau_0, L) > m) = \sum_{m=0}^{L-1} (\sqrt{m+1} - \sqrt{m}) \mathbb{P}(\tau_0 > m).$$

Since $\sqrt{m+1} - \sqrt{m} \leq (2\sqrt{m})^{-1}$ for $m \geq 1$, the above tail bound gives

$$\mathbb{E}[\sqrt{\min(\tau_0, L)}] \leq 1 + \sum_{m=1}^{L-1} \frac{1}{2\sqrt{m}} \cdot \frac{c_2}{\sqrt{m}} = 1 + \frac{c_2}{2} \sum_{m=1}^{L-1} \frac{1}{m} \leq C_{\text{ret}} \log(L+1).$$

If \mathcal{G} is any σ -field independent of the future increments $(X_k)_{k \geq 1}$, then conditionally on \mathcal{G} the increments remain i.i.d. with the same law, so the same bound holds for $\mathbb{E}[\sqrt{\min(\tau_0, L)} | \mathcal{G}]$. \square

Having proved Proposition 3, we now relate it back to the cumulative bucket noise in our setup.

Lemma 19 (Excursions have logarithmic expected root-length). *Under the assumptions of Theorem 3 and with the choice of ε fixed above, there exists a constant $C_{\text{ret}} > 0$ such that*

$$\mathbb{E}\left[\sum_v \sqrt{n_v}\right] \leq C_{\text{ret}} \log(L+1) \mathbb{E}[L_\varepsilon]. \quad (21)$$

Proof. We will first prove an intermediate claim, which implies the result of the lemma: for every bucket v and every excursion index $j \in \{1, \dots, R_v\}$, we will now show that

$$\mathbb{E}[\sqrt{\ell_j^v} | \mathcal{F}_{t_{v,\kappa_j^v}-1}] \leq C_{\text{ret}} \log(L+1). \quad (22)$$

Towards this, fix v, j and set $\mathcal{G} := \mathcal{F}_{t_{v,\kappa_j^v}-1}$. By definition of κ_j^v we have $|S_{\kappa_j^v-1}^v| \leq \varepsilon$, and by the standing lattice assumption this implies $S_{\kappa_j^v-1}^v = 0$.

Because the bucketing strategy is predictable (each v_t is \mathcal{F}_{t-1} -measurable), the update times satisfy $\{t_{v,k} = t\} \in \mathcal{F}_{t-1}$ for every t .

Thus, conditionally on $\mathcal{F}_{t_{v,k}-1}$, the increment $Z_{t_{v,k}}$ has the same $\{\pm h\}$ law as Z_1 . Since each event $\{t_{v,k} = t\}$ is \mathcal{F}_{t-1} -measurable, Z_t is independent of \mathcal{F}_{t-1} , and the Z_t 's are independent across t , it follows by induction that $X_1 := Z_{t_{v,\kappa_j^v}}, X_2 := Z_{t_{v,\kappa_j^v+1}}, \dots$ are i.i.d. given $\mathcal{F}_{t_{v,\kappa_j^v}-1}$, with $\mathbb{P}(X_n = h) = \mathbb{P}(X_n = -h) = 1/2$.

Define the local random walk by $T_0 := 0$ and $T_n := \sum_{r=1}^n X_r$, and let the first return time be $\tau_{v,j}^0 := \inf\{n \geq 1 : T_n = 0\}$. For $j < R_v$ we have $\ell_j^v = \tau_{v,j}^0$, while for the final excursion $j = R_v$ we have $\ell_j^v \leq \min(\tau_{v,j}^0, L)$. Hence $\sqrt{\ell_j^v} \leq \sqrt{\min(\tau_{v,j}^0, L)}$.

Conditional on $\mathcal{F}_{t_{v,\kappa_j^v}-1}$, the process $(T_n)_{n \geq 0}$ satisfies the assumptions of Proposition 3. Hence, by Proposition 3, for a universal constant C_{ret} we have:

$$\mathbb{E}[\sqrt{\ell_j^v} | \mathcal{F}_{t_{v,\kappa_j^v}-1}] \leq \mathbb{E}[\sqrt{\min(\tau_{v,j}^0, L)} | \mathcal{F}_{t_{v,\kappa_j^v}-1}] \leq C_{\text{ret}} \log(L+1),$$

and our intermediate claim (22) follows.

To now establish (21), we start from Lemma 18:

$$\mathbb{E}\left[\sum_v \sqrt{n_v}\right] \leq \mathbb{E}\left[\sum_v \sum_{j=1}^{R_v} \sqrt{\ell_j^v}\right].$$

Since $R_v \leq n_v \leq L$, we may write

$$\mathbb{E}\left[\sum_{j=1}^{R_v} \sqrt{\ell_j^v}\right] = \sum_{j=1}^L \mathbb{E}\left[\mathbf{1}[j \leq R_v] \sqrt{\ell_j^v}\right] = \sum_{j=1}^L \mathbb{E}\left[\mathbf{1}[j \leq R_v] \mathbb{E}\left[\sqrt{\ell_j^v} | \mathcal{F}_{t_{v,\kappa_j^v}-1}\right]\right] \leq C_{\text{ret}} \log(L+1) \mathbb{E}[R_v],$$

where the inner conditional expectation is invoked only on the event $\{j \leq R_v\}$ (on which κ_j^v and ℓ_j^v are defined), and the inequality uses Equation (22) and $\sum_{j=1}^L \mathbf{1}[j \leq R_v] = R_v$.

Summing the last display over v and recalling that $\sum_v R_v = L_\varepsilon$ concludes the proof. \square

4.6.4 Finishing the proof of Theorem 3

Combining Lemma 17 and Lemma 19, we obtain:

$$\mathbb{E}\left[\sum_v \sqrt{n_v}\right] \leq C_{\text{ret}} \log(L+1) \mathbb{E}[L_\varepsilon] \leq \frac{C_{\text{ret}}}{h} \log(L+1) \mathbb{E}\left[\sum_v |B_v|\right].$$

Setting $C_{\text{rev}} := 1/C_{\text{ret}}$ and rearranging, we obtain the claimed bound of Theorem 3.

4.6.5 Applying Theorem 3 to the Noise Contribution

For block J_a and block sign $\psi_{a,j}$, define the signed noise sums

$$N_v^{(a,j)} := \sum_{t \in J_a : p^t = v} \psi_{a,j}(t - (a-1)L - 1) (x^t - y^t).$$

For a fixed block J_a and block sign $\psi_{a,j}$, the quantities $N_v^{(a,j)}$ collect the contribution of the “noise” terms $x^t - y^t$ on that block, restricted to a single prediction bucket v and modulated by the block sign. Equivalently, if we view the block in its local time coordinate $s = 1, \dots, L$, define increments

$$Z_s := \psi_{a,j}(s-1) (x^{t_s} - y^{t_s})$$

and bucket assignments $v_s := p^{t_s}$, then $N_v^{(a,j)}$ is exactly the bucket sum $\sum_{s: v_s = v} Z_s$.

The adaptive noise bucketing theorem (Theorem 3) therefore applies to each pair (a, j) with increments $(Z_s)_{s=1}^L$ and predictable bucket choices $(v_s)_{s=1}^L$.

The next corollary records the resulting lower bound: for every block and every Hadamard element, the total signed noise $\sum_v |N_v^{(a,j)}|$ cannot be much smaller (up to a logarithmic factor) than the “natural scale” $\sum_v \sqrt{n_{v,a}} = N_a$ of that block.

Corollary 2 (Noise floor for each signed Hadamard functional). *For all a, j and a universal $c_2 > 0$,*

$$\mathbb{E} \left[\sum_v |N_v^{(a,j)}| \right] \geq \frac{c_2}{\log(L+1)} \cdot \mathbb{E}[N_a].$$

Proof. Fix a and j and adopt the local-time notation introduced above: $t_s := (a-1)L + s$, $Z_s := \psi_{a,j}(s-1)(x^{t_s} - y^{t_s})$, and $v_s := p^{t_s}$, so that

$$N_v^{(a,j)} = \sum_{s: v_s = v} Z_s.$$

Under the environment (13) we have $x^{t_s} - y^{t_s} = -\xi^{t_s}/4$, so $(Z_s)_{s=1}^L$ are i.i.d. with $Z_s \in \{\pm h\}$ for $h := 1/4$ and $\mathbb{E}[Z_s] = 0$. We now verify the predictability condition in Theorem 3. Let

$$\mathcal{H} := \sigma(x^1, \dots, x^T, y^1, \dots, y^{(a-1)L}, \text{algorithm's internal randomness}).$$

Then \mathcal{H} is independent of the future labels $y^{(a-1)L+1}, \dots, y^{aL}$ and hence of $(Z_s)_{s=1}^L$. For each s , the bucket choice $v_s := p^{t_s}$ is measurable with respect to the past history $\sigma(x^1, \dots, x^{t_s}, y^1, \dots, y^{t_s-1})$, and on J_a the past labels are measurable from (Z_1, \dots, Z_{s-1}) via $y^{t_r} = x^{t_r} - \psi_{a,j}(r-1)Z_r$ for all $r < s$. Therefore v_s is measurable with respect to $\mathcal{F}_{s-1} := \mathcal{H} \vee \sigma(Z_1, \dots, Z_{s-1})$.

Next, for each bucket value v , the signed noise sum on block J_a can be written as

$$N_v^{(a,j)} = \sum_{s=1}^L \mathbf{1}[p^{t_s} = v] Z_s,$$

so $N_v^{(a,j)}$ is exactly the bucket sum B_v obtained by applying Theorem 3 to the blockwise sequence $(Z_s)_{s=1}^L$ with predictable assignment $(v_s)_{s=1}^L$ and initial σ -field \mathcal{H} .

The hypotheses of Theorem 3 are thus satisfied with $h = 1/4$, so we obtain, for $c_2 := C_{\text{rev}}/4$,

$$\mathbb{E} \left[\sum_v |N_v^{(a,j)}| \right] \geq \frac{C_{\text{rev}}}{4 \log(L+1)} \cdot \mathbb{E} \left[\sum_v \sqrt{n_{v,a}} \right] = \frac{c_2}{\log(L+1)} \mathbb{E}[N_a]. \quad \square$$

4.7 Proof of Theorem 2

Finally, we are ready to prove our main theorem, which lower-bounds the multicalibration error rate that any algorithm can guarantee for prediction-independent groups.

The proof combines all the ingredients developed above:

1. The calibration error for each signed Hadamard functional $h_{a,j}$ decomposes into noise minus bias (Equation 23).
2. The noise term is lower-bounded by $\tilde{\Omega}(N_a)$ via Corollary 2.
3. The bias term is upper-bounded on average over j by $O(\sqrt{N_a E_a})$ via Lemma 16.
4. Summing over blocks and using the ℓ_1 -truthfulness constraint to bound $\sum_a E_a$, we show that the noise dominates the bias, yielding the $\tilde{\Omega}(T^{2/3})$ lower bound.

By Lemma 9, $\text{MCerr}_{T'}(G) \geq \frac{1}{2} \max_{a,j} \text{Err}_{T'}(h_{a,j})$ so it suffices to lower-bound $\max_{a,j} \text{Err}_{T'}(h_{a,j})$. Fix (a, j) . For each bucket v on block J_a , the signed Hadamard functional satisfies:

$$\begin{aligned} & \sum_{t \leq T'} \mathbf{1}[p^t = v] h_{a,j}(c^t) (p^t - y^t) \\ &= \sum_{t \in J_a : p^t = v} \psi_{a,j}(t - (a-1)L-1) (p^t - x^t) + \sum_{t \in J_a : p^t = v} \psi_{a,j}(t - (a-1)L-1) (x^t - y^t) \\ &= D_v^{(a,j)} + N_v^{(a,j)}. \end{aligned}$$

Hence by the triangle inequality,

$$\text{Err}_{T'}(h_{a,j}) = \sum_v |D_v^{(a,j)} + N_v^{(a,j)}| \geq \sum_v |N_v^{(a,j)}| - \sum_v |D_v^{(a,j)}|. \quad (23)$$

Averaging over block Hadamard functions and blocks. We now average over Hadamard directions j within each block, and then over blocks a . The key insight is that while the *noise* term is large for every j (by Corollary 2), the *bias* term is small on average over j (by Lemma 16). This averaging argument is what allows us to find at least one Hadamard direction where noise dominates bias.

Take expectations in (23), apply Corollary 2, and then average over j :

$$\frac{1}{L} \sum_{j=0}^{L-1} \mathbb{E}[\text{Err}_{T'}(h_{a,j})] \geq \frac{c_2}{\log(L+1)} \mathbb{E}[N_a] - \frac{1}{L} \sum_{j=0}^{L-1} \mathbb{E}\left[\sum_v |D_v^{(a,j)}|\right].$$

Using Lemma 16 and then averaging over blocks a yields

$$\frac{1}{KL} \sum_{a=1}^K \sum_{j=0}^{L-1} \mathbb{E}[\text{Err}_{T'}(h_{a,j})] \geq \frac{c_2}{K \log(L+1)} \mathbb{E}\left[\sum_{a=1}^K N_a\right] - \frac{1}{K} \mathbb{E}\left[\sum_{a=1}^K \sqrt{N_a E_a}\right]. \quad (24)$$

Bounding the bias penalty. By Cauchy–Schwarz over a (pathwise),

$$\frac{1}{K} \sum_{a=1}^K \sqrt{N_a E_a} \leq \sqrt{\frac{1}{K} \sum_a N_a} \cdot \sqrt{\frac{1}{K} \sum_a E_a}.$$

By Lemma 15, $\sum_a N_a \leq \sqrt{K} N$, hence $\sqrt{\frac{1}{K} \sum_a N_a} \leq K^{-1/4} \sqrt{N}$. Also $\sum_a E_a = \sum_{t \leq T'} \delta_t^2 =: S$, so $\sqrt{\frac{1}{K} \sum_a E_a} = K^{-1/2} \sqrt{S}$. Therefore, pathwise,

$$\frac{1}{K} \sum_a \sqrt{N_a E_a} \leq \frac{1}{K^{3/4}} \sqrt{NS}.$$

Taking expectations and applying Cauchy–Schwarz gives

$$\frac{1}{K} \mathbb{E}\left[\sum_a \sqrt{N_a E_a}\right] \leq \frac{1}{K^{3/4}} \sqrt{\mathbb{E}[N] \mathbb{E}[S]}. \quad (25)$$

Finish: lower bound multicalibration error. We now combine the pieces. Since the maximum over (a, j) is at least the average, and using $\sum_a N_a \geq N$ (Lemma 15), we obtain from (24) and (25):

$$\mathbb{E} \left[\max_{a,j} \text{Err}_{T'}(h_{a,j}) \right] \geq \frac{c_2}{\log(L+1)} \cdot \frac{\mathbb{E}[N]}{K} - \frac{1}{K^{3/4}} \sqrt{\mathbb{E}[N] \mathbb{E}[S]}. \quad (26)$$

Let

$$MC := \mathbb{E}[\text{MCerr}_{T'}(G)].$$

By Lemma 9, $\text{MCerr}_{T'}(G) \geq \frac{1}{2} \max_{a,j} \text{Err}_{T'}(h_{a,j})$, so taking expectations and using (26) yields

$$MC \geq \frac{c_2}{2 \log(L+1)} \cdot \frac{\mathbb{E}[N]}{K} - \frac{1}{2K^{3/4}} \sqrt{\mathbb{E}[N] \mathbb{E}[S]}. \quad (27)$$

Define the ℓ_1 deviation from honesty

$$A := \sum_{t \leq T'} |p^t - x^t|.$$

Since $|p^t - x^t| \leq 1$ we have $(p^t - x^t)^2 \leq |p^t - x^t|$ pointwise and thus $S \leq A$. Therefore, by Lemma 11,

$$\mathbb{E}[S] \leq \mathbb{E}[A] \leq C_{\ell_1} \log(m+1) MC. \quad (28)$$

Also, by Corollary 1,

$$\mathbb{E}[N] \geq \frac{T'}{4\sqrt{\mathbb{E}[A] + \frac{T'}{m} + 1}}. \quad (29)$$

Set $\Lambda := \log(T+1)$. Recall $K = \lceil \Lambda^{10} \rceil$, so $K \geq \Lambda^{10}$ and (for $\Lambda \geq 1$) $K \leq 2\Lambda^{10}$. Also $\log(m+1) \leq \Lambda$ and $\log(L+1) \leq \Lambda$ since $m, L \leq T$.

Define

$$B := \frac{c_2}{64\sqrt{3}} \cdot \frac{\sqrt{mT'}}{K\Lambda}.$$

We show that for all sufficiently large T , one has $MC \geq B$. Suppose for contradiction that $MC < B$. Then by (28) and $\log(m+1) \leq \Lambda$,

$$\mathbb{E}[A] \leq C_{\ell_1} \log(m+1) MC < C_{\ell_1} \Lambda B = \frac{C_{\ell_1} c_2}{64\sqrt{3}} \cdot \frac{\sqrt{mT'}}{K}.$$

For T large, we have $T' = KL \geq T/2$ and $m \leq T^{1/3} + 1 \leq 2T^{1/3}$, hence

$$\frac{T'}{m} \geq \frac{T/2}{2T^{1/3}} = \frac{1}{4} T^{2/3} \geq \frac{1}{4\sqrt{2}} \sqrt{mT'}.$$

Since $K \rightarrow \infty$ with T , these inequalities imply that for all sufficiently large T ,

$$\mathbb{E}[A] \leq \frac{T'}{m} \quad \text{and} \quad \frac{T'}{m} \geq 1.$$

Plugging this into (29) gives

$$\mathbb{E}[N] \geq \frac{T'}{4\sqrt{\mathbb{E}[A] + \frac{T'}{m} + 1}} \geq \frac{T'}{4\sqrt{3 \cdot \frac{T'}{m}}} = \frac{1}{4\sqrt{3}} \sqrt{mT'}.$$

From (28) and $MC < B$ we also have

$$\mathbb{E}[S] \leq \mathbb{E}[A] \leq C_{\ell_1} \Lambda B = \frac{C_{\ell_1} c_2}{64\sqrt{3}} \cdot \frac{\sqrt{mT'}}{K}.$$

Combining with the lower bound on $\mathbb{E}[N]$ yields

$$\sqrt{\frac{\mathbb{E}[S]}{\mathbb{E}[N]}} \leq \frac{\sqrt{C_{\ell_1} c_2}}{8} \cdot \frac{1}{\sqrt{K}}.$$

Write the right-hand side of (27) as $\frac{1}{2}(\text{Noise} - \text{Bias})$ with

$$\text{Noise} := \frac{c_2}{\log(L+1)} \cdot \frac{\mathbb{E}[N]}{K}, \quad \text{Bias} := \frac{1}{K^{3/4}} \sqrt{\mathbb{E}[N] \cdot \mathbb{E}[S]}.$$

Then

$$\frac{\text{Bias}}{\text{Noise}} = \frac{\log(L+1)}{c_2} K^{1/4} \sqrt{\frac{\mathbb{E}[S]}{\mathbb{E}[N]}} \leq \frac{\sqrt{C_{\ell_1}}}{8c_2} \cdot \frac{\log(L+1)}{K^{1/4}}.$$

Using $\log(L+1) \leq \Lambda$ and $K^{1/4} \geq \Lambda^{5/2}$ gives $\text{Bias}/\text{Noise} \leq \frac{\sqrt{C_{\ell_1}}}{8c_2} \Lambda^{-3/2}$, which is at most $1/2$ for all sufficiently large T . Plugging $\text{Bias} \leq \frac{1}{2}\text{Noise}$ into (27) yields

$$MC \geq \frac{1}{4} \text{Noise} = \frac{c_2}{4\log(L+1)} \cdot \frac{\mathbb{E}[N]}{K} \geq \frac{c_2}{4\Lambda} \cdot \frac{1}{4\sqrt{3}} \cdot \frac{\sqrt{mT'}}{K} = 4B,$$

contradicting $MC < B$. Therefore $MC \geq B$ for all sufficiently large T .

Finally, for T large we have $m \geq T^{1/3}/2$ and $T' \geq T/2$, hence $\sqrt{mT'} \geq \frac{1}{2}T^{2/3}$. Also $K \leq 2\Lambda^{10}$. Thus, for all sufficiently large T , for a universal constant $c > 0$,

$$MC \geq B \geq c \cdot \frac{T^{2/3}}{\log^{11}(T+1)}.$$

This completes the proof of Theorem 2.

5 Discussion

We have established tight $\Theta(T^{2/3})$ bounds (up to logarithmic factors) for online multicalibration, separating it from marginal calibration. Several natural questions remain open. We highlight one:

Intermediate group family sizes. For prediction-independent groups, we show that constant-sized families reduce to marginal calibration and hence can be solved at $O(T^{2/3-\varepsilon})$ rates by [Dagan et al., 2025], while families of size $|G| = \Theta(T)$ are subject to the $\tilde{\Omega}(T^{2/3})$ lower bound. What happens for intermediate sizes $|G|$? Is there a sharp threshold, or does the complexity interpolate smoothly? What about for families of size $|G| = \text{polylog}(T)$? One strategy for giving better upper bounds in this case might be to give a more refined reduction from multicalibration to marginal calibration, that unlike the reduction in Appendix A incurs overhead $\text{poly}(|G|)$ rather than $2^{|G|}$. However, in Appendix B we give an ‘oracle’ lower bound that serves as an obstruction to this approach: we demonstrate an instance in which $|G| = \Theta(\log T)$ such that every reduction from multicalibration to marginal calibration in a natural family of ‘proper’ reductions that we define must instantiate at least exponentially many marginal calibration oracles in $|G|$ to obtain non-trivial multicalibration error. There may nevertheless be ways to circumvent this obstruction.

Acknowledgments

We gratefully acknowledge funding from the Simons Foundation, the UK AI Safety Institute, and the NSF ENCoRE TRIPODS Institute.

We thank Princewill Okoroafor for enlightening discussions at an early stage of this work. The authors used AI tools including GPT 5.1 Pro in the development of this paper; all of the final theorems and proofs are written and verified by the authors, and all of the exposition and discussion of related work was written without AI assistance.

References

- Jacob Abernethy, Peter L Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 27–46. JMLR Workshop and Conference Proceedings, 2011.
- Krishna Acharya, Eshwar Ram Arunachaleswaran, Sampath Kannan, Aaron Roth, and Juba Ziani. Oracle efficient algorithms for groupwise regret. In *The Twelfth International Conference on Learning Representations*, 2024.
- Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In *Innovations in Theoretical Computer Science Conference (ITCS)*, volume 11, 2020.
- Donald L Burkholder. Distribution function inequalities for martingales. *The Annals of Probability*, 1(1):19–42, 1973.
- Silvia Casacuberta, Cynthia Dwork, and Salil Vadhan. Complexity-theoretic implications of multicalibration. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1071–1082, 2024.
- Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1532–1543, 2025.
- Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. Collaborative prediction: Tractable information aggregation via agreement. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2026.
- Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. Breaking the $T^{2/3}$ barrier for sequential calibration. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 2007–2018, 2025.
- A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.
- A Philip Dawid. Calibration-based empirical probability. *The Annals of Statistics*, 13(4):1251–1274, 1985.
- Zhun Deng, Cynthia Dwork, and Linjun Zhang. Happymap: A generalized multicalibration method. In *14th Innovations in Theoretical Computer Science Conference, ITCS 2023*, page 41. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2023.

- Cynthia Dwork and Pranay Tankala. Supersimulators. *arXiv preprint arXiv:2509.17994*, 2025.
- William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, New York, 3 edition, 1968.
- Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.
- Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2725–2792. SIAM, 2024.
- Rohan Ghuge, Vidya Muthukumar, and Sahil Singla. Improved and oracle-efficient online ℓ_1 -multicalibration. *arXiv preprint arXiv:2505.17365*, 2025.
- Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In *International Conference on Machine Learning*, pages 11459–11492. PMLR, 2023.
- Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pages 79–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022a.
- Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022b.
- Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, 2023a.
- Parikshit Gopalan, Michael Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. *Advances in Neural Information Processing Systems*, 36: 39936–39956, 2023b.
- Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pages 82–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022.
- Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. *Advances in Neural Information Processing Systems*, 36: 72464–72506, 2023.
- Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. Truthfulness of calibration measures. *Advances in Neural Information Processing Systems*, 37:117237–117290, 2024.
- Sergiu Hart. Calibrated forecasts: The minimax proof. In *Matching, Dynamics and Games for the Allocation of Resources: Essays in Celebration of David Gale’s 100th Birthday*, pages 153–159. Springer, 2025.
- Jason Hartline, Lunjia Hu, and Yifan Wu. A perfectly truthful calibration measure. *arXiv preprint arXiv:2508.13100*, 2025.

- Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.
- Lunjia Hu, Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Efficient swap multicalibration of elicitable properties. *arXiv preprint arXiv:2511.04907*, 2025.
- Sham M Kakade and Dean P Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.
- Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022.
- Daniel Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. *Advances in Neural Information Processing Systems*, 35:29051–29063, 2022.
- Ehud Lehrer. Approachability in infinite dimensional spaces. *International Journal of Game Theory*, 31(2):253–268, 2003.
- Georgy Noarov and Aaron Roth. The statistical scope of multicalibration. In *International Conference on Machine Learning*, pages 26283–26310. PMLR, 2023.
- Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. In *Forty-second International Conference on Machine Learning*, 2025.
- Juan Carlos Perdomo and Benjamin Recht. In defense of defensive forecasting. *arXiv preprint arXiv:2506.11848*, 2025.
- Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 456–466, 2021.
- Mingda Qiao and Eric Zhao. Truthfulness of decision-theoretic calibration measures. *arXiv preprint arXiv:2503.02384*, 2025.
- Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28(1):141–153, 2003.

A Reducing Constant Sized Families of Binary Prediction Independent Groups to Marginal Calibration

The lower bound in Theorem 1 crucially exploits the fact that the groups g_1, g_2, g_3 are allowed to depend on the prediction value v . In this section we show that an analogous lower bound cannot hold for a constant number (i.e. not growing with T) of binary *prediction-independent* groups—groups that depend only on the context x and not on the prediction v .

To do this we show a simple reduction from the problem of sequential adversarial multicalibration (for prediction independent groups) to the problem of sequential adversarial *marginal* calibration, with rates that degrade at most exponentially with the number of groups to be multicalibrated. An implication of this is that for any constant number of groups, the best rate for multicalibration is the same (up to constant factors depending on the number of groups) as it is for marginal calibration — and in particular, by the recent result of Dagan et al. [2025], $O(T^{2/3-\epsilon})$.

The idea is extremely simple: if we have k prediction independent binary groups, then before we make our prediction, we know which of the at most 2^k *combinations of groups* are active at each round t before we must make our prediction. The sequence of rounds on which each combination of groups is active is by construction disjoint from every other. Thus we can instantiate a separate copy of any marginal calibration algorithm for each of the 2^k possible combinations of groups and run each on the corresponding subsequence. Naively this results in a blow-up in rates of 2^k . In this section we give a somewhat more refined bound that depends for each group on the number of distinct combinations of groups that it participates in, and further takes advantage of the convexity of calibration error upper bounds.

Theorem 4 (Structure-aware prediction-independent multicalibration). *Let $k \in \mathbb{N}$ and let $G = \{g_1, \dots, g_k\}$ be a family of prediction-independent groups. Suppose there exists an online prediction algorithm A^{marg} and a function $R : \mathbb{N} \rightarrow [0, \infty)$ that is nondecreasing and satisfies $R(0) = 0$, such that for every horizon $n \geq 1$ and every sequence $(x^t, y^t)_{t=1}^n$, when A^{marg} is run for n rounds we have:*

$$\mathbb{E}\left[\sum_{v \in V_n} \left| \sum_{t=1}^n \mathbf{1}[p^t = v] (p^t - y^t) \right| \right] \leq R(n).$$

where the expectation is over the internal randomness of A^{marg} . Assume moreover that R extends to a concave, nondecreasing function on $[0, \infty)$ (for instance, by piecewise linear interpolation of its values on \mathbb{N}). Then there exists an online algorithm A^{multi} such that for every horizon $T \geq 0$ and every sequence $(x^t, y^t)_{t=1}^T$, the following holds.

Partition the rounds by membership patterns: for $x \in X$, let $z(x) \in \{0, 1\}^k$ with $z_j(x) = g_j(x)$. For each realized pattern z , define the cell $C_z := \{t : z(x^t) = z\}$ and its size $T_z := |C_z|$. For each $j \in \{1, \dots, k\}$ set

$$T_j := \sum_{z:z_j=1} T_z, \quad K_j := |\{z : z_j = 1, T_z > 0\}|.$$

Then, for every j , we have

$$\mathbb{E}[\text{Err}_T(g_j)] \leq \sum_{z:z_j=1} R(T_z) \leq K_j R(T_j/K_j) \quad \text{if } R \text{ is concave}, \quad (30)$$

with the convention that $K_j = 0$ implies both sides are 0 (since $T_j = 0$ and $R(0) = 0$). In addition, the multicalibration error satisfies the following valid bounds:

$$\mathbb{E}[\text{MCerr}_T(G)] \leq \sum_{z \in \{0, 1\}^k} R(T_z) \leq 2^k R(T/2^k) \quad \text{if } R \text{ is concave}. \quad (31)$$

Proof. For each pattern $z \in \{0, 1\}^k$, create an independent copy A^z of A^{marg} . On each round t , compute $z_t := z(x^t)$, query A^{z_t} for a prediction p^t , then reveal y^t and feed (x^t, y^t, p^t) only to the copy A^{z_t} . This defines the algorithm A^{multi} .

For each z and $v \in [0, 1]$, define the cell-wise bias

$$B_z(v) := \sum_{t \in C_z} \mathbf{1}[p^t = v] (p^t - y^t).$$

Write $V(C_z)$ for the set of prediction values output by A^z on its own rounds C_z . By construction, on the rounds C_z the copy A^z runs exactly as A^{marg} would on the length- T_z sequence $((x^t, y^t))_{t \in C_z}$. Therefore, conditioning on the value $T_z = n$ and on the realized subsequence $((x^t, y^t))_{t \in C_z}$, the marginal guarantee gives

$$\mathbb{E} \left[\sum_{v \in V(C_z)} |B_z(v)| \mid C_z, T_z = n \right] \leq R(n).$$

Taking expectations and using the law of total expectation yields

$$\mathbb{E} \left[\sum_{v \in V(C_z)} |B_z(v)| \right] \leq R(T_z). \quad (32)$$

Fix $j \in \{1, \dots, k\}$. Since $g_j(x) = 1$ if and only if $z_j(x) = 1$, we have for each v ,

$$B_T(v, g_j) = \sum_{t=1}^T \mathbf{1}[p^t = v] g_j(x^t) (p^t - y^t) = \sum_{z:z_j=1} \sum_{t \in C_z} \mathbf{1}[p^t = v] (p^t - y^t) = \sum_{z:z_j=1} B_z(v).$$

By the triangle inequality and summing over v , we obtain

$$\text{Err}_T(g_j) = \sum_{v \in V_T} |B_T(v, g_j)| \leq \sum_{z:z_j=1} \sum_{v \in V(C_z)} |B_z(v)|.$$

Taking expectations and applying (32),

$$\mathbb{E}[\text{Err}_T(g_j)] \leq \sum_{z:z_j=1} R(T_z),$$

which proves the first inequality in (30).

For the second inequality in (30), fix a realization of $(T_z)_{z:z_j=1}$ and let $K = K_j$ and $a_i := T_{z_i}$ for the K patterns with $z_j = 1$. Since R is concave and nondecreasing on $[0, \infty)$ with $R(0) = 0$, Jensen's inequality gives

$$\frac{1}{K} \sum_{i=1}^K R(a_i) \leq R\left(\frac{1}{K} \sum_{i=1}^K a_i\right) = R(T_j/K).$$

Multiplying by K yields the pathwise bound $\sum_{z:z_j=1} R(T_z) \leq K_j R(T_j/K_j)$. Taking expectations establishes the second inequality in (30).

To bound the multicalibration error, define $A_z := \sum_{v \in V(C_z)} |B_z(v)| \geq 0$ for each pattern z . For each j we showed pathwise that $\text{Err}_T(g_j) \leq \sum_{z:z_j=1} A_z$, hence

$$\text{MCerr}_T(G) = \max_{1 \leq j \leq k} \text{Err}_T(g_j) \leq \max_{1 \leq j \leq k} \sum_{z:z_j=1} A_z \leq \sum_{z \in \{0,1\}^k} A_z,$$

where the last inequality holds because each $A_z \geq 0$ and the maximum of partial sums is at most the total sum. Taking expectations and using (32) yields

$$\mathbb{E}[\text{MCerr}_T(G)] \leq \sum_{z \in \{0,1\}^k} \mathbb{E}[A_z] \leq \sum_{z \in \{0,1\}^k} R(T_z).$$

Since R is concave and nondecreasing with $R(0) = 0$ and $\sum_z T_z = T$, Jensen’s inequality gives the pathwise bound

$$\sum_{z \in \{0,1\}^k} R(T_z) \leq 2^k R\left(\frac{1}{2^k} \sum_{z \in \{0,1\}^k} T_z\right) = 2^k R(T/2^k),$$

establishing (31). \square

The theorem shows that the difficulty of prediction-independent multicalibration is governed by the intersection structure of the groups, quantified by K_j and T_j . In particular, when $|G|$ is fixed independently of T , the crude bound $K_j \leq 2^{k-1}$ recovers the same rate as the marginal algorithm.

This simple reduction crucially relies on the fact that the groups are *binary* and *prediction independent*, since it needs to identify which subset of groups are active *before* it decides which algorithm will be assigned to make a prediction each day. It establishes that marginal calibration rates are the same as multicalibration rates (as a function of T) for any collection of groups whose cardinality is independent of T — and hence, by the result of Dagan et al. [2025], $o(T^{2/3})$. Contrast this with Theorem 1 which establishes an $\Omega(T^{2/3})$ lower bound for multicalibration of even 3 groups when the groups can depend on the prediction.

B An Oracle Lower Bound for Better Black-Box Reductions

This section formalizes a natural class of *proper* black-box reductions from multicalibration to marginal calibration, and proves an oracle lower bound showing that such reductions require *exponentially many* oracle copies (in the number of groups) in the worst case, even for prediction independent groups — showing that our reduction in Appendix A is in a sense tight. Like the lower bounds in Sections 3 and 4, the environment here is completely oblivious, but our instance here is even more benign: the labels are now deterministic given the context; the lower bound follows purely from the properness constraint and the “context-blindness” of marginal calibration algorithms. We show that any “proper” black-box reduction from marginal calibration to multicalibration must itself incur multicalibration error $\Omega(T^{1-\gamma})$ for any constant $\gamma > 0$ unless it uses exponentially many oracle copies. Our construction uses a binary group family of cardinality $|G| = \Theta(\log T)$, and so serves as an obstruction to a reductions-based strategy to giving $o(T^{2/3})$ upper bounds for multicalibration in this regime.

B.1 Context-blind oracles and proper reductions

We define the notion of a “context-blind” oracle:

Definition 9 (Context-blind oracle). A (possibly randomized) forecasting algorithm A is *context-blind* if for every round t its output distribution depends only on its internal state (i.e. on its own past transcript) and not on the current context x^t . Equivalently, for any realized internal randomness of A , the mapping $x^t \mapsto Q^t$ produced by A on round t is constant.

Remark 1 (Context-Blinding). If A is any algorithm with a worst-case marginal calibration guarantee that holds for *all* context sequences, then feeding A a fixed dummy context \bar{x} on every round (and otherwise running it unchanged) produces a context-blind algorithm with the same marginal guarantee. Thus, any black-box reduction that claims to work for *every* marginal calibration algorithm must in particular work for some context-blind marginal oracle. In what follows we fix such a context-blind oracle A and treat it purely as a black box satisfying a marginal calibration guarantee.

We define the family of reductions that our barrier result applies to below. Informally, it corresponds to algorithms that can run m copies of some marginal calibration algorithm A , update each algorithm in arbitrary ways, potentially differently for each copy of A , and then use prediction distributions that are somewhere in the convex hull of the predictive distributions proposed by each copy of A , where the weights of the convex mixture can depend both on context and history. This is e.g. the form that reductions from multigroup regret to marginal (external) regret via sleeping experts constructions take [Blum and Lykouris, 2020, Acharya et al., 2024]. These reductions run one copy of the oracle for each group — i.e. setting $m = \Theta(|G|)$ and update each oracle for marginal regret on the subsequence corresponding to the rounds at which the corresponding group is active. Our barrier will rule out any similar reduction obtaining sublinear multicalibration error.

Definition 10 (Proper m -copy black-box reduction). Fix an integer $m \geq 1$. A *proper m -copy black-box reduction* B is a meta-algorithm with oracle access to a context-blind forecasting algorithm A , and it is allowed to run m independent copies $A^{(1)}, \dots, A^{(m)}$ of A .

On each round t :

1. The context x^t is revealed to B .
2. Each copy $A^{(i)}$ outputs a distribution Q_i^t on $[0, 1]$. Because A is context-blind, each Q_i^t is a (possibly randomized) function only of the past transcript of copy i , and does not depend on x^t .
3. The reduction outputs a distribution P^t satisfying the *properness* constraint

$$P^t \in \text{conv}\{Q_1^t, \dots, Q_m^t\}, \quad \text{i.e.} \quad P^t = \sum_{i=1}^m \alpha_{i,t} Q_i^t \quad \text{for some } \alpha_t \in \Delta_m,$$

where the weights α_t may depend on x^t and all past history.

4. The outcome y^t is revealed.
5. The reduction may choose, for each copy i , whether and how to update the state of copy i using information available up to this point. (Our lower bound will not depend on any particular update scheme; it uses only the context-blindness of the Q_i^t at the moment they are produced and the properness constraint on P^t .)
6. Finally, the prediction $p^t \sim P^t$ is drawn.

B.2 Hard instance and a logarithmic-size group family

Fix integers $T \geq 1$ and $k \geq 1$, and set $N := 2^k$. Let the context space be the k -bit hypercube

$$X := \{0, 1\}^k,$$

and identify each $x = (x_1, \dots, x_k) \in X$ with the integer

$$\text{val}(x) := \sum_{r=1}^k x_r 2^{k-r} \in \{0, 1, \dots, N-1\}.$$

Define the partition of $[0, 1]$ into N intervals

$$J_b := \begin{cases} \left[\frac{b}{N}, \frac{b+1}{N}\right), & b = 0, 1, \dots, N-2, \\ \left[\frac{N-1}{N}, 1\right], & b = N-1. \end{cases}$$

Define the mean map $\mu : X \rightarrow (0, 1)$ by

$$\mu(x) := \frac{\text{val}(x) + \frac{1}{2}}{N}.$$

Note that $\mu(x)$ is the midpoint of $J_{\text{val}(x)}$.

Distribution over contexts and labels. Let $\mathcal{D}_{T,N}$ be the oblivious distribution over $(x^t, y^t)_{t=1}^T$ defined by contexts x^1, \dots, x^T that are i.i.d. uniform on X and labels that are deterministically $y^t := \mu(x^t)$ for each t . Thus (x^t, y^t) are independent of the forecaster, and $y^t \in (0, 1)$ always.

Group family. We use $k+1 = \log_2 N + 1$ binary prediction-independent groups:

$$g_0(x) := 1, \quad g_r(x) := x_r \in \{0, 1\} \quad (r = 1, \dots, k).$$

Let

$$G_{\text{bits}} := \{g_0, g_1, \dots, g_k\}, \quad |G_{\text{bits}}| = k+1.$$

B.3 Main theorem

Theorem 5 (Oracle lower bound with $|G| = \Theta(\log N)$). *Fix integers $T \geq 1$, $k \geq 1$, and set $N := 2^k$. Let B be any proper m -copy black-box reduction (Definition 10), and let A be any context-blind oracle (Definition 9). Run the induced forecaster B^A for T rounds on $\mathcal{D}_{T,N}$ and evaluate multicalibration error with respect to G_{bits} . Then*

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{1}{8} \left(1 - \frac{m}{N}\right) \frac{T}{N^2}.$$

In particular, if $m \leq N/2$ then

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{T}{16N^2}.$$

Equivalently, fix any constant $\gamma \in (0, \frac{1}{2})$ and let

$$k := \lfloor \gamma \log_2 T \rfloor, \quad N := 2^k,$$

so that $|G_{\text{bits}}| = k + 1 = \Theta(\log T)$. Then there exists $T_0 = T_0(\gamma)$ such that for all $T \geq T_0$ and all $m = \text{poly}(|G_{\text{bits}}|)$ we have

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{1}{16} T^{1-2\gamma}.$$

The proof is a combination of three ingredients: (i) a properness lemma implying the reduction cannot put much probability mass on the correct interval $J_{\text{val}(x^t)}$ on average; (ii) the fact that missing the correct interval implies a squared-loss penalty of order $1/N^2$; and (iii) a deterministic inequality showing that squared loss is controlled by calibration error for the bit groups.

Remark 2. The parameter γ in Theorem 5 can be chosen arbitrarily small, so the exponent $1 - 2\gamma$ can be made arbitrarily close to 1 while still working with only $|G_{\text{bits}}| = \Theta(\log T)$ binary prediction-independent groups and using only $m = \text{poly}(|G_{\text{bits}}|)$ oracle copies.

B.4 Proof of Theorem 5

Lemma 20 (Correct-interval mass bound). *Let B^A be any proper m -copy reduction with a context-blind oracle A , run on $\mathcal{D}_{T,N}$. For each round t , let P^t be the external distribution output by B , and define*

$$\pi_t := P^t(J_{\text{val}(x^t)}) \in [0, 1].$$

Then for every t ,

$$\mathbb{E}[\pi_t] \leq \frac{m}{N}.$$

Consequently,

$$\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}[p^t \in J_{\text{val}(x^t)}]\right] = \mathbb{E}\left[\sum_{t=1}^T \pi_t\right] \leq \frac{m}{N} T.$$

Proof. Fix a round t and condition on the full transcript up to the end of round $t - 1$, including all internal randomness of B and of all oracle copies. Under $\mathcal{D}_{T,N}$, the fresh context x^t is uniform on X and independent of this past transcript.

Because the oracle A is context-blind, each copy's distribution Q_i^t depends only on the past transcript of copy i . Hence, under the conditioning, the distributions Q_1^t, \dots, Q_m^t are fixed probability measures on $[0, 1]$ and do not depend on the random draw of x^t .

Since B is proper, it outputs

$$P^t = \sum_{i=1}^m \alpha_{i,t} Q_i^t \quad \text{for some } \alpha_t \in \Delta_m$$

(possibly chosen as a function of x^t). For any $b \in \{0, \dots, N - 1\}$ we have

$$P^t(J_b) = \sum_{i=1}^m \alpha_{i,t} Q_i^t(J_b) \leq \max_{1 \leq i \leq m} Q_i^t(J_b).$$

Averaging over the uniform x^t (equivalently $b = \text{val}(x^t)$, which is uniform on $\{0, \dots, N-1\}$) and using $\max_i a_i \leq \sum_i a_i$ for nonnegative a_i , we obtain

$$\begin{aligned}\mathbb{E}[P^t(J_{\text{val}(x^t)}) \mid \text{past transcript up to } t-1] &\leq \frac{1}{N} \sum_{b=0}^{N-1} \max_{1 \leq i \leq m} Q_i^t(J_b) \\ &\leq \frac{1}{N} \sum_{b=0}^{N-1} \sum_{i=1}^m Q_i^t(J_b) = \frac{1}{N} \sum_{i=1}^m \sum_{b=0}^{N-1} Q_i^t(J_b) \\ &= \frac{1}{N} \sum_{i=1}^m 1 = \frac{m}{N},\end{aligned}$$

since the intervals J_0, \dots, J_{N-1} form a partition of $[0, 1]$, so $\sum_b Q_i^t(J_b) = 1$ for each i . Taking expectations over the past transcript yields $\mathbb{E}[\pi_t] \leq m/N$.

For the second step, note that conditional on (P^t, x^t) the realized draw satisfies

$$\mathbb{E}[\mathbf{1}[p^t \in J_{\text{val}(x^t)}] \mid P^t, x^t] = P^t(J_{\text{val}(x^t)}) = \pi_t,$$

because $p^t \sim P^t$. Taking expectations and summing over t gives the claim. \square

Lemma 21 (Misses force squared loss). *Under $\mathcal{D}_{T,N}$, for every realization we have*

$$\sum_{t=1}^T (p^t - y^t)^2 \geq \frac{1}{4N^2} \sum_{t=1}^T \mathbf{1}[p^t \notin J_{\text{val}(x^t)}].$$

Consequently,

$$\mathbb{E}\left[\sum_{t=1}^T (p^t - y^t)^2\right] \geq \frac{1}{4N^2} \left(1 - \frac{m}{N}\right) T.$$

Proof. Fix a round t . Under $\mathcal{D}_{T,N}$, the label is $y^t = \mu(x^t)$, which is the midpoint of the interval $J_{\text{val}(x^t)}$. The distance from the midpoint of an interval of width $1/N$ to the complement of the interval is exactly $1/(2N)$, so

$$p^t \notin J_{\text{val}(x^t)} \implies |p^t - y^t| \geq \frac{1}{2N} \implies (p^t - y^t)^2 \geq \frac{1}{4N^2}.$$

Multiplying by the indicator $\mathbf{1}[p^t \notin J_{\text{val}(x^t)}]$ and summing over t gives the first claim.

For the second claim, take expectations and apply Lemma 20:

$$\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}[p^t \notin J_{\text{val}(x^t)}]\right] = T - \mathbb{E}\left[\sum_{t=1}^T \mathbf{1}[p^t \in J_{\text{val}(x^t)}]\right] \geq \left(1 - \frac{m}{N}\right) T.$$

\square

Thus far we have established that proper oracle reductions with $m \ll N$ must frequently mispredict the true label with non-negligible margin, and hence incur large squared loss. The next lemma establishes that strong multicalibration bounds force small squared loss.

Lemma 22 (Squared loss controlled by bit-group calibration). *For every realization under $\mathcal{D}_{T,N}$,*

$$\sum_{t=1}^T (p^t - y^t)^2 \leq \left(2 - \frac{1}{2N}\right) \text{MCerr}_T(G_{\text{bits}}) < 2 \text{MCerr}_T(G_{\text{bits}}).$$

Proof. For each realized prediction value $v \in V_T$, let $S_v := \{t \in \{1, \dots, T\} : p^t = v\}$ and note that

$$\sum_{t=1}^T (p^t - y^t)^2 = \sum_{v \in V_T} \sum_{t \in S_v} (v - y^t)^2.$$

Fix $v \in V_T$. For each $t \in S_v$ we have the identity

$$(v - y^t)^2 = (v - y^t)v - (v - y^t)y^t,$$

so summing over $t \in S_v$ gives

$$\sum_{t \in S_v} (v - y^t)^2 = v \sum_{t \in S_v} (v - y^t) - \sum_{t \in S_v} y^t (v - y^t).$$

Since $0 \leq v \leq 1$, the triangle inequality yields

$$\sum_{t \in S_v} (v - y^t)^2 \leq \left| \sum_{t \in S_v} (v - y^t) \right| + \left| \sum_{t \in S_v} y^t (v - y^t) \right|. \quad (33)$$

Next we expand $y^t = \mu(x^t)$ in terms of the bit groups. Because $N = 2^k$ and $\text{val}(x^t) = \sum_{r=1}^k x_r^t 2^{k-r}$, we have

$$y^t = \mu(x^t) = \frac{\text{val}(x^t) + \frac{1}{2}}{N} = \frac{1}{2N} + \frac{\text{val}(x^t)}{2^k} = \frac{1}{2N} + \sum_{r=1}^k 2^{-r} x_r^t.$$

Therefore

$$\sum_{t \in S_v} y^t (v - y^t) = \frac{1}{2N} \sum_{t \in S_v} (v - y^t) + \sum_{r=1}^k 2^{-r} \sum_{t \in S_v} x_r^t (v - y^t),$$

and applying the triangle inequality gives

$$\left| \sum_{t \in S_v} y^t (v - y^t) \right| \leq \frac{1}{2N} \left| \sum_{t \in S_v} (v - y^t) \right| + \sum_{r=1}^k 2^{-r} \left| \sum_{t \in S_v} x_r^t (v - y^t) \right|. \quad (34)$$

Now relate these terms to calibration error. By definition, for a group g and a value $v \in V_T$,

$$B_T(v, g) := \sum_{t=1}^T \mathbf{1}[p^t = v] g(x^t) (p^t - y^t) = \sum_{t \in S_v} g(x^t) (v - y^t),$$

and $\text{Err}_T(g) = \sum_{v \in V_T} |B_T(v, g)|$. Thus

$$\sum_{t \in S_v} (v - y^t) = B_T(v, g_0), \quad \sum_{t \in S_v} x_r^t (v - y^t) = B_T(v, g_r).$$

Plugging these into (33)–(34) yields

$$\sum_{t \in S_v} (v - y^t)^2 \leq \left(1 + \frac{1}{2N}\right) |B_T(v, g_0)| + \sum_{r=1}^k 2^{-r} |B_T(v, g_r)|.$$

Summing over $v \in V_T$ gives

$$\sum_{t=1}^T (p^t - y^t)^2 \leq \left(1 + \frac{1}{2N}\right) \text{Err}_T(g_0) + \sum_{r=1}^k 2^{-r} \text{Err}_T(g_r).$$

Finally, since $\sum_{r=1}^k 2^{-r} = 1 - 2^{-k} = 1 - \frac{1}{N}$ and each $\text{Err}_T(g_r) \leq \text{MCerr}_T(G_{\text{bits}})$, we obtain

$$\sum_{t=1}^T (p^t - y^t)^2 \leq \left(1 + \frac{1}{2N} + 1 - \frac{1}{N}\right) \text{MCerr}_T(G_{\text{bits}}) = \left(2 - \frac{1}{2N}\right) \text{MCerr}_T(G_{\text{bits}}),$$

as claimed. \square

It remains to put the pieces together:

Proof of Theorem 5. Combine Lemma 21 and Lemma 22. Lemma 22 implies pathwise

$$\text{MCerr}_T(G_{\text{bits}}) \geq \frac{1}{2} \sum_{t=1}^T (p^t - y^t)^2,$$

since $2 - \frac{1}{2N} < 2$. Taking expectations and applying Lemma 21 gives

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{1}{2} \mathbb{E}\left[\sum_{t=1}^T (p^t - y^t)^2\right] \geq \frac{1}{2} \cdot \frac{1}{4N^2} \left(1 - \frac{m}{N}\right) T = \frac{1}{8} \left(1 - \frac{m}{N}\right) \frac{T}{N^2}.$$

If $m \leq N/2$ then $(1 - m/N) \geq 1/2$, yielding

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{T}{16N^2}.$$

For the parametric statement in the theorem, fix any constant $\gamma \in (0, \frac{1}{2})$ and set

$$k := \lfloor \gamma \log_2 T \rfloor, \quad N := 2^k.$$

Then $|G_{\text{bits}}| = k + 1 = \Theta(\log T)$, and

$$N \leq 2^{\gamma \log_2 T} = T^\gamma$$

implies

$$\frac{T}{N^2} \geq \frac{T}{T^{2\gamma}} = T^{1-2\gamma}.$$

If in addition $m = \text{poly}(|G_{\text{bits}}|)$, then since $N = 2^k = T^{\Theta(\gamma)}$ grows superpolynomially in $|G_{\text{bits}}|$, there exists $T_0 = T_0(\gamma)$ such that $m \leq N/2$ for all $T \geq T_0$. For such T we obtain

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{1}{16} T^{1-2\gamma},$$

as claimed in Theorem 5. \square

Interpretation Since $|G_{\text{bits}}| = k + 1$ and $N = 2^k$, the bound in Theorem 5 can be rewritten as

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{1}{8} \left(1 - \frac{m}{2^{|G_{\text{bits}}|-1}}\right) \frac{T}{2^{2(|G_{\text{bits}}|-1)}} = \Omega\left(\left(1 - \frac{m}{2^{|G|-1}}\right) \frac{T}{2^{2(|G|-1)}}\right),$$

where $|G| := |G_{\text{bits}}|$. In particular, making this lower bound vacuous requires $m \approx N = 2^{\Theta(|G_{\text{bits}}|)}$, so any proper black-box reduction using $m = \text{poly}(|G|)$ copies (e.g. one copy per group) fails on this instance for large T .

A complementary view is obtained by optimizing N as a function of m . Fix $T \geq 1$ and $m \geq 1$, and choose

$$k := \lceil \log_2(2m) \rceil, \quad N := 2^k.$$

Then $2m \leq N < 4m$, and for any proper m -copy reduction B with context-blind oracle A run on $\mathcal{D}_{T,N}$, Theorem 5 gives

$$\mathbb{E}[\text{MCerr}_T(G_{\text{bits}})] \geq \frac{1}{8} \left(1 - \frac{m}{N}\right) \frac{T}{N^2} \geq \frac{1}{8} \cdot \frac{1}{2} \cdot \frac{T}{(4m)^2} = \frac{T}{256m^2} = \Omega\left(\frac{T}{m^2}\right).$$

Here $|G_{\text{bits}}| = k + 1 = \Theta(\log m)$, so any proper reduction with m context-blind oracle copies can be forced to incur multicalibration error $\Omega(T/m^2)$ on only $O(\log m)$ groups. This rules out any non-trivial reduction from multicalibration to marginal calibration in the “sleeping experts style” (as in Blum and Lykouris [2020], Acharya et al. [2024]) which use one oracle per group: $m = \Theta(|G|)$, and forces non-trivial reductions to choose m growing polynomially with T .