Sample Complexity of Multicalibration for Multilevel Properties

Jiuyao Lu*[1], Krishnakumar Balasubramanian[2,3], Aleksandr Podkopaev[2], and Shiva Prasad Kasiviswanathan[2]

[1]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania
[2]Amazon
[3]Department of Statistics, University of California, Davis

Abstract

Calibration requires a predictor to be unbiased after conditioning on its own predictions. Multicalibration asks for this guarantee simultaneously across a collection of groups. Many prediction tasks ask for several related features of the same conditional outcome distribution: variance is defined relative to the mean, skewness relative to both mean and variance, and conditional value at risk relative to a quantile. We study multicalibration for a sequence of k properties in which each property is identifiable once the preceding properties are fixed. This framework includes Bayes pairs but does not require the properties to arise from a single loss.

For every fixed $k \geq 2$, we establish matching upper and lower sample-complexity bounds up to logarithmic factors under regularity conditions. Even with only polylogarithmically many binary groups, achieving multicalibration error ε requires $\widetilde{\Omega}(\varepsilon^{-(k+2)})$ samples. Conversely, for any finite group family $\mathcal{G}$, we give a randomized learner using $O(\varepsilon^{-(k+2)} + \varepsilon^{-2} \log |\mathcal{G}|)$ samples. Thus the sample complexity is $\widetilde{\Theta}(\varepsilon^{-(k+2)})$ for polynomial-size group families. We instantiate the theory for three canonical examples.

Contents

*This work was done while interning at Amazon.

1 Introduction

Calibration asks whether predictions have their intended statistical meaning. For a predictor of the conditional mean, the basic requirement is that among the individuals who receive prediction p, the average outcome is also p [11]. This guarantee can still hide systematic errors: a predictor may underestimate outcomes on one structured subpopulation and compensate by overestimating them elsewhere. Given a family $\mathcal{G}$ of groups, multicalibration requires the predictor to remain calibrated after restricting attention to any one of them [29]. Multicalibration is useful when the same predictions inform many downstream decisions. It supports predictors that perform well for many losses, as in omniprediction [21, 23, 39], and it has found applications in complexity theory and information aggregation [3, 6, 7, 14].

Prior work has developed multicalibration for conditional means, moments, and quantiles [12, 27, 32, 33, 37]. In many applications, however, several related features of the conditional outcome law must be predicted together, and some are defined relative to others. Quantiles describe coverage and tail events, dispersion measures describe uncertainty, and measures of tail risk summarize rare outcomes with potentially large costs. For example, a three-level example arises in hospital capacity planning, where a model may need to predict a patient's expected length of stay, the variability around that expectation, and the degree of right-tail asymmetry caused by occasional very long stays. The model can output a prediction vector $p = (m, \sigma^2, \gamma)$, representing the conditional mean,

variance, and skewness of length of stay. These quantities have a natural sequential structure. At the first level, the mean prediction m separates patient groups whose stays are centered at different values. At the second level, after the mean has been fixed, the variance prediction σ^2 describes variability around that mean, so that differences in expected length of stay are not mistaken for uncertainty. At the third level, after the mean and variance have been fixed, the skewness prediction γ describes asymmetry relative to that center and scale. The third coordinate is important because two patient groups may have the same expected stay and the same overall variability but very different tail behavior: one may have roughly symmetric fluctuations around its mean, while the other may usually be discharged near the expected date but occasionally require a much longer hospitalization. Calibrating skewness without also fixing the mean and variance could pool groups with different centers or scales, causing those lower-order differences to appear spuriously as tail asymmetry. Three-level multicalibration therefore ensures that each prediction describes a distinct feature of the conditional outcome distribution: its location, its dispersion around that location, and the shape of its tail after location and dispersion have been accounted for.

Formally, let $\mathcal{M}$ be a class of probability laws on the outcome space $\mathcal{Y}$, and let $\Gamma_j : \mathcal{M} \to \mathbb{R}$ denote the j-th distributional property of interest. For each context x, write $\mu_x = \mathcal{L}(Y \mid X = x)$ for the conditional law of the outcome. Calibrating the properties $\Gamma_1, \Gamma_2, \Gamma_3$ separately need not yield a jointly coherent predictor. Indeed, consider the oracle scalar predictor $f_j(x) = \Gamma_j(\mu_x)$. Conditioning on the event $f_j(X) = a$ pools the conditional laws μ_x into the mixture

$$\bar{\mu}_a = \mathcal{L}(Y \mid f_j(X) = a) = \int \mu_x \, d\mathbb{P}(x \mid f_j(X) = a).$$

Although every component of this mixture satisfies $\Gamma_j(\mu_x) = a$, calibration additionally requires $\Gamma_j(\bar{\mu}_a) = a$. This implication holds only when the level set $\Gamma_j^{-1}(a) = \{\mu \in \mathcal{M} : \Gamma_j(\mu) = a\}$ is closed under mixtures. Thus, if Γ_j lacks convex level sets, even an oracle predictor can fail to be calibrated after conditioning on its own output [37]. In a multilevel setting, the failure may arise because distributions that agree on a higher-level property Γ_j, but differ in lower-level properties $\Gamma_1, \ldots, \Gamma_{j-1}$, are pooled together, and these lower-level differences change the value of Γ_j under mixing.

Variance provides a concrete example of this failure. Suppose

$$\mathbb{P}[(X, Y) = (1, 1)] = \mathbb{P}[(X, Y) = (2, 2)] = \frac{1}{2},$$

and consider the groups

$$g_1(x) = \mathbf{1}\{x = 1\}, \qquad g_2(x) = \mathbf{1}\{x = 2\}, \qquad g_{\text{all}}(x) \equiv 1.$$

The oracle conditional-variance predictor outputs zero at both contexts because the outcome is deterministic conditional on X. It is calibrated within g_1 and g_2, but not within g_{all}. Its single prediction cell pools the two outcomes, whose variance is $1/4$. In fact, g_1 and g_2 force any variance-only predictor to output zero at both contexts, whereas calibration within g_{all} would require their common prediction to be $1/4$. Thus no scalar variance predictor can be multicalibrated with respect to all three groups. Predicting the mean together with the variance separates the two contexts through their mean coordinates and removes this conflict.

This motivates predicting the ordered vector $f(x) = \big(f_1(x), f_2(x), f_3(x)\big)$ and calibrating its coordinates jointly. For simplicity, suppose that there are residual functions R_1, R_2, R_3 such that R_1 identifies Γ_1, R_2 identifies Γ_2 once the value of Γ_1 is fixed, and R_3 identifies Γ_3 once the values of both Γ_1 and Γ_2 are fixed. That is,

$$\mathbb{E}_\mu[R_1(p_1, Y)] = 0 \quad \Longleftrightarrow \quad p_1 = \Gamma_1(\mu),$$

and, conditional on $p_1 = \Gamma_1(\mu)$,

$$\mathbb{E}_\mu[R_2(p_1, p_2, Y)] = 0 \quad \Longleftrightarrow \quad p_2 = \Gamma_2(\mu),$$

while, conditional on $p_1 = \Gamma_1(\mu)$ and $p_2 = \Gamma_2(\mu)$,

$$\mathbb{E}_\mu[R_3(p_1, p_2, p_3, Y)] = 0 \quad \Longleftrightarrow \quad p_3 = \Gamma_3(\mu).$$

Three-level multicalibration then requires these residuals to have mean zero on every relevant subgroup and within each prediction cell indexed by $f = (f_1, f_2, f_3)$. The earlier coordinates therefore prevent contexts with different predictions at earlier levels from being pooled when calibrating a later coordinate, allowing a property that is not calibratable in isolation to become calibratable relative to the preceding predictions.

This motivates us to study the sample complexity of multilevel multicalibration under an ECE-type metric defined from these residuals, where ECE denotes expected calibration error [9, 36]. Let $(\Gamma_1, \ldots, \Gamma_k)$ be an ordered collection of distributional properties such that, for each level $j \in [k]$, the property Γ_j is identifiable conditional on the preceding properties $(\Gamma_1, \ldots, \Gamma_{j-1})$. Given n independent samples and a target accuracy ε, we ask when a learner can produce a predictor $f = (f_1, \ldots, f_k)$ with multicalibration error at most ε over every relevant subgroup. Our goal is to characterize how the required sample size n scales with ε and the number of groups, for each fixed k. In particular, we ask whether exploiting the conditional structure permits sample complexity comparable to scalar or vector-valued mean multicalibration, or whether the dependence among levels introduces an unavoidable additional statistical cost.

1.1 Main Results

We prove upper and lower bounds on the sample complexity of multicalibration that match up to logarithmic factors for sequentially conditionally identifiable multilevel properties. For finitely supported predictors, our multicalibration error is the maximum over groups of the ℓ_1 expected calibration error computed by bucketing according to the full prediction vector. (See Section 2 for the definitions of sequential conditional identifiability and multicalibration error for general predictors.) The lower and upper bounds rely on different regularity assumptions, all of which are satisfied by our canonical examples.

Lower Bound. For every sufficiently small ε, we construct a finite context space, polylogarithmically many binary groups, and a collection of data distributions such that any learner achieving multicalibration error at most ε with probability at least $2/3$ must have sample size

$$n = \Omega\left(\frac{\varepsilon^{-(k+2)}}{\log^{k+2}(1/\varepsilon)}\right) = \widetilde{\Omega}\left(\varepsilon^{-(k+2)}\right).$$

The result holds against randomized learners. Since the constructed group family has polylogarithmic size, it lies within every group-size budget of the form $\varepsilon^{-\kappa}$ with fixed $\kappa > 0$, once ε is sufficiently small.

Upper Bound. We give an online forecaster and convert its sequence of prediction rules into a randomized predictor. For a finite family $\mathcal{G}$ of groups, the resulting learner uses

$$n = O\left(\varepsilon^{-(k+2)} + \varepsilon^{-2} \log |\mathcal{G}|\right).$$

Thus, whenever $|\mathcal{G}|$ is polynomial in $1/\varepsilon$, the upper and lower bounds have the same exponent $k+2$. We also give conditions under which the learner can be implemented in polynomial time.

Canonical Properties. We verify the conditions of both bounds in three cases: mean and mean absolute deviation; mean, variance, and skewness; and quantile and conditional value at risk. The two-level examples have sample complexity $\widetilde{\Theta}(\varepsilon^{-4})$, while the three-level example has sample complexity $\widetilde{\Theta}(\varepsilon^{-5})$.

The results of Collina et al. [9] are most closely related to ours. They establish the exponent 3 for scalar mean multicalibration and $d+2$ for d-dimensional vector means, and also treat regular scalar elicitable properties. When $\Gamma(\mu) = \mathbb{E}_\mu[\psi(Y)]$ for a fixed vector-valued function ψ, vector mean multicalibration applies directly by treating $\psi(Y)$ as the outcome. More generally, a multilevel property need not admit such a representation, and a residual at a later level may depend on the earlier predictions. Our bounds allow this structure.

1.2 Proof Overview

Lower Bound. The lower bound proceeds by considering the task of determining which distribution generated the samples. The proof has four steps. First, we construct a large finite collection of candidate distributions. The true distribution is an unknown member of this collection, and the samples are drawn from it. Second, we show that a predictor with small multicalibration error must have small prediction error. Third, we use such a predictor to determine which member generated the samples. Finally, we show that determining the true distribution requires many samples.

Fix a resolution parameter Q. We take the context space to be the discrete grid $\{0, \dots, Q-1\}^k$, which contains Q^k contexts. We associate each context with an unperturbed vector in the k-dimensional property space; these vectors form a Cartesian grid. We assign one hidden sign in $\{-1, +1\}$ to each context. According to this sign assignment, we shift the final coordinate of the corresponding property vector by an amount of order $1/Q$, while leaving the first $k-1$ coordinates unchanged. We select $\exp(\Omega(Q^k))$ sign assignments such that any two of them disagree on a constant fraction of the contexts. Our regularity assumptions (Assumption 3.1(i)) ensure that every resulting vector is the property value of some outcome distribution. Taking the context to be uniform and using the corresponding outcome distribution conditionally on each context produces one candidate data distribution for every sign assignment. The resulting statistical task is to identify which candidate distribution generated the data, or equivalently, its sign assignment.

The second step converts multicalibration error into prediction error. Consider a given level, and suppose first that the predictions at all preceding levels are correct. Under our regularity assumptions (Assumption 3.1(ii)), the expected residual at the current level has the same sign as the difference between the prediction at that level and its true value, and its magnitude is at least proportional to the absolute value of this difference. Incorrect predictions at preceding levels can change the current residual, but the change is controlled by the corresponding prediction errors (see Assumption 3.1(iv)). Thus, the signed residual controls the current prediction error up to the prediction errors at preceding levels. The required sign depends on both the prediction and the context, whereas a group can depend only on the context. For a fixed prediction, however, this sign is a threshold function of one context coordinate. We construct a common collection of only polylogarithmically many binary groups whose linear combinations approximate all the threshold functions needed in the proof (see Lemma 3.3). Multiplying the residual by such an approximation produces, up to the approximation error, a linear combination of group-weighted residuals. Each of these residual terms is controlled by the multicalibration error (see Lemmas 3.4 and 3.6). Since the sum of the absolute values of the coefficients is $O(\log Q)$, the signed residual, and hence the current prediction error, is controlled by $O(\log Q)$ times the multicalibration error, together with the prediction errors at preceding levels.

Because a residual at level j may depend on all preceding predictions, we apply this argument

successively. After controlling the errors in the first $j-1$ coordinates, the Lipschitz dependence of the level-j residual on preceding predictions allows us to control the error in coordinate j. This gives bounds for the first $k-1$ coordinates. For the final coordinate, we additionally separate predictions according to whether they lie inside or outside disjoint neighborhoods of the unperturbed property vectors. Threshold functions control the predictions outside these neighborhoods, while the group containing all contexts controls those inside. Altogether, multicalibration error at most ε forces the expected ℓ_1 distance between the prediction and the true property vector to be $O(\varepsilon \log Q)$, where the expectation is over a uniformly random context and the predictor's randomization (see Proposition 3.7).

The third step uses this prediction guarantee to recover the sign assignment. Because any two candidate sign assignments disagree on a constant fraction of the contexts, their associated property vectors are separated by order $1/Q$ on average. Given a possibly randomized predictor, we take its mean output at every context and select the sign assignment whose associated property vectors are closest to these mean predictions. If the average prediction error is smaller than a sufficiently small constant multiple of $1/Q$, the true sign assignment is the unique closest one (see Lemmas 3.8 and 3.9). Consequently, any learner that produces a predictor with multicalibration error at most ε also provides a way to determine the true distribution whenever

$$\varepsilon \log Q \lesssim \frac{1}{Q}.$$

Finally, our regularity assumptions (Assumption 3.1(v)) ensure that the Kullback–Leibler divergence between nearby outcome distributions is at most a constant times the squared distance between their property vectors. Since the candidates differ by only $O(1/Q)$ at each context, the pairwise Kullback–Leibler divergence between their one-sample data distributions is $O(1/Q^2)$, and for n independent samples it is $O(n/Q^2)$ (see Lemma 3.10). On the other hand, the logarithm of the number of candidates is $\Omega(Q^k)$. Using Fano's inequality, we show in Proposition 3.11 that identifying the true candidate with constant success probability requires

$$\frac{n}{Q^2} \gtrsim Q^k, \qquad \text{and hence} \qquad n = \Omega(Q^{k+2}).$$

Choosing Q at the largest scale allowed by $\varepsilon \log Q \lesssim 1/Q$, namely

$$Q \asymp \frac{1}{\varepsilon \log(1/\varepsilon)},$$

gives

$$n = \Omega\left(\frac{\varepsilon^{-(k+2)}}{\log^{k+2}(1/\varepsilon)}\right).$$

The constructed family contains only polylogarithmically many groups, so it lies within every polynomial group-size budget for sufficiently small ε.

Upper Bound. Our upper-bound construction proceeds through an online forecasting problem. Fix a finite grid $\mathcal{P}_Q$ of k-dimensional prediction vectors, where Q controls the resolution and $|\mathcal{P}_Q| = O(Q^k)$. At round t, the forecaster uses the history from the preceding rounds to construct a randomized prediction rule

$$\pi_t : \mathcal{X} \to \Delta(\mathcal{P}_Q),$$

where $\Delta(\mathcal{P}_Q)$ denotes the set of distributions over the grid. The current context X_t is then revealed, and the forecaster predicts the distribution $\pi_t(\cdot \mid X_t)$. After observing the outcome Y_t, it

evaluates the residuals associated with this distribution. We first control the empirical analogue of multicalibration error accumulated over these online rounds.

Given a dataset consisting of T i.i.d. context–outcome pairs, we run this online forecasting procedure for T rounds, using the t-th pair on round t. The context X_t is revealed before the prediction and the outcome Y_t afterward. This run produces T prediction rules $\pi_1, \ldots, \pi_T$. We return their average:

$$\Pi(x) = \frac{1}{T} \sum_{t=1}^{T} \pi_t(\cdot \mid x).$$

An online-to-batch argument transfers the empirical guarantee for the online transcript to a population multicalibration guarantee for Π.

The empirical objective contains many calibration requirements. For every group g, point $p \in \mathcal{P}_Q$, and level j, it records the cumulative level-j residual associated with g, with the residual on round t weighted by the probability $\pi_t(p \mid X_t)$ assigned to p. For each group, the empirical error sums the absolute values of these cumulative residuals over all pairs (p, j). We can express this sum as a maximum over sign arrays $s = (s_{p,j})$, with one sign assigned to each pair (p, j). After also taking the maximum over groups, the empirical error becomes the maximum of a collection of signed residual objectives indexed by pairs (g, s). Controlling all of these objectives is therefore equivalent to controlling the empirical multicalibration error.

The forecaster assigns a weight to each objective at the beginning of every round. These weights depend on the signed residuals accumulated on previous rounds. If a calibration violation has grown large for some group and some pattern of signs across prediction cells and levels, the corresponding objective receives greater weight, so the next prediction places more emphasis on reducing that violation. Objectives whose cumulative violations remain small receive less relative emphasis. Once the weights have been fixed, every candidate distribution on $\mathcal{P}_Q$ has a weighted expected residual under each possible conditional outcome law. The forecaster chooses among these distributions in a minimax manner by minimizing the largest such weighted expected residual over all admissible outcome laws.

Formally, we use the framework of online learning with expert advice. We treat each objective as an expert and its signed residual on each round as the expert's gain. The exponential weights algorithm assigns weights to the experts according to their cumulative gains, producing the objective weights described above. Its regret guarantee bounds the largest cumulative gain of any expert by the sum of the expected gains under these weights, plus a regret term (Lemma 4.4). Because the empirical multicalibration error is exactly the largest expert gain divided by T, it remains to control these expected gains. Conditional on the past and the current context, the expected gain on each round is bounded by the value of the minimax problem solved on that round.

We bound this value by reversing the order of play using Sion's minimax theorem (Lemma 4.5). In the reversed game, an admissible outcome law μ is fixed before the prediction is selected. For high-dimensional mean prediction, it suffices to choose a grid point close to the revealed outcome vector [9, 38]. In our setting, the residual need not be the difference between a prediction and an outcome, and a residual at a later level may depend on the entire preceding prediction. Instead, we assume that, for every μ, there is a single grid point at which all expected residuals are simultaneously $O(1/Q)$ (Assumption 4.1(iii)). Choosing this point in the reversed game makes every weighted objective $O(1/Q)$. The minimax theorem then guarantees that, in the original order of play, there is a distribution on $\mathcal{P}_Q$ whose worst-case weighted expected residual is equally small, even though the forecaster does not know μ.

Combining this bound with the regret guarantee of exponential weights, Theorem 4.3 gives

$$O\left(\frac{1}{Q}+\sqrt{\frac{Q^k+\log|\mathcal{G}|}{T}}\right)$$

expected empirical multicalibration error, apart from the chosen accuracy for solving the minimax problem.

Finally, we convert the online guarantee into a batch guarantee. A martingale concentration argument shows that the population multicalibration error of the averaged predictor Π exceeds the empirical multicalibration error of the online transcript by at most a term of the same order (Lemma 4.7). Taking $Q=\Theta(1/\varepsilon)$ and

$$T=O\left(\varepsilon^{-(k+2)}+\varepsilon^{-2}\log|\mathcal{G}|\right)$$

gives the upper bound on sample complexity in Corollary 4.9.

Although the expert family is exponentially large, its weights can be computed without enumerating all experts. The minimax step reduces to linear optimization over $\mathcal{M}$, as specified in Assumption 4.10. For each of our canonical examples, we implement this oracle in polynomial time (Propositions 5.3, 5.7, and 5.12), yielding the polynomial-time learner in Theorem 4.11.

1.3 Related Work

Multicalibration and Sample Complexity. Hebert-Johnson et al. [29] introduced multicalibration. Subsequent work developed online algorithms and efficient implementations, and connected multicalibration to multiobjective optimization and omniprediction [17, 18, 28, 34]. Other formulations measure calibration differently. Gopalan et al. [22] consider a multicalibration condition that tests residuals after restricting attention to observations whose predictions fall in an interval. Globus-Harris et al. [20] study multicalibration under a weighted L_2 metric and give a boosting algorithm for controlling it. When $|\mathcal{G}|$ is polynomial in $1/\varepsilon$, the guarantees of Gopalan et al. [22] and Globus-Harris et al. [20] translate into sample complexities of $\widetilde{O}(\varepsilon^{-8})$ and $\widetilde{O}(\varepsilon^{-10})$ for mean ECE, respectively [9].

Gopalan et al. [24] introduce swap multicalibration, a stronger condition in which the group function used to test residual bias may vary across prediction values. Luo et al. [35] improve the online rates and sample-complexity upper bounds for swap multicalibration of conditional means when residual bias is tested using bounded linear functions of the context.

The weaker notion of calibrated multiaccuracy asks a predictor to be globally calibrated and multiaccurate with respect to a class of groups, meaning that its residuals are uncorrelated with every group function in the class [4]. Gibbs and Tibshirani [19] establish an $\Omega(\varepsilon^{-5/2})$ sample-complexity lower bound for learning deterministic predictors of conditional means under this requirement.

Collina et al. [9] study batch sample complexity for scalar and vector means, weighted error metrics, and regular scalar elicitable properties. Collina et al. [8] study the online setting, establish sharp lower bounds, and develop the subsampled Walsh family used in our group construction.

A related but distinct statistical question is whether empirical multicalibration error reliably estimates population multicalibration error throughout a chosen predictor class. Shabat et al. [44] establish bounds depending on the size or graph dimension of the class, while Rosenberg et al. [41] connect this question to standard ERM generalization bounds. This guarantee concerns estimation within the class, not the existence of a low-error predictor in it.

Calibration of Distributional Properties. Jung et al. [32] introduce mean-conditioned moment multicalibration, and Gupta et al. [27] give an online algorithm for this notion. Multivalid prediction intervals have also been studied in online and batch settings [2, 27, 33]. Noarov and Roth [37] connect property multicalibration to elicitation and identification. They characterize which continuous scalar properties are sensible for calibration under mild conditions and give an algorithm for jointly multicalibrating two-level conditionally elicitable properties, including Bayes pairs. Hu et al. [31] give an oracle-efficient online algorithm for swap multicalibration of elicitable properties. We study the sample complexity of joint calibration for an arbitrary fixed number of conditional levels, including sequences that are not generated by a single loss and its Bayes risk.

High-Dimensional Prediction. Recent work studies calibration and related unbiasedness guarantees for vector-valued predictions, including computational questions and online rates that depend on dimension [16, 25, 38, 40]. Noarov et al. [38] give an online algorithm for predicting high-dimensional states such that, on every event in a specified family, the accumulated prediction error is small in every coordinate. These events may depend on the past transcript, the current context, and the prediction. In our setting, each coordinate predicts one property in an ordered sequence rather than one component of a state vector, and the residual for a later property may depend on the preceding predictions.

Organization. Section 2 defines sequential conditional identifiability and the multicalibration error used throughout. Sections 3 and 4 prove the lower and upper bounds, respectively. Section 5 verifies the assumptions in our three canonical examples.

2 Basic Setup

2.1 Sequential Conditional Identifiability

Let $\mathcal{Y}$ be a standard Borel outcome space and let $\mathcal{M}$ be a class of probability measures on $\mathcal{Y}$. Fix an integer $k \geq 2$ and a compact rectangular prediction space

$$\mathcal{P} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_k \subset \mathbb{R}^k.$$

Throughout, k is treated as fixed. Constants suppressed by asymptotic notation may depend on k, as well as on the stated regularity parameters. A prediction value is denoted $p = (p_1, \ldots, p_k)$. For $j \in [k]$, write

$$p_{<j} := (p_1, \ldots, p_{j-1}), \qquad \mathcal{P}_{<j} := \mathcal{P}_1 \times \cdots \times \mathcal{P}_{j-1},$$

with $\mathcal{P}_{<1}$ interpreted as a one-point set.

Definition 2.1 (Sequentially Conditionally Identifiable Property)**.** *A k-level property is a map*

$$\Gamma = (\Gamma_1, \ldots, \Gamma_k) : \mathcal{M} \to \mathcal{P}.$$

For $\mu \in \mathcal{M}$, write

$$\Gamma_{<j}(\mu) := (\Gamma_1(\mu), \ldots, \Gamma_{j-1}(\mu)).$$

The property Γ is sequentially conditionally identifiable if, for every level $j \in [k]$, there is a jointly measurable residual function

$$R_j : \mathcal{P}_{<j} \times \mathcal{P}_j \times \mathcal{Y} \to \mathbb{R}$$

with the following properties. For every $\mu \in \mathcal{M}$ and every $(a, q) \in \mathcal{P}_{<j} \times \mathcal{P}_j$, the function $R_j(a, q, \cdot)$ is μ-integrable. Moreover, for every $\mu \in \mathcal{M}$ and every $q \in \mathcal{P}_j$,

$$\mathbb{E}_{Y\sim\mu} R_j(\Gamma_{<j}(\mu), q, Y) = 0 \quad \Longleftrightarrow \quad q = \Gamma_j(\mu).$$

Thus level j is identified after the previous true levels have been fixed. The residual functions are also evaluated away from the true prefix, because predictions need not have correct earlier coordinates. The residual vector used throughout the paper is

$$R(p, y) := \big(R_1(p_{<1}, p_1, y), \ldots, R_k(p_{<k}, p_k, y)\big).$$

When the final argument is a distribution, expectation over that distribution is denoted by:

$$R_j(a, q, \mu) := \mathbb{E}_{Y\sim\mu} R_j(a, q, Y), \qquad R(p, \mu) := \mathbb{E}_{Y\sim\mu} R(p, Y).$$

Sequential conditional identifiability implies

$$R(\Gamma(\mu), \mu) = 0 \qquad \forall \mu \in \mathcal{M}.$$

The familiar Bayes pair setting is the special case $k = 2$ [37]. If a scalar property γ has identification function r_γ, and if L is a strictly consistent loss for γ (so $\gamma(\mu)$ uniquely minimizes $\mathbb{E}_\mu L(r, Y)$), then its Bayes risk is

$$\mathcal{B}(\mu) := \mathbb{E}_{Y\sim\mu} L(\gamma(\mu), Y),$$

and the Bayes pair is $\Gamma(\mu) = (\gamma(\mu), \mathcal{B}(\mu))$. It is sequentially conditionally identifiable with residual functions

$$R_1(\varnothing, p_1, y) = r_\gamma(p_1, y), \qquad R_2(p_1, p_2, y) = p_2 - L(p_1, y).$$

For $\tau \in (0, 1)$, the quantile and conditional-value-at-risk example in Section 5.3 uses the loss

$$L_\tau(q, y) := q + \frac{(y - q)_+}{1 - \tau}$$

which is strictly consistent for q_τ on the distribution class considered there and has Bayes risk CVaR_τ.

2.2 Multicalibration Error

Let $\mathcal{X}$ be a standard Borel context space and let D be a distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\mathsf{D}_{Y|X=x}$ denote a regular conditional law of Y given $X = x$. We call D $\mathcal{M}$-compatible if

$$\mathsf{D}_{Y|X=x} \in \mathcal{M} \qquad \text{for } \mathsf{D}_X\text{-almost every } x.$$

A randomized predictor assigns to each context $x \in \mathcal{X}$ a distribution $\Pi_x \in \Delta(\mathcal{P})$ over prediction values. This assignment is measurable in the sense that, for every Borel set $A \subseteq \mathcal{P}$, the map $x \mapsto \Pi_x(A)$ is measurable. We write

$$\Pi = (\Pi_x)_{x\in\mathcal{X}}.$$

For a given D and Π, consider the absolute-integrability condition

$$\mathbb{E}_{(X,Y)\sim\mathsf{D}} \left[\int_{\mathcal{P}} \|R(p, Y)\|_1 \, \Pi_X(dp) \right] < \infty. \tag{1}$$

When (1) holds, a measurable weight $w : \mathcal{X} \to [-1, 1]$ defines a finite vector signed measure on $\mathcal{P}$ by

$$\begin{aligned}\nu_w^{\mathsf{D},\Pi}(A) &:= \mathbb{E}_{(X,Y)\sim\mathsf{D}}\left[w(X)\int_A R(p,Y)\,\Pi_X(dp)\right] \\ &= \mathbb{E}_{\substack{(X,Y)\sim\mathsf{D}\\ P\sim\Pi_X}}\left[w(X)\mathbf{1}\{P\in A\}R(P,Y)\right], \qquad \text{for measurable } A\subseteq\mathcal{P}.\end{aligned}$$

The population expected calibration error for Γ (Γ-ECE) of Π with respect to w is the coordinatewise total variation

$$\begin{aligned}\mathrm{Err}_{\mathsf{D}}^{\Gamma}(\Pi; w) &:= \sum_{j=1}^{k} |\nu_{w,j}^{\mathsf{D},\Pi}|(\mathcal{P}) \\ &= \sum_{j=1}^{k} \mathbb{E}_{\substack{(X,Y)\sim\mathsf{D}\\ P\sim\Pi_X}}\left[\left|\mathbb{E}\left[w(X)R_j(P_{<j}, P_j, Y) \mid P\right]\right|\right].\end{aligned}$$

If R is not absolutely integrable under (D, Π), we set $\mathrm{Err}_{\mathsf{D}}^{\Gamma}(\Pi; w) = +\infty$. A group function is a measurable map $g : \mathcal{X} \to [0, 1]$. It is binary if it takes values in $\{0, 1\}$. For a finite family $\mathcal{G}$ of group functions, define

$$\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi; \mathcal{G}) := \max_{g\in\mathcal{G}} \mathrm{Err}_{\mathsf{D}}^{\Gamma}(\Pi; g).$$

Remark 2.2 (Bucketing by the Full Prediction Vector)**.** *When Π is supported on a finite set $S(\Pi) \subset \mathcal{P}$, the total-variation definition becomes*

$$\mathrm{Err}_{\mathsf{D}}^{\Gamma}(\Pi; g) = \sum_{p\in S(\Pi)} \left\|\mathbb{E}\left[g(X)\Pi_X(\{p\})R(p,Y)\right]\right\|_1.$$

Thus each bucket is indexed by $p = (p_1, \ldots, p_k)$, rather than by one coordinate at a time. Noarov and Roth [37] use a different error metric for two-level joint multicalibration. For each coordinate, they fix the other prediction coordinate and average squared calibration errors over the values of the coordinate being evaluated. Our ECE metric follows the definition for scalar and high-dimensional mean prediction studied by Collina et al. [9], summing the absolute residual mass over prediction values and levels. This joint bucketing gives the k-dimensional exponent.

2.3 Minimax Sample Complexity

We use the minimax formulation for batch multicalibration of Collina et al. [9], with the distributions restricted to those that are $\mathcal{M}$-compatible. A learner $\mathsf{A} = (\mathsf{A}_n)_{n\geq 1}$ receives a finite group family $\mathcal{G}$, a sample $S \in (\mathcal{X} \times \mathcal{Y})^n$, and an independent random seed ζ, and returns a randomized predictor $\mathsf{A}_n(\mathcal{G}, S, \zeta)$ with prediction space $\mathcal{P}$.

Let $B : (0, 1) \to \mathbb{N}$ be a group-size budget. For a learner A, let $n_{\mathsf{A}}^{\Gamma}(\varepsilon; B, \mathcal{M})$ be the smallest $n \geq 1$ such that, for every standard Borel context space $\mathcal{X}$, every $\mathcal{M}$-compatible distribution D on $\mathcal{X} \times \mathcal{Y}$, and every finite group family $\mathcal{G}$ with $1 \leq |\mathcal{G}| \leq B(\varepsilon)$,

$$\mathbb{P}_{S\sim\mathsf{D}^n,\,\zeta}\left[\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}\big(\mathsf{A}_n(\mathcal{G}, S, \zeta); \mathcal{G}\big) \leq \varepsilon\right] \geq \frac{2}{3}.$$

If no such n exists, set $n_{\mathsf{A}}^{\Gamma}(\varepsilon; B, \mathcal{M}) = +\infty$. The minimax sample complexity is

$$\mathrm{SC}_{\Gamma,\mathcal{M}}(\varepsilon; B) := \inf_{\mathsf{A}} n_{\mathsf{A}}^{\Gamma}(\varepsilon; B, \mathcal{M}).$$

For a fixed $\kappa > 0$, define

$$B_\kappa(\varepsilon) := \left\lceil \varepsilon^{-\kappa} \right\rceil, \qquad \mathrm{SC}^{(\kappa)}_{\Gamma,\mathcal{M}}(\varepsilon) := \mathrm{SC}_{\Gamma,\mathcal{M}}(\varepsilon; B_\kappa).$$

Constants suppressed by asymptotic notation may depend on the fixed parameter κ. Thus the context space, distribution, and group family may all vary with ε, subject to the compatibility and size requirements above.

3 The Lower Bound

This section proves the lower bound by reducing multicalibration to the task of identifying which distribution in a finite collection generated the observed data. We construct this collection so that a predictor with small multicalibration error is accurate enough to reveal the true member. We then use an information-theoretic argument to show that this identification cannot be accomplished from too few samples.

The construction must therefore satisfy two requirements. The same candidate distributions must yield sufficiently different property vectors at the contexts for an accurate predictor to distinguish them, yet remain statistically close enough that distinguishing them from the observed data requires many samples. We also construct a small family of groups such that small multicalibration error with respect to these groups guarantees the required prediction accuracy. The regularity assumptions below make these ingredients possible.

3.1 Regularity Assumptions

Sequential conditional identifiability determines the value at which the expected residual at each level vanishes, but it does not quantify how the residual changes when the prediction moves away from that value. For the lower bound, we require this quantitative control for a family of outcome distributions indexed by property vectors in a fixed rectangle $\mathcal{R}_0$. Specifically, for every $v \in \mathcal{R}_0$, the family contains a distribution μ_v on $\mathcal{Y}$ satisfying $\Gamma(\mu_v) = v$ (Condition (i) below). We will later use these outcome distributions as conditional laws at different contexts to construct the finite collection of candidate data distributions.

The remaining conditions govern how the residuals and distributions vary within this family. When the predictions at preceding levels are correct, the expected residual at the current level must have the same sign as the current prediction error, and its magnitude must be bounded above and below by constant multiples of that error (Conditions (ii) and (iii)). Errors in preceding predictions may change a later residual, but only in proportion to those errors (Condition (iv)). Finally, distributions with nearby property vectors must have small Kullback–Leibler divergence, so that the candidate distributions we construct are difficult to distinguish (Condition (v)).

Formally, let

$$\mathcal{R}_0 = I_1^0 \times \cdots \times I_k^0 \subset \mathrm{int}(\mathcal{P})$$

be a nondegenerate compact rectangle. A point in $\mathcal{R}_0$ is denoted

$$v = (v_1, \ldots, v_k).$$

Write

$$I_j^0 = [\ell_j, \ell_j + W_j], \qquad W_j > 0, \qquad j \in [k].$$

For $v \in \mathcal{R}_0$, write $v_{<j} := (v_1, \ldots, v_{j-1})$. We interpret $v_{<1}$ as the empty prefix. Thus $R_j(v_{<j}, q, \mu_v)$ is the expected level-j residual evaluated with the true prefix.

Assumption 3.1 (Regularity of the Witness Family)**.** *There exists a family of outcome distributions*

$$\{\mu_v : v \in \mathcal{R}_0\} \subset \mathcal{M}$$

and constants $c_{\mathrm{anti}}, C_{\mathrm{prev}}, C_{\mathrm{lip}}, C_{\mathrm{KL}} \in (0, \infty)$ *such that the following hold.*

(i) For every $v \in \mathcal{R}_0$*,*

$$\Gamma(\mu_v) = v.$$

(ii) For every $v \in \mathcal{R}_0$*, every* $j \in [k]$*, and every* $q \in \mathcal{P}_j$*,*

$$\operatorname{sgn}(q - v_j)\,(R_j(v_{<j}, q, \mu_v) - R_j(v_{<j}, v_j, \mu_v)) \geq c_{\mathrm{anti}}|q - v_j|.$$

Here and throughout this section, we use the convention $\operatorname{sgn}(0) = 0$*.*

(iii) For every $v \in \mathcal{R}_0$*, every* $j \in [k]$*, and every* $q \in \mathcal{P}_j$*,*

$$|R_j(v_{<j}, q, \mu_v) - R_j(v_{<j}, v_j, \mu_v)| \leq C_{\mathrm{lip}}|q - v_j|.$$

(iv) For every $v \in \mathcal{R}_0$*, every* $p \in \mathcal{P}$*, and every* $j \in [k]$*,*

$$|R_j(p_{<j}, p_j, \mu_v) - R_j(v_{<j}, p_j, \mu_v)| \leq C_{\mathrm{prev}} \sum_{i<j} |p_i - v_i|.$$

The sum over $i < 1$ *is interpreted as zero.*

(v) For every $v, v' \in \mathcal{R}_0$*,*

$$D_{\mathrm{KL}}(\mu_v \,\|\, \mu_{v'}) \leq C_{\mathrm{KL}}\|v - v'\|_2^2.$$

Collina et al. [9] use counterparts of Conditions (i), (ii), and (v) in their lower bounds for scalar properties. Our proof additionally uses (iii) to bound the magnitude of a residual by the current prediction error and (iv) to bound the effect of errors at earlier levels.

As a quick sanity check, for mean and mean absolute deviation, the expected level-j residual at the true prefix is exactly $q - v_j$. Thus, (ii) and (iii) hold as equalities with constant 1, and (iv) also holds with constant 1. For quantile and CVaR, (ii) and (iii) in the quantile coordinate follow directly from lower and upper bounds on the density, while in the CVaR coordinate they hold as equalities with constant 1. Section 5 gives the complete verifications.

3.2 Main Result

The conditions above allow us to construct a finite collection of candidate data distributions for every sufficiently small ε. Any learner that achieves multicalibration error at most ε on every distribution in the collection must use $\widetilde{\Omega}(\varepsilon^{-(k+2)})$ samples. The corresponding group family has only polylogarithmically many members and therefore satisfies $|\mathcal{G}_\varepsilon| \leq \varepsilon^{-\kappa}$ for every fixed $\kappa > 0$ once ε is sufficiently small. The following theorem states this result precisely.

Theorem 3.2 (Lower Bound for Multicalibration of k-Level Properties)**.** *Suppose Assumption 3.1 holds. Fix any* $\kappa > 0$*. Then, for every sufficiently small* $\varepsilon > 0$*, there exist*

(i) a finite context space $\mathcal{X}_\varepsilon$*;*

(ii) a binary group family $\mathcal{G}_\varepsilon$ on $\mathcal{X}_\varepsilon$ satisfying

$$|\mathcal{G}_\varepsilon| \le \varepsilon^{-\kappa};$$

(iii) a finite collection $\mathfrak{D}_\varepsilon$ of $\mathcal{M}$-compatible distributions on $\mathcal{X}_\varepsilon \times \mathcal{Y}$;

such that any possibly randomized learner which, given n i.i.d. samples from any $\mathsf{D} \in \mathfrak{D}_\varepsilon$, outputs a predictor Π satisfying

$$\mathrm{MCErr}^{\Gamma}_{\mathsf{D}}(\Pi; \mathcal{G}_\varepsilon) \le \varepsilon$$

with probability at least $2/3$, must use

$$n = \Omega\left(\frac{\varepsilon^{-(k+2)}}{\log^{k+2}(1/\varepsilon)}\right) = \widetilde{\Omega}\left(\varepsilon^{-(k+2)}\right).$$

The implicit constants, and the threshold for sufficiently small ε, depend only on k, κ, $\mathcal{R}_0$, and the constants in Assumption 3.1.

Equivalently, in the notation of Section 2.3,

$$\mathrm{SC}^{(\kappa)}_{\Gamma,\mathcal{M}}(\varepsilon) = \Omega\left(\frac{\varepsilon^{-(k+2)}}{\log^{k+2}(1/\varepsilon)}\right) = \widetilde{\Omega}\left(\varepsilon^{-(k+2)}\right).$$

We prove Theorem 3.2 in the remainder of this section, starting with the construction of the candidate distributions.

3.3 The Candidate Distributions

Fix a power of two $Q \ge 2$. We take the context space to be the discrete grid

$$\mathcal{X}_Q = \{0, \dots, Q-1\}^k.$$

We write a context as

$$a = (a_1, \dots, a_k) \in \mathcal{X}_Q.$$

Recall that $I^0_j = [\ell_j, \ell_j + W_j]$ is the j-th coordinate interval of $\mathcal{R}_0$, with lower endpoint ℓ_j and width $W_j > 0$. For $j \in [k]$ and $u \in \{0, \dots, Q-1\}$, define

$$b_j(u) := \ell_j + \frac{W_j}{3} + \frac{u\, W_j}{3(Q-1)}.$$

Thus $b_j(0), \dots, b_j(Q-1)$ are Q evenly spaced values in the j-th property coordinate. For $a \in \mathcal{X}_Q$, define the unperturbed property vector associated with context a

$$b(a) := (b_1(a_1), \dots, b_k(a_k)).$$

The vectors $b(a)$, as a ranges over $\mathcal{X}_Q$, form a Cartesian grid in the property space. This grid lies in the central third

$$\prod_{j=1}^{k} \left[\ell_j + \frac{W_j}{3},\, \ell_j + \frac{2W_j}{3}\right] \subset \mathcal{R}_0.$$

The distance between consecutive values $b_j(u)$ is $W_j/[3(Q-1)]$. Let d_Q be the smallest of these distances:

$$d_Q := \frac{\min_{j\in[k]} W_j}{3(Q-1)}.$$

In particular,

$$d_Q \asymp Q^{-1},$$

with constants depending only on $\mathcal{R}_0$.

As in the construction used by Collina et al. [9] for vector means, we assign one hidden sign to each context. At each context, this sign determines the direction in which we perturb the final property coordinate. Keeping the preceding coordinates at their unperturbed values allows their prediction errors to be controlled successively. The direction in which the final coordinate is perturbed can be chosen separately at each of the Q^k contexts, yielding exponentially many candidate data distributions.

We choose the hidden perturbation to be small relative to d_Q. This keeps every perturbed property vector inside the witness rectangle and ensures that the property vectors assigned to different contexts remain well separated. Choose a constant $c_\eta > 0$ satisfying

$$c_\eta < \frac{1}{20},$$

and set

$$\eta := c_\eta d_Q.$$

Then, for every $a \in \{0,\dots,Q-1\}^k$ and every $\theta_a \in \{-1,+1\}$,

$$b(a) + \eta\theta_a e_k \in \mathcal{R}_0.$$

We need to choose exponentially many sign assignments while ensuring that any two disagree on a constant fraction of the contexts. Let

$$\Theta_Q \subseteq \{-1,+1\}^{\mathcal{X}_Q}$$

satisfy

$$\log|\Theta_Q| \ge c_{\mathrm{pack}} Q^k, \qquad \big|\{a \in \mathcal{X}_Q : \theta_a \ne \theta'_a\}\big| \ge \rho_{\mathrm{pack}} Q^k \quad \forall \theta \ne \theta',$$

for universal constants $c_{\mathrm{pack}}, \rho_{\mathrm{pack}} > 0$. Such a collection exists by the Gilbert packing bound [42, 46].

Each sign assignment specifies the true property vector at every context, and hence one candidate distribution. For $\theta \in \Theta_Q$, define the true property vector at context a by

$$v_\theta(a) := b(a) + \eta\theta_a e_k.$$

Thus

$$v_{\theta,j}(a) = b_j(a_j) \quad (j < k), \qquad v_{\theta,k}(a) = b_k(a_k) + \eta\theta_a.$$

The distribution D_θ on $\mathcal{X}_Q \times \mathcal{Y}$ is

$$X \sim \mathrm{Unif}(\mathcal{X}_Q), \qquad Y \mid X = a \sim \mu_{v_\theta(a)}.$$

We next construct a group family for which small multicalibration error forces a predictor's outputs to be close to $v_\theta(a)$. This prediction guarantee will allow us to identify θ.

3.4 Approximation of Threshold Signs with Walsh Functions

The key relation, formalized later in Lemma 3.5, has the schematic form

$$|p_j - v_{\theta,j}(a)| \lesssim \operatorname{sgn}\big(p_j - v_{\theta,j}(a)\big) R_j\big(p_{<j}, p_j, \mu_{v_\theta(a)}\big) + \sum_{i<j} |p_i - v_{\theta,i}(a)|.$$

Thus, once the errors in the preceding coordinates have been controlled, the error in coordinate j can be bounded through a residual multiplied by the sign of the current error.

Multicalibration, on the other hand, controls residuals multiplied by groups $g(a)$. We therefore seek to approximate the sign in the inequality above by a linear combination of a common family of functions of the context. This will allow us to bound the prediction error by the multicalibration error multiplied by a factor. The sum of the absolute values of the coefficients in the linear combination determines this factor, while the number of functions determines the size of the resulting group family.

For $j < k$, we have $v_{\theta,j}(a) = b_j(a_j)$, so for a fixed prediction p_j, the sign

$$a \mapsto \operatorname{sgn}\big(p_j - b_j(a_j)\big)$$

is a threshold function of a_j. For the final coordinate, the two possible true values $b_k(a_k) - \eta$ and $b_k(a_k) + \eta$ lie in a small interval around $b_k(a_k)$. Outside a slightly larger interval, the sign of the error is the same for both values. We use a two-sided threshold that agrees with this common sign outside the interval and is zero inside it, where predictions will be handled separately.

Because the group family must be fixed independently of the prediction and of θ, we need a common set of functions that can represent all these one-sided and two-sided threshold signs, with uniformly controlled coefficients. The full Walsh basis gives an exact representation with coefficient bound $O(\log Q)$, but it would require $O(Q)$ groups. By retaining a suitable common subset of the Walsh functions, we can uniformly approximate all the required threshold signs while preserving the $O(\log Q)$ coefficient bound. Only polylogarithmically many Walsh functions are needed.

Because Q is a power of two, identify each $u, \ell \in \{0, \dots, Q-1\}$ with its length-$\log_2 Q$ binary expansion. Define the Walsh function

$$\psi_\ell(u) := (-1)^{\langle \ell, u \rangle_{\mathbb{F}_2}}.$$

Here

$$\langle \ell, u \rangle_{\mathbb{F}_2} := \sum_{i=1}^{\log_2 Q} \ell_i u_i \pmod 2.$$

The Walsh functions form an orthonormal basis for real-valued functions on $\{0, \dots, Q-1\}$.

For $\sigma \in \{-1, +1\}$ and $t \in \mathbb{R}$, define the one-sided function

$$s_{\sigma,t}(u) := \sigma \operatorname{sgn}(t - u).$$

For $\sigma \in \{-1, +1\}$ and $t_- \le t_+$, define the two-sided function

$$s_{\sigma,t_-,t_+}(u) := \frac{\sigma}{2}\big(\operatorname{sgn}(t_- - u) + \operatorname{sgn}(t_+ - u)\big).$$

Let $\mathcal{T}_Q$ be the class of all these functions. A one-sided function changes sign at one threshold. A two-sided function is zero between two thresholds and has opposite signs beyond them, with the boundary values determined by $\operatorname{sgn}(0) = 0$. A two-sided function tests whether the final prediction lies outside a prescribed interval.

Collina et al. [8] show that one common polylogarithmic subset of the Walsh basis approximates every one-sided discrete threshold sign. Their coefficients satisfy the uniform $O(\log Q)$ bound in Lemma 3.3.

Lemma 3.3 (Walsh Approximation of Threshold Signs)**.** *There is a universal constant C_{W} such that the following holds. Let $Q \geq 2$ be a power of two. For every fixed accuracy $\alpha \in (0,1)$, there exists a set*

$$\mathcal{S}_\alpha \subseteq \{1, \dots, Q-1\}$$

of indices of Walsh functions satisfying

$$|\mathcal{S}_\alpha| \leq C_{\mathrm{W}} \alpha^{-2} \log^3(Q+1)$$

and coefficient functions

$$\beta_0 : \mathcal{T}_Q \to \mathbb{R}, \qquad \beta_\ell : \mathcal{T}_Q \to \mathbb{R} \quad (\ell \in \mathcal{S}_\alpha),$$

such that, for every $s \in \mathcal{T}_Q$,

$$\widehat{s}(u) = \beta_0(s) + \sum_{\ell \in \mathcal{S}_\alpha} \beta_\ell(s) \psi_\ell(u)$$

is a uniform approximation:

$$\|s - \widehat{s}\|_\infty \leq \alpha$$

and the coefficients satisfy

$$\sup_{s \in \mathcal{T}_Q} |\beta_0(s)| + \sum_{\ell \in \mathcal{S}_\alpha} \sup_{s \in \mathcal{T}_Q} |\beta_\ell(s)| \leq C_{\mathrm{W}} \log(Q+1).$$

Proof. For each $r \in \{0, \dots, Q\}$, define

$$f_r : \{0, \dots, Q-1\} \to \{-1, +1\}$$

by

$$f_r(u) := \begin{cases} +1, & u < r, \\ -1, & u \geq r. \end{cases}$$

Collina et al. [8] construct a common set

$$\mathcal{S}_\alpha \subseteq \{1, \dots, Q-1\}, \qquad |\mathcal{S}_\alpha| \leq C \alpha^{-2} \log^3(Q+1),$$

together with coefficients $a_0(r)$ and $c_\ell(r)$ for all $r \in \{0, \dots, Q\}$. The corresponding approximations

$$\widetilde{f}_r(u) := a_0(r) + \sum_{\ell \in \mathcal{S}_\alpha} c_\ell(r) \psi_\ell(u)$$

satisfy

$$\max_{r \in \{0,\dots,Q\}} \|f_r - \widetilde{f}_r\|_\infty \leq \alpha$$

and

$$\max_{r \in \{0,\dots,Q\}} |a_0(r)| + \sum_{\ell \in \mathcal{S}_\alpha} \max_{r \in \{0,\dots,Q\}} |c_\ell(r)| \leq C \log(Q+1)$$

for a universal constant C.

We use these coefficients to define β_0 and β_ℓ on $\mathcal{T}_Q$. For each function s that has a one-sided representation, fix one such representation $s = s_{\sigma,t}$. If $t \notin \{0, \dots, Q-1\}$, the restriction of $u \mapsto \operatorname{sgn}(t-u)$ to $\{0, \dots, Q-1\}$ equals a unique f_r. Define

$$\beta_0(s) := \sigma a_0(r), \qquad \beta_\ell(s) := \sigma c_\ell(r) \quad (\ell \in \mathcal{S}_\alpha).$$

If $t \in \{0, \dots, Q-1\}$, then

$$\operatorname{sgn}(t-u) = \frac{f_t(u) + f_{t+1}(u)}{2}.$$

In this case, define

$$\beta_0(s) := \frac{\sigma}{2}\big(a_0(t) + a_0(t+1)\big)$$

and

$$\beta_\ell(s) := \frac{\sigma}{2}\big(c_\ell(t) + c_\ell(t+1)\big) \quad (\ell \in \mathcal{S}_\alpha).$$

For each function $s \in \mathcal{T}_Q$ not already covered, fix a two-sided representation $s = s_{\sigma,t_-,t_+}$. By definition,

$$s_{\sigma,t_-,t_+}(u) = \frac{s_{\sigma,t_-}(u) + s_{\sigma,t_+}(u)}{2}.$$

Using the coefficients already defined for these two one-sided functions, set

$$\beta_0(s) := \frac{\beta_0(s_{\sigma,t_-}) + \beta_0(s_{\sigma,t_+})}{2}$$

and

$$\beta_\ell(s) := \frac{\beta_\ell(s_{\sigma,t_-}) + \beta_\ell(s_{\sigma,t_+})}{2} \quad (\ell \in \mathcal{S}_\alpha).$$

In each case, $\widehat{s}$ is obtained by applying the same multiplication and averaging operations to the corresponding approximations $\widetilde{f_r}$. The triangle inequality therefore gives

$$\|s - \widehat{s}\|_\infty \le \alpha.$$

Every coefficient $\beta_0(s)$ is a signed average of coefficients $a_0(r)$, and every $\beta_\ell(s)$ is the corresponding signed average of coefficients $c_\ell(r)$. Consequently,

$$\sup_{s \in \mathcal{T}_Q} |\beta_0(s)| \le \max_{r \in \{0,\dots,Q\}} |a_0(r)|$$

and, for every $\ell \in \mathcal{S}_\alpha$,

$$\sup_{s \in \mathcal{T}_Q} |\beta_\ell(s)| \le \max_{r \in \{0,\dots,Q\}} |c_\ell(r)|.$$

The required coefficient bound now follows from the corresponding bound on $a_0(r)$ and $c_\ell(r)$, after enlarging the universal constant if necessary. □

3.5 The Group Family

We now use the Walsh approximation above to construct a family of binary groups.

Fix a sufficiently small constant $\alpha \in (0,1)$, and let $\mathcal{S}_\alpha$ be the set supplied by Lemma 3.3. For each coordinate $j \in [k]$ and each $\ell \in \mathcal{S}_\alpha$, define the signed Walsh weight

$$w_{j,\ell}(a) := \psi_\ell(a_j).$$

Convert it into two binary groups

$$g^+_{j,\ell}(a) := \frac{1 + w_{j,\ell}(a)}{2}, \qquad g^-_{j,\ell}(a) := \frac{1 - w_{j,\ell}(a)}{2}.$$

Let

$$g_{\text{all}} \equiv 1,$$

and define

$$\mathcal{G}_Q := \{g_{\mathrm{all}}\} \cup \{g^+_{j,\ell}, g^-_{j,\ell} : \ j \in [k], \ \ell \in \mathcal{S}_\alpha\}.$$

Then $\mathcal{G}_Q$ is binary-valued and satisfies

$$|\mathcal{G}_Q| \le C_{\alpha,k} \log^3(Q+1).$$

The analysis uses the signed Walsh weights $w_{j,\ell}$, although the group family itself is binary-valued. The following elementary lemma passes between these two forms.

Lemma 3.4 (Signed Weights from Binary Groups)**.** *For every $j \in [k]$ and $\ell \in \mathcal{S}_\alpha$, every distribution D on $\mathcal{X}_Q \times \mathcal{Y}$, and every predictor Π,*

$$\mathrm{Err}^\Gamma_{\mathsf{D}}(\Pi; w_{j,\ell}) \le 2 \ \mathrm{MCErr}^\Gamma_{\mathsf{D}}(\Pi; \mathcal{G}_Q).$$

Proof. Since $w_{j,\ell} = g^+_{j,\ell} - g^-_{j,\ell}$, its vector signed measure is the difference of the vector signed measures for $g^+_{j,\ell}$ and $g^-_{j,\ell}$. Both groups belong to $\mathcal{G}_Q$, and coordinatewise total variation is subadditive. □

3.6 From Multicalibration to Prediction Accuracy

Assumption 3.1(ii) gives a lower bound when the preceding predictions equal their true values, but a predictor need not use the true prefix. The following lemma relates the residual at an arbitrary prediction to the prediction error at the current level and the errors at preceding levels.

Lemma 3.5 (Relating Residuals to Prediction Errors)**.** *Under Assumption 3.1, for every $v \in \mathcal{R}_0$, every $p \in \mathcal{P}$, and every $j \in [k]$,*

$$\mathrm{sgn}(p_j - v_j)\left(R_j(p_{<j}, p_j, \mu_v) - R_j(v_{<j}, v_j, \mu_v)\right) \ge c_{\mathrm{anti}}|p_j - v_j| - C_{\mathrm{prev}} \sum_{i<j} |p_i - v_i|,$$

and

$$|R_j(p_{<j}, p_j, \mu_v) - R_j(v_{<j}, v_j, \mu_v)| \le C_{\mathrm{lip}}|p_j - v_j| + C_{\mathrm{prev}} \sum_{i<j} |p_i - v_i|.$$

Proof. Because $\Gamma(\mu_v) = v$ by Assumption 3.1(i), $R_j(v_{<j}, v_j, \mu_v) = 0$. Thus it suffices to prove the same two inequalities with $R_j(p_{<j}, p_j, \mu_v)$ in place of $R_j(p_{<j}, p_j, \mu_v) - R_j(v_{<j}, v_j, \mu_v)$. For the lower bound,

$$\mathrm{sgn}(p_j - v_j) R_j(p_{<j}, p_j, \mu_v) \ge \mathrm{sgn}(p_j - v_j) R_j(v_{<j}, p_j, \mu_v) - \left|R_j(p_{<j}, p_j, \mu_v) - R_j(v_{<j}, p_j, \mu_v)\right|.$$

Assumption 3.1(ii) controls the first term, and condition (iv) controls the second. For the upper bound, use the triangle inequality and $R_j(v_{<j}, v_j, \mu_v) = 0$, followed by the bounds in conditions (iii) and (iv). □

The preceding lemma relates prediction error to a residual multiplied by the sign of the prediction error. We approximate the threshold signs that arise in this argument by linear combinations of the fixed Walsh functions that define $\mathcal{G}_Q$. Lemma 3.4 shows that Γ-ECE controls residuals weighted by each of these Walsh functions. The next lemma shows that Γ-ECE also controls a residual weighted by such a linear combination.

Lemma 3.6 (Multicalibration Controls Weighted Residuals)**.** *Fix θ, a predictor Π, and a level $j \in [k]$. Suppose a function $h : \mathcal{P} \times \mathcal{X}_Q \to \mathbb{R}$ has the form*

$$h(p, a) = \beta_0(p) + \sum_{\ell \in \mathcal{S}_\alpha} \beta_\ell(p) w_{i,\ell}(a),$$

for some context coordinate $i \in [k]$ and measurable coefficient functions

$$\beta_0 : \mathcal{P} \to \mathbb{R}, \qquad \beta_\ell : \mathcal{P} \to \mathbb{R} \quad (\ell \in \mathcal{S}_\alpha),$$

satisfying

$$\sup_{p \in \mathcal{P}} |\beta_0(p)| + \sum_{\ell \in \mathcal{S}_\alpha} \sup_{p \in \mathcal{P}} |\beta_\ell(p)| \le C_{\mathrm{coef}}.$$

Then

$$\left| \frac{1}{Q^k} \sum_{a \in \mathcal{X}_Q} \int_{\mathcal{P}} h(p, a)\, R_j(p_{<j}, p_j, \mu_{v_\theta(a)})\, \Pi_a(dp) \right| \le 2C_{\mathrm{coef}}\ \mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q).$$

Proof. For a signed weight w, the j-th coordinate signed measure under (D_θ, Π) is

$$\nu^{\mathsf{D}_\theta, \Pi}_{w,j}(A) := \frac{1}{Q^k} \sum_{a \in \mathcal{X}_Q} w(a) \int_A R_j(p_{<j}, p_j, \mu_{v_\theta(a)})\, \Pi_a(dp), \qquad A \subseteq \mathcal{P}.$$

Therefore

$$\frac{1}{Q^k} \sum_a \int h(p, a) R_j(p_{<j}, p_j, \mu_{v_\theta(a)}) \Pi_a(dp) = \int \beta_0(p)\, \nu^{\mathsf{D}_\theta, \Pi}_{g_{\mathrm{all}}, j}(dp) + \sum_{\ell \in \mathcal{S}_\alpha} \int \beta_\ell(p)\, \nu^{\mathsf{D}_\theta, \Pi}_{w_{i,\ell}, j}(dp).$$

By total variation,

$$\left| \int \beta_0\, d\nu^{\mathsf{D}_\theta, \Pi}_{g_{\mathrm{all}}, j} \right| \le \sup_p |\beta_0(p)|\, |\nu^{\mathsf{D}_\theta, \Pi}_{g_{\mathrm{all}}, j}|(\mathcal{P}) \le \sup_p |\beta_0(p)|\ \mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q).$$

For each signed Walsh weight $w_{i,\ell} = g^+_{i,\ell} - g^-_{i,\ell}$, Lemma 3.4 gives

$$|\nu^{\mathsf{D}_\theta, \Pi}_{w_{i,\ell}, j}|(\mathcal{P}) \le \mathrm{Err}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; w_{i,\ell}) \le 2\, \mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q).$$

Therefore

$$\begin{aligned} &\left| \frac{1}{Q^k} \sum_a \int h(p, a) R_j(p_{<j}, p_j, \mu_{v_\theta(a)}) \Pi_a(dp) \right| \\ &\le \left(\sup_p |\beta_0(p)| + 2 \sum_{\ell \in \mathcal{S}_\alpha} \sup_p |\beta_\ell(p)| \right) \mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q) \\ &\le 2C_{\mathrm{coef}}\, \mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q). \end{aligned}$$

□

We now combine the residual inequalities with the group family constructed above. To establish their lower bound for multicalibration of vector-valued means, Collina et al. [9] divide the prediction error into two parts. Coordinatewise signs control the contribution from predictions that are far from the unperturbed mean for their context, while the all-ones group controls the contribution

from predictions that are nearby. Splitting predictions into these two cases is not enough here, because an error in an earlier coordinate can alter every later residual. We first use threshold signs to control coordinates $1, \dots, k-1$ successively. With the prefix errors controlled, two-sided threshold signs bound the error in the final coordinate outside the correct box, and the all-ones group handles predictions inside it. The resulting bound controls the average prediction error by Γ-ECE.

For a predictor $\Pi = (\Pi_a)_{a \in \mathcal{X}_Q}$, define its average error in predicting the property under θ by

$$\mathrm{PredErr}_\theta(\Pi) := \frac{1}{Q^k} \sum_{a \in \mathcal{X}_Q} \int_{\mathcal{P}} \|p - v_\theta(a)\|_1 \, \Pi_a(dp).$$

Also define the coordinate prediction errors

$$\mathrm{PredErr}_{\theta,j}(\Pi) := \frac{1}{Q^k} \sum_{a \in \mathcal{X}_Q} \int_{\mathcal{P}} |p_j - v_{\theta,j}(a)| \, \Pi_a(dp), \qquad j \in [k].$$

Thus

$$\mathrm{PredErr}_\theta(\Pi) = \sum_{j=1}^{k} \mathrm{PredErr}_{\theta,j}(\Pi).$$

Proposition 3.7 (Γ-ECE Controls Prediction Error)**.** *Under Assumption 3.1, there exists a constant $C_{\mathrm{pred}} < \infty$, depending only on k, $\mathcal{R}_0$, and the constants in Assumption 3.1, such that for every sufficiently large power of two Q, every $\theta \in \Theta_Q$, and every randomized predictor Π,*

$$\mathrm{PredErr}_\theta(\Pi) \le C_{\mathrm{pred}} \log(Q+1) \; \mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q).$$

Proof. All constants in this proof may depend on k, $\mathcal{R}_0$, and the constants in Assumption 3.1, but not on Q, θ, or Π. The accuracy α in the Walsh approximation was fixed small enough that the terms $C\alpha \, \mathrm{PredErr}_{\theta,j}(\Pi)$ that arise can be absorbed into the left-hand side. For brevity, write

$$e_i(a,p) := |p_i - v_{\theta,i}(a)|, \qquad \mathsf{R}_{a,i}(p) := R_i(p_{<i}, p_i, \mu_{v_\theta(a)}).$$

Because $\Gamma(\mu_{v_\theta(a)}) = v_\theta(a)$ by Assumption 3.1(i), $\mathsf{R}_{a,i}(v_\theta(a)) = 0$.

Controlling the First $k-1$ Coordinates. For $j < k$ and $p_j \in \mathcal{P}_j$, define

$$s^{(j)}_{p_j}(a) := \mathrm{sgn}(p_j - b_j(a_j)).$$

Since $v_{\theta,j}(a) = b_j(a_j)$ for $j < k$, this is $\mathrm{sgn}(p_j - v_{\theta,j}(a))$. As a function of a_j, it is one of the one-sided threshold signs in $\mathcal{T}_Q$. Let $\widehat{s}^{(j)}_{p_j}$ be its Walsh approximation from Lemma 3.3. Lemma 3.5 gives

$$s^{(j)}_{p_j}(a) \mathsf{R}_{a,j}(p) \ge c_{\mathrm{anti}} e_j(a,p) - C_{\mathrm{prev}} \sum_{i<j} e_i(a,p),$$

and also

$$|\mathsf{R}_{a,j}(p)| \le C e_j(a,p) + C \sum_{i<j} e_i(a,p).$$

Therefore, using $\|\widehat{s}^{(j)}_{p_j} - s^{(j)}_{p_j}\|_\infty \le \alpha$,

$$\widehat{s}^{(j)}_{p_j}(a) \mathsf{R}_{a,j}(p) \ge c e_j(a,p) - C \sum_{i<j} e_i(a,p)$$

after fixing α sufficiently small and decreasing $c > 0$. Integrating over (a, p) and rearranging gives

$$c\,\mathrm{PredErr}_{\theta,j}(\Pi) \le \left| \frac{1}{Q^k} \sum_a \int \widehat{s}_{p_j}^{(j)}(a) \mathsf{R}_{a,j}(p)\, \Pi_a(dp) \right| + C \sum_{i<j} \mathrm{PredErr}_{\theta,i}(\Pi).$$

The Walsh coefficients of $\widehat{s}_{p_j}^{(j)}(a)$, viewed as a function of a, are step functions of p_j and hence measurable. They also satisfy the uniform bound $C\log(Q+1)$. Lemma 3.6 therefore bounds the term inside the absolute value by

$$C\log(Q+1)\,\mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q),$$

and hence

$$\mathrm{PredErr}_{\theta,j}(\Pi) \le C\log(Q+1)\,\mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q) + C \sum_{i<j} \mathrm{PredErr}_{\theta,i}(\Pi).$$

Induction over $j = 1, \dots, k-1$ yields

$$\sum_{j<k} \mathrm{PredErr}_{\theta,j}(\Pi) \le C\log(Q+1)\,\mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q). \tag{2}$$

Predictions Outside the Boxes. Set

$$r := 5\eta.$$

Since $c_\eta < 1/20$, we have $r < d_Q/4$. Define boxes around the unperturbed property vectors by

$$C_a := \{p \in \mathcal{P} :\ |p_i - b_i(a_i)| \le r \text{ for every } i \in [k]\}.$$

Each true property vector $v_\theta(a)$ lies in C_a, and the boxes are pairwise disjoint because $2r < d_Q$. Let

$$E_k^{\mathrm{out}} := \frac{1}{Q^k} \sum_a \int_{\mathcal{P}\setminus C_a} |p_k - v_{\theta,k}(a)|\, \Pi_a(dp).$$

Decompose this error according to whether the final prediction is near $b_k(a_k)$. Define

$$E_k^{\mathrm{out,near}} := \frac{1}{Q^k} \sum_a \int_{\mathcal{P}\setminus C_a} \mathbf{1}\{|p_k - b_k(a_k)| \le r\} |p_k - v_{\theta,k}(a)|\, \Pi_a(dp),$$
$$E_k^{\mathrm{out,far}} := \frac{1}{Q^k} \sum_a \int_{\mathcal{P}\setminus C_a} \mathbf{1}\{|p_k - b_k(a_k)| > r\} |p_k - v_{\theta,k}(a)|\, \Pi_a(dp).$$

Then

$$E_k^{\mathrm{out}} = E_k^{\mathrm{out,near}} + E_k^{\mathrm{out,far}}.$$

On the near part, $p \notin C_a$ implies that some prefix coordinate $i < k$ has $|p_i - b_i(a_i)| > r$, while

$$|p_k - v_{\theta,k}(a)| \le r + \eta \le C \sum_{i<k} |p_i - v_{\theta,i}(a)|.$$

The second inequality uses $v_{\theta,i}(a) = b_i(a_i)$ for $i < k$, and the fact that one prefix error is at least r while r and η are fixed constant multiples of one another. Therefore

$$E_k^{\mathrm{out,near}} \le C \sum_{i<k} \mathrm{PredErr}_{\theta,i}(\Pi).$$

For the far part, define the two-sided threshold sign

$$s_{p_k}^{(k)}(a) := \frac{1}{2}\left(\operatorname{sgn}(p_k - r - b_k(a_k)) + \operatorname{sgn}(p_k + r - b_k(a_k))\right).$$

As a function of a_k, this belongs to $\mathcal{T}_Q$. On the event $|p_k - b_k(a_k)| > r$, it agrees with $\operatorname{sgn}(p_k - v_{\theta,k}(a))$, because $|v_{\theta,k}(a) - b_k(a_k)| = \eta < r$. Lemma 3.5 gives

$$c_{\text{anti}}\mathbf{1}\{|p_k - b_k(a_k)| > r\}|p_k - v_{\theta,k}(a)| \le s_{p_k}^{(k)}(a)\mathsf{R}_{a,k}(p) + C\sum_{i<k}|p_i - v_{\theta,i}(a)|.$$

Let $\widehat{s}_{p_k}^{(k)}$ be the Walsh approximation to $s_{p_k}^{(k)}$. Replacing $s_{p_k}^{(k)}$ by $\widehat{s}_{p_k}^{(k)}$, integrating, and controlling the approximation error gives

$$\begin{aligned} c_{\text{anti}}E_k^{\text{out,far}} \le &\left|\frac{1}{Q^k}\sum_a \int \widehat{s}_{p_k}^{(k)}(a)\mathsf{R}_{a,k}(p)\,\Pi_a(dp)\right| \\ &+ C\sum_{i<k}\operatorname{PredErr}_{\theta,i}(\Pi) + C\alpha\operatorname{PredErr}_{\theta,k}(\Pi). \end{aligned}$$

Indeed, the extra $C\alpha\operatorname{PredErr}_{\theta,k}(\Pi)$ term comes from

$$\left|\widehat{s}_{p_k}^{(k)}(a) - s_{p_k}^{(k)}(a)\right|\,|\mathsf{R}_{a,k}(p)| \le \alpha\left(Ce_k(a,p) + C\sum_{i<k}e_i(a,p)\right),$$

where the residual bound is the upper inequality in Lemma 3.5. The Walsh coefficients again satisfy the conditions of Lemma 3.6, with coefficient bound $C\log(Q+1)$, so that lemma bounds the term inside the absolute value by

$$C\log(Q+1)\operatorname{MCErr}_{\mathsf{D}_\theta}^{\Gamma}(\Pi;\mathcal{G}_Q).$$

Therefore

$$E_k^{\text{out,far}} \le C\log(Q+1)\operatorname{MCErr}_{\mathsf{D}_\theta}^{\Gamma}(\Pi;\mathcal{G}_Q) + C\sum_{i<k}\operatorname{PredErr}_{\theta,i}(\Pi) + C\alpha\operatorname{PredErr}_{\theta,k}(\Pi).$$

Combining the near and far bounds with (2),

$$E_k^{\text{out}} \le C\log(Q+1)\operatorname{MCErr}_{\mathsf{D}_\theta}^{\Gamma}(\Pi;\mathcal{G}_Q) + C\alpha\operatorname{PredErr}_{\theta,k}(\Pi). \tag{3}$$

Predictions Inside the Boxes. Inside the boxes, we use the all-ones group rather than another threshold sign. Because the boxes C_a are pairwise disjoint, the residual contributions localized to different contexts occupy disjoint regions of the prediction space and therefore do not cancel in total variation. The all-ones residual also includes predictions outside the boxes, whose contribution is controlled by the preceding bounds.

Let

$$E_k^{\text{loc}} := \frac{1}{Q^k}\sum_a \int_{C_a} |p_k - v_{\theta,k}(a)|\,\Pi_a(dp),$$

so that

$$\operatorname{PredErr}_{\theta,k}(\Pi) = E_k^{\text{loc}} + E_k^{\text{out}}.$$

Let $\nu^{\mathsf{D}_\theta,\Pi}_{g_{\mathrm{all}},k}$ be the signed measure for the final coordinate and the all-ones group, and define its localized part by

$$\nu^{\mathrm{loc}}_k(A) := \frac{1}{Q^k}\sum_a \int_{A\cap C_a} \mathsf{R}_{a,k}(p)\,\Pi_a(dp).$$

Because the boxes C_a are pairwise disjoint,

$$|\nu^{\mathrm{loc}}_k|(\mathcal{P}) = \frac{1}{Q^k}\sum_a \int_{C_a} |\mathsf{R}_{a,k}(p)|\,\Pi_a(dp).$$

The lower inequality in Lemma 3.5 implies, pointwise,

$$c_{\mathrm{anti}} e_k(a,p) \le |\mathsf{R}_{a,k}(p)| + C\sum_{i<k} e_i(a,p).$$

Integrating this bound over C_a, summing over a, and dividing by Q^k gives

$$E^{\mathrm{loc}}_k \le C|\nu^{\mathrm{loc}}_k|(\mathcal{P}) + C\sum_{i<k}\mathrm{PredErr}_{\theta,i}(\Pi).$$

Write

$$\nu^{\mathsf{D}_\theta,\Pi}_{g_{\mathrm{all}},k} = \nu^{\mathrm{loc}}_k + \nu^{\mathrm{out}}_k,$$

where ν^{out}_k is the contribution from $p \notin C_a$. Lemma 3.5 gives

$$|\nu^{\mathrm{out}}_k|(\mathcal{P}) \le CE^{\mathrm{out}}_k + C\sum_{i<k}\mathrm{PredErr}_{\theta,i}(\Pi).$$

Here the total variation of the outside contribution is bounded by the integral of $|\mathsf{R}_{a,k}(p)|$ over $p \notin C_a$; the boxes C_a are disjoint, but their complements need not be. The upper inequality in Lemma 3.5 then bounds this integral by E^{out}_k and the prefix errors. Since

$$|\nu^{\mathsf{D}_\theta,\Pi}_{g_{\mathrm{all}},k}|(\mathcal{P}) \le \mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi;\mathcal{G}_Q),$$

the measure decomposition also gives

$$|\nu^{\mathrm{loc}}_k|(\mathcal{P}) \le |\nu^{\mathsf{D}_\theta,\Pi}_{g_{\mathrm{all}},k}|(\mathcal{P}) + |\nu^{\mathrm{out}}_k|(\mathcal{P}).$$

Combining these bounds gives

$$E^{\mathrm{loc}}_k \le C\,\mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi;\mathcal{G}_Q) + CE^{\mathrm{out}}_k + C\sum_{i<k}\mathrm{PredErr}_{\theta,i}(\Pi).$$

Using (2) and (3),

$$\mathrm{PredErr}_{\theta,k}(\Pi) \le C\log(Q+1)\,\mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi;\mathcal{G}_Q) + C\alpha\,\mathrm{PredErr}_{\theta,k}(\Pi).$$

Absorbing the final term into the left-hand side gives

$$\mathrm{PredErr}_{\theta,k}(\Pi) \le C\log(Q+1)\,\mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi;\mathcal{G}_Q).$$

Together with (2), this proves the proposition. □

3.7 Recovering the Sign Assignment

Proposition 3.7 converts small Γ-ECE into small average prediction error. Distinct sign assignments produce different true property vectors on a constant fraction of the contexts, so sufficiently small prediction error identifies the true assignment. The KL bound in Assumption 3.1(v) limits how much information n samples contain about that assignment, and Fano's inequality then gives the required lower bound on sample complexity.

Lemma 3.8 (Different Sign Assignments Produce Separated Property Vectors)**.** *For every distinct* $\theta, \theta' \in \Theta_Q$,

$$\frac{1}{Q^k} \sum_{a \in \mathcal{X}_Q} \|v_\theta(a) - v_{\theta'}(a)\|_1 \geq 2\rho_{\text{pack}}\eta.$$

Proof. For every context a, the vectors $v_\theta(a)$ and $v_{\theta'}(a)$ agree in their first $k-1$ coordinates. Whenever $\theta_a \neq \theta'_a$, their final coordinates satisfy

$$|v_{\theta,k}(a) - v_{\theta',k}(a)| = 2\eta.$$

The packing condition gives at least $\rho_{\text{pack}} Q^k$ such contexts. □

Lemma 3.9 (Prediction Error Implies Exact Decoding)**.** *There exists a decoder* $\widehat{\theta}(\Pi)$ *such that, if*

$$\text{PredErr}_\theta(\Pi) \leq \frac{\rho_{\text{pack}}\eta}{2},$$

then

$$\widehat{\theta}(\Pi) = \theta.$$

Proof. For each a, define the mean prediction value

$$\bar{p}_\Pi(a) := \int_{\mathcal{P}} p \, \Pi_a(dp).$$

By Jensen's inequality,

$$\frac{1}{Q^k} \sum_a \|\bar{p}_\Pi(a) - v_\theta(a)\|_1 \leq \text{PredErr}_\theta(\Pi).$$

Decode by nearest neighbor:

$$\widehat{\theta}(\Pi) \in \arg\min_{\theta' \in \Theta_Q} \frac{1}{Q^k} \sum_a \|\bar{p}_\Pi(a) - v_{\theta'}(a)\|_1.$$

If $\theta' \neq \theta$, then Lemma 3.8 and the triangle inequality imply

$$\frac{1}{Q^k} \sum_a \|\bar{p}_\Pi(a) - v_{\theta'}(a)\|_1 \geq 2\rho_{\text{pack}}\eta - \frac{\rho_{\text{pack}}\eta}{2} > \frac{\rho_{\text{pack}}\eta}{2} \geq \frac{1}{Q^k} \sum_a \|\bar{p}_\Pi(a) - v_\theta(a)\|_1.$$

Thus the unique nearest neighbor is θ. □

Lemma 3.10 (Pairwise KL Bound)**.** *For every* $\theta, \theta' \in \Theta_Q$ *and every sample size* n,

$$D_{\text{KL}}(\mathsf{D}_\theta^n \,\|\, \mathsf{D}_{\theta'}^n) \leq 4C_{\text{KL}}\, n\eta^2.$$

Proof. Since X is uniform on $\mathcal{X}_Q$ under both distributions,

$$D_{\mathrm{KL}}(\mathsf{D}_\theta \,\|\, \mathsf{D}_{\theta'}) = \frac{1}{Q^k} \sum_a D_{\mathrm{KL}}\left(\mu_{v_\theta(a)} \,\middle\|\, \mu_{v_{\theta'}(a)}\right).$$

By Assumption 3.1(v),

$$D_{\mathrm{KL}}\left(\mu_{v_\theta(a)} \,\middle\|\, \mu_{v_{\theta'}(a)}\right) \le C_{\mathrm{KL}} \|v_\theta(a) - v_{\theta'}(a)\|_2^2 \le 4C_{\mathrm{KL}}\eta^2.$$

Tensorization over n i.i.d. samples gives the claim. □

Combining Proposition 3.7, Lemma 3.9, and Lemma 3.10 gives a lower bound on sample complexity at a fixed resolution.

Proposition 3.11 (Lower Bound at Resolution Q)**.** *There are constants $c_*, C_* > 0$, depending only on k, $\mathcal{R}_0$, and the constants in Assumption 3.1, such that the following holds for all sufficiently large powers of two Q. If a possibly randomized learner, given n i.i.d. samples from D_θ, outputs a predictor Π satisfying*

$$\mathrm{MCErr}^{\Gamma}_{\mathsf{D}_\theta}(\Pi; \mathcal{G}_Q) \le c_* \frac{\eta}{\log(Q+1)}$$

with probability at least $2/3$ for every $\theta \in \Theta_Q$, then

$$n \ge C_* Q^{k+2}.$$

Proof. Choose $c_* > 0$ small enough that Proposition 3.7 implies

$$\mathrm{PredErr}_\theta(\Pi) \le \frac{\rho_{\mathrm{pack}}\eta}{2}$$

whenever the learner achieves the multicalibration error required in the proposition. By Lemma 3.9, the learner then induces an exact decoder for θ with success probability at least $2/3$.

Let Θ be uniformly distributed over Θ_Q, and let the sample be drawn from D_Θ^n. If the learner is randomized, include its independent random seed in the decoder's observation; this does not change the KL bound. Fano's inequality [10] gives

$$\mathbb{P}[\widehat{\Theta} \ne \Theta] \ge 1 - \frac{I(\Theta; \mathrm{sample}) + \log 2}{\log |\Theta_Q|}.$$

The standard bound on mutual information in terms of pairwise KL divergence, together with Lemma 3.10, implies

$$I(\Theta; \mathrm{sample}) \le 4C_{\mathrm{KL}} n \eta^2.$$

A decoder with error probability at most $1/3$ therefore requires

$$4C_{\mathrm{KL}} n\eta^2 + \log 2 \ge \frac{2}{3} \log |\Theta_Q| \ge cQ^k$$

for a constant $c > 0$. For all sufficiently large Q, the $\log 2$ term is absorbed into the right-hand side, so

$$4C_{\mathrm{KL}} n\eta^2 \ge cQ^k$$

after decreasing c. Since $\eta \asymp Q^{-1}$,

$$n \ge C_* Q^{k+2}.$$

□

3.8 Proof of the Main Theorem

Choosing Q as a function of the target error ε converts Proposition 3.11 into Theorem 3.2.

Proof of Theorem 3.2. For each large power of two Q, Proposition 3.11 gives a finite collection of candidate distributions for which $n \geq C_* Q^{k+2}$ samples are required at error level

$$\varepsilon_Q := c_* \frac{\eta}{\log(Q+1)} \asymp \frac{1}{Q \log Q}.$$

Given sufficiently small $\varepsilon > 0$, choose Q to be the largest power of two such that

$$\varepsilon \leq \varepsilon_Q.$$

Then

$$Q \geq c' \frac{1}{\varepsilon \log(C'/\varepsilon)}$$

for constants $c', C' > 0$. Therefore

$$n \geq C_* Q^{k+2} \geq c \frac{\varepsilon^{-(k+2)}}{\log^{k+2}(C/\varepsilon)}$$

for constants $c, C > 0$ depending only on k, κ, $\mathcal{R}_0$, and the constants in Assumption 3.1. The group family satisfies

$$|\mathcal{G}_Q| \leq C_{\alpha,k} \log^3(Q+1) \leq \varepsilon^{-\kappa}$$

for every fixed $\kappa > 0$ and all sufficiently small ε. Taking

$$\mathcal{X}_\varepsilon := \mathcal{X}_Q, \qquad \mathcal{G}_\varepsilon := \mathcal{G}_Q, \qquad \mathfrak{D}_\varepsilon := \{\mathsf{D}_\theta : \theta \in \Theta_Q\}$$

proves the theorem. □

4 The Upper Bound

Our route to the upper bound is through an online forecasting problem. On each round, the forecaster constructs a prediction rule from the preceding rounds, applies it to the current context, and then observes the outcome. We first control the empirical multicalibration error accumulated over these rounds.

Given a dataset consisting of T i.i.d. context–outcome pairs, we run the online procedure for T rounds, using one pair on each round. The context component of the t-th pair is revealed before the prediction and its outcome component afterward. The procedure produces T prediction rules. We return their average as our predictor. A martingale argument then transfers the empirical guarantee for the online transcript to a population multicalibration guarantee for this predictor.

We begin by formulating the online problem and expressing its empirical multicalibration error as a collection of objectives that the forecaster must control simultaneously. We then state the regularity conditions, construct and analyze the forecaster, and carry out the online-to-batch reduction. Finally, we show how to implement the forecaster efficiently using linear optimization over $\mathcal{M}$.

4.1 The Online Formulation and Its Objectives

We first describe the online formulation for an arbitrary finite set $\mathcal{P}_Q \subset \mathcal{P}$ of possible predictions. Fix a horizon T and a finite group family $\mathcal{G}$.

Let $\mathcal{F}_{t-1}$ be the sigma-field generated by the observations through round $t-1$ and any internal randomness used by the forecaster before round t. At the beginning of round t, the forecaster uses this history to define an $\mathcal{F}_{t-1}$-measurable rule $\pi_t : \mathcal{X} \to \Delta(\mathcal{P}_Q)$. After the context X_t is revealed, it outputs $\pi_t(\cdot \mid X_t)$ as the randomized prediction and then observes the outcome Y_t.

For any sequence of such rules, define the empirical multicalibration error over the T online rounds by

$$\widehat{\mathrm{MCErr}}_T^{\,\Gamma} := \frac{1}{T} \max_{g \in \mathcal{G}} \sum_{p \in \mathcal{P}_Q} \left\| \sum_{t=1}^{T} g(X_t)\pi_t(p \mid X_t) R(p, Y_t) \right\|_1 .$$

Let $\Sigma := \{\pm 1\}^{\mathcal{P}_Q \times [k]}$, and write $s_{p,j}$ for the sign assigned by $s \in \Sigma$ to the pair (p, j). Representing each absolute value in the ℓ_1-norm as a maximum over its sign gives

$$\begin{aligned} \widehat{\mathrm{MCErr}}_T^{\,\Gamma} &= \frac{1}{T} \max_{\substack{g \in \mathcal{G} \\ s \in \Sigma}} \sum_{p \in \mathcal{P}_Q} \sum_{j=1}^{k} s_{p,j} \sum_{t=1}^{T} g(X_t)\pi_t(p \mid X_t) R_j(p_{<j}, p_j, Y_t) \\ &= \frac{1}{T} \max_{\substack{g \in \mathcal{G} \\ s \in \Sigma}} \sum_{t=1}^{T} g(X_t) \mathbb{E}_{p \sim \pi_t(\cdot \mid X_t)} \left[\sum_{j=1}^{k} s_{p,j} R_j(p_{<j}, p_j, Y_t) \right] . \end{aligned}$$

The above equalities express the empirical error as the largest cumulative signed residual among the objectives indexed by $(g, s) \in \mathcal{G} \times \Sigma$. Thus, the online task is a multiobjective learning problem: the forecaster must control all of these signed objectives simultaneously [34]. We use the framework of online learning with expert advice. Each pair (g, s) is treated as an expert, whose gain on round t is

$$h_t(g, s) := g(X_t) \mathbb{E}_{p \sim \pi_t(\cdot \mid X_t)} \left[\sum_{j=1}^{k} s_{p,j} R_j(p_{<j}, p_j, Y_t) \right] .$$

Since the empirical error is the largest cumulative gain of any expert divided by T, our goal is to control the cumulative gain of every expert. The exponential weights algorithm chooses a distribution w_t over the experts on each round. We call the expectation of $h_t(g, s)$ under this distribution the expected gain under w_t. Its regret guarantee bounds the cumulative gain of every individual expert by the sum of the expected gains under w_t, plus a sublinear regret term. It therefore remains to keep these expected gains small.

We use the same weights w_t to choose the forecaster's distribution on $\mathcal{P}_Q$. For any candidate distribution $\pi \in \Delta(\mathcal{P}_Q)$ and any admissible outcome law, we can evaluate the expected gain under w_t from the corresponding signed residuals. The forecaster chooses π in a minimax manner, minimizing the largest possible value of this expected gain over admissible outcome laws. We allow an additive error $\rho \geq 0$ in this minimization. Noarov et al. [38] use a related construction for high-dimensional unbiased prediction, in which an online learning algorithm weights signed objectives and those weights guide the randomized forecast.

4.2 Regularity Assumptions

The online formulation can be stated for any finite set $\mathcal{P}_Q$, but its analysis requires additional structure. Convexity, compactness, and continuity (Condition (i) below) allow us to apply Sion's

minimax theorem. Although the forecaster must choose a distribution on $\mathcal{P}_Q$ without knowing the conditional outcome law, the theorem allows us to analyze the one-round objective by first fixing that law and then choosing the distribution. A uniform bound on the residual (Condition (ii)) controls both the expert gains and the martingale deviations. Finally, Condition (iii) places two requirements on the grid. It has only $O(Q^k)$ points, and for every admissible outcome law μ, some $p \in \mathcal{P}_Q$ makes every coordinate of $R(p, \mu)$ small. After μ is fixed in the reversed analysis, choosing the point mass at such a p makes the expected gain under w_t small, which bounds the one-round minimax value. The following assumption makes these requirements precise.

Assumption 4.1 (Regularity for the Upper Bound)**.** *There are constants $R_{\max}, C_{\text{grid}} < \infty$. For each integer $Q \geq 1$, there is a finite grid $\mathcal{P}_Q \subset \mathcal{P}$ such that the following hold.*

(i) $\mathcal{M}$ is a convex compact subset of a topological vector space of finite signed measures on $\mathcal{Y}$, and for every $p \in \mathcal{P}_Q$ and coordinate $j \in [k]$, the map

$$\mu \mapsto R_j(p_{<j}, p_j, \mu)$$

is affine and continuous on $\mathcal{M}$.

(ii) For every $p \in \mathcal{P}$ and $y \in \mathcal{Y}$,

$$\|R(p, y)\|_\infty \leq R_{\max}.$$

(iii) The grid size satisfies

$$|\mathcal{P}_Q| \leq C_{\text{grid}} Q^k,$$

and the expected residual can be rounded at rate $1/Q$:

$$\Delta_Q := \sup_{\mu \in \mathcal{M}} \min_{p \in \mathcal{P}_Q} \|R(p, \mu)\|_\infty \leq \frac{C_{\text{grid}}}{Q}.$$

Remark 4.2 (A Sufficient Lipschitz Condition)**.** *A convenient way to ensure Assumption 4.1(iii) is to require a constant L_R such that*

$$\|R(p, \mu) - R(p', \mu)\|_\infty \leq L_R \|p - p'\|_\infty \qquad \forall \mu \in \mathcal{M}, \quad p, p' \in \mathcal{P},$$

together with a finite grid $\mathcal{P}_Q \subset \mathcal{P}$ whose ℓ_∞-mesh is $O(1/Q)$. Because the residual vanishes at the true property, every $\Gamma(\mu)$ can then be rounded to some $p \in \mathcal{P}_Q$ with $\|R(p, \mu)\|_\infty = O(1/Q)$.

4.3 The Online Forecaster and Its Empirical Guarantee

Fix an integer $Q \geq 1$, and let $\mathcal{P}_Q$ be the grid supplied by Assumption 4.1. Algorithm 1 specifies how the forecaster operates on each round.

Algorithm 1 Online Forecaster

Input: Horizon T, group family $\mathcal{G}$, grid $\mathcal{P}_Q$, law class $\mathcal{M}$, level count k, additive optimization error ρ

Set

$$\eta_{\mathrm{EW}} \leftarrow \frac{\sqrt{2\log|\mathcal{G}\times\Sigma|}}{kR_{\max}\sqrt{T}}.$$

Initialize cumulative gains $C_0(g,s) \leftarrow 0$ for all $(g,s) \in \mathcal{G}\times\Sigma$.

for $t = 1,\dots,T$ **do**

Compute

$$w_t(g,s) \leftarrow \frac{\exp\{\eta_{\mathrm{EW}} C_{t-1}(g,s)\}}{\sum_{(g',s')\in\mathcal{G}\times\Sigma}\exp\{\eta_{\mathrm{EW}} C_{t-1}(g',s')\}}.$$

Define $\pi_t(\cdot \mid x)$ pointwise for $x \in \mathcal{X}$ to satisfy

$$\begin{aligned}\max_{\mu\in\mathcal{M}} \mathbb{E}_{p\sim\pi_t(\cdot|x)}\left[\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma} w_t(g,s)g(x)\sum_{j=1}^{k} s_{p,j}R_j(p_{<j},p_j,\mu)\right] \\ \le \min_{\pi\in\Delta(\mathcal{P}_Q)}\max_{\mu\in\mathcal{M}}\mathbb{E}_{p\sim\pi}\left[\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma} w_t(g,s)g(x)\sum_{j=1}^{k} s_{p,j}R_j(p_{<j},p_j,\mu)\right] + \rho.\end{aligned} \tag{4}$$

Observe X_t and output $\pi_t(\cdot \mid X_t)$ as the randomized prediction.

Observe Y_t.

For every $(g,s) \in \mathcal{G}\times\Sigma$, set

$$h_t(g,s) \leftarrow g(X_t)\mathbb{E}_{p\sim\pi_t(\cdot|X_t)}\left[\sum_{j=1}^{k} s_{p,j}R_j(p_{<j},p_j,Y_t)\right].$$

Set $C_t(g,s) \leftarrow C_{t-1}(g,s) + h_t(g,s)$ for every $(g,s) \in \mathcal{G}\times\Sigma$.

end for

Output: Prediction rules $\pi_1,\dots,\pi_T$.

The prediction rules in Algorithm 1 can be chosen jointly measurable, as shown in Appendix B.

The forecaster chooses π_t to keep the expected gain under w_t small on each round, while the no-regret guarantee of the exponential weights algorithm bounds the cumulative gain of every expert by the sum of these expected gains plus a sublinear regret term. Since the empirical multicalibration error is the largest cumulative expert gain divided by T, this gives the following guarantee.

Theorem 4.3 (Empirical Guarantee for the Online Forecaster)**.** *Suppose Assumption 4.1 holds. Let $(X_t,Y_t)_{t=1}^T$ be any stochastic process such that, conditionally on $\mathcal{F}_{t-1}$ and X_t, the conditional law of Y_t belongs to $\mathcal{M}$. Then Algorithm 1 satisfies*

$$\widehat{\mathbb{E}\mathrm{MCErr}}_T^{\Gamma} \le \rho + k\Delta_Q + C_k R_{\max}\sqrt{\frac{\log|\mathcal{G}| + k|\mathcal{P}_Q|}{T}},$$

where C_k depends only on k. Consequently, the two bounds in Assumption 4.1(iii) give

$$\widehat{\mathbb{E}\mathrm{MCErr}}_T^{\Gamma} \le \rho + O\left(\frac{1}{Q} + R_{\max}\sqrt{\frac{Q^k + \log|\mathcal{G}|}{T}}\right). \tag{5}$$

The expectation is over the stochastic transcript and any internal randomization of the forecaster.

We prove Theorem 4.3 in the next subsection.

4.4 Analysis of the Online Forecaster

Throughout the analysis, π_t denotes the $\mathcal{F}_{t-1}$-measurable rule selected at round t. Since every group function takes values in $[0,1]$ and the residual bound in Assumption 4.1(ii) gives $\|R\|_\infty \le R_{\max}$, every signed gain satisfies

$$|h_t(g,s)| \le kR_{\max}.$$

By the signed-objective representation of the empirical multicalibration error,

$$\widehat{\mathrm{MCErr}}_T^{\Gamma} = \frac{1}{T}\max_{(g,s)\in\mathcal{G}\times\Sigma}\sum_{t=1}^{T} h_t(g,s).$$

Thus controlling the gains of all experts $(g,s)\in\mathcal{G}\times\Sigma$ controls empirical Γ-ECE.

4.4.1 Exponential Weights

The first step compares the largest cumulative signed residual with the sum of the expected gains under the distributions w_t produced by exponential weights.

Lemma 4.4 (Exponential Weights Bound)**.** *For every realized transcript of Algorithm 1,*

$$\max_{(g,s)\in\mathcal{G}\times\Sigma}\sum_{t=1}^{T} h_t(g,s) \le \sum_{t=1}^{T}\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma} w_t(g,s)h_t(g,s) + C_k R_{\max}\sqrt{T\big(\log|\mathcal{G}| + k|\mathcal{P}_Q|\big)},$$

where C_k depends only on k and $h_t(g,s)$ is the signed gain from Algorithm 1,

$$h_t(g,s) = g(X_t)\mathbb{E}_{p\sim\pi_t(\cdot|X_t)}\left[\sum_{j=1}^{k} s_{p,j}R_j(p_{<j},p_j,Y_t)\right].$$

Proof. Using the learning rate η_{EW} from Algorithm 1, define the exponential weights potential

$$W_t := \sum_{g\in\mathcal{G}}\sum_{s\in\Sigma}\exp\{\eta_{\mathrm{EW}}C_{t-1}(g,s)\}.$$

Since $C_t(g,s) = C_{t-1}(g,s) + h_t(g,s)$, and since $|h_t(g,s)| \le kR_{\max}$, Hoeffding's lemma [30] gives

$$\log\frac{W_{t+1}}{W_t} = \log\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma} w_t(g,s)\exp\{\eta_{\mathrm{EW}}h_t(g,s)\} \le \eta_{\mathrm{EW}}\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma} w_t(g,s)h_t(g,s) + \frac{\eta_{\mathrm{EW}}^2k^2R_{\max}^2}{2}.$$

Summing over t yields

$$\log W_{T+1} \le \log|\mathcal{G}\times\Sigma| + \eta_{\mathrm{EW}}\sum_{t=1}^{T}\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma} w_t(g,s)h_t(g,s) + \frac{\eta_{\mathrm{EW}}^2k^2R_{\max}^2T}{2}.$$

On the other hand, for every $(g,s)\in\mathcal{G}\times\Sigma$,

$$\log W_{T+1} \ge \eta_{\mathrm{EW}}C_T(g,s) = \eta_{\mathrm{EW}}\sum_{t=1}^{T} h_t(g,s).$$

Combining the two inequalities and maximizing over (g, s) gives

$$\max_{(g,s)\in\mathcal{G}\times\Sigma} \sum_{t=1}^{T} h_t(g,s) \le \sum_{t=1}^{T} \sum_{g\in\mathcal{G}} \sum_{s\in\Sigma} w_t(g,s) h_t(g,s) + \frac{\log|\mathcal{G}\times\Sigma|}{\eta_{\mathrm{EW}}} + \frac{\eta_{\mathrm{EW}} k^2 R_{\max}^2 T}{2}.$$

The expert class satisfies

$$\log|\mathcal{G}\times\Sigma| = \log|\mathcal{G}| + k|\mathcal{P}_Q|\log 2.$$

Substituting this expression and the value of η_{EW} from Algorithm 1 into the regret terms completes the proof. □

4.4.2 The Minimax Value

Conditional on the history and the current context, the expected gain under w_t is obtained by evaluating the objective in (4) at the conditional law of the outcome. By the choice of π_t, it is bounded by the worst-case value of the one-round minimax problem, up to the additive accuracy ρ. We next bound this value.

Lemma 4.5 (Upper Bound on the Minimax Value)**.** *Under Assumption 4.1, the minimax term on the right-hand side of* (4) *is at most $k\Delta_Q$ for every history and every context x.*

Proof. Fix the history and the context x, and let $F(\pi, \mu)$ denote the objective in (4). We first bound the value when the outcome law is chosen before the prediction:

$$\max_{\mu\in\mathcal{M}} \min_{\pi\in\Delta(\mathcal{P}_Q)} F(\pi,\mu).$$

Fix $\mu \in \mathcal{M}$. Assumption 4.1(iii) provides $p_\mu \in \mathcal{P}_Q$ such that

$$\|R(p_\mu,\mu)\|_\infty \le \Delta_Q.$$

Choose the point mass $\pi = \delta_{p_\mu}$ at p_μ. For every (g, s), the bound $0 \le g(x) \le 1$ gives

$$g(x)\sum_{j=1}^{k} s_{p_\mu,j} R_j((p_\mu)_{<j},(p_\mu)_j,\mu) \le \sum_{j=1}^{k} |R_j((p_\mu)_{<j},(p_\mu)_j,\mu)| \le k\Delta_Q.$$

Taking the expectation under w_t preserves the bound. Since the argument holds for every μ,

$$\max_{\mu\in\mathcal{M}} \min_{\pi\in\Delta(\mathcal{P}_Q)} F(\pi,\mu) \le k\Delta_Q.$$

Since F is affine in π, Assumption 4.1(i) allows us to apply Sion's minimax theorem [45]. It identifies this value with the value in which the prediction is chosen first:

$$\min_{\pi}\max_{\mu} F(\pi,\mu) = \max_{\mu}\min_{\pi} F(\pi,\mu) \le k\Delta_Q.$$

□

The exponential weights lemma bounds the largest total gain of any expert by the sum of the expected gains under w_t, together with a regret term. Lemma 4.5 bounds each round's expected gain after conditioning on the history and the current context. We now combine the two bounds.

Proof of Theorem 4.3. Lemma 4.4 gives

$$\mathbb{E}\left[\max_{(g,s)\in\mathcal{G}\times\Sigma}\sum_{t=1}^{T}h_t(g,s)\right]\le\mathbb{E}\left[\sum_{t=1}^{T}\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma}w_t(g,s)h_t(g,s)\right]+C_kR_{\max}\sqrt{T\big(\log|\mathcal{G}|+k|\mathcal{P}_Q|\big)}.$$

Condition on $\mathcal{F}_{t-1}$ and X_t, and let μ_t be the conditional law of Y_t. By the hypothesis of Theorem 4.3, $\mu_t\in\mathcal{M}$. The conditional expected gain under w_t is the objective in (4) evaluated at $\pi_t(\cdot\mid X_t)$ and μ_t. It is therefore at most the worst-case value for $\pi_t(\cdot\mid X_t)$. The choice of π_t bounds this worst-case value by the minimax value plus ρ, and Lemma 4.5 bounds the minimax value by $k\Delta_Q$. Hence

$$\mathbb{E}\left[\sum_{g\in\mathcal{G}}\sum_{s\in\Sigma}w_t(g,s)h_t(g,s)\,\middle|\,\mathcal{F}_{t-1},X_t\right]\le k\Delta_Q+\rho.$$

Summing over t and taking expectations bounds the cumulative expected gain by $T(k\Delta_Q+\rho)$. Dividing the exponential weights bound by T and using the representation at the beginning of this subsection proves the first bound in Theorem 4.3. The second follows from the two bounds in Assumption 4.1(iii). □

4.5 The Averaged Batch Predictor and Sample Complexity

Let D be an $\mathcal{M}$-compatible distribution, and let $S=((X_1,Y_1),\dots,(X_T,Y_T))\sim\mathsf{D}^T$. We run Algorithm 1 on the sample sequence and return the averaged predictor $\Pi_S:\mathcal{X}\to\Delta(\mathcal{P}_Q)$ defined by

$$\Pi_S(p\mid x):=\frac{1}{T}\sum_{t=1}^{T}\pi_t(p\mid x).\tag{6}$$

4.5.1 Online-to-Batch Reduction

The population multicalibration error of the averaged predictor need not equal the empirical multicalibration error of the online transcript. After expressing both errors as maxima over groups and sign arrays, we compare the corresponding population and empirical objectives. As in Collina et al. [9], for each fixed group and sign array, the difference between these objectives is a martingale sum. The following maximal inequality controls all these sums simultaneously.

Lemma 4.6 (A Maximal Inequality for Finitely Many Martingales)**.** *Let $\mathcal{A}$ be a finite set of size N. For each $e\in\mathcal{A}$, let $(M_t(e))_{t=1}^T$ be martingale differences with respect to the same filtration, and assume $|M_t(e)|\le G$ almost surely for all t,e. Then*

$$\mathbb{E}\max_{e\in\mathcal{A}}\sum_{t=1}^{T}M_t(e)\le G\sqrt{2T\log N}.$$

Proof. For any $\eta>0$, conditional Hoeffding gives

$$\mathbb{E}\left[\exp\left(\eta\sum_{t=1}^{T}M_t(e)\right)\right]\le\exp\left(\frac{\eta^2G^2T}{2}\right).$$

Therefore

$$\mathbb{E}\max_{e}\sum_{t}M_t(e)\le\frac{1}{\eta}\log\mathbb{E}\sum_{e}\exp\left(\eta\sum_{t}M_t(e)\right)\le\frac{\log N}{\eta}+\frac{\eta G^2T}{2}.$$

Optimizing over η proves the bound. □

We now apply Lemma 4.6 to compare the population multicalibration error of the averaged predictor with the empirical multicalibration error of the online transcript.

Lemma 4.7 (Online-to-Batch Transfer)**.** *Under Assumption 4.1, for the averaged predictor* Π_S *in* (6)*,*

$$\mathbb{E}\left[\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi_S;\mathcal{G})\right] \le \mathbb{E}\widehat{\mathrm{MCErr}}_T^{\Gamma} + C_k R_{\max}\sqrt{\frac{Q^k + \log|\mathcal{G}|}{T}},$$

where C_k *depends only on* k*. Both expectations are over the sample and any internal randomization used by the forecaster.*

Proof. For $(g,s) \in \mathcal{G}\times\Sigma$, set

$$\mathcal{L}_{g,s}(\Pi_S) := \sum_{p\in\mathcal{P}_Q}\sum_{j=1}^{k} s_{p,j}\nu_{g,j}^{\mathsf{D},\Pi_S}(\{p\}).$$

Maximizing over each $s_{p,j} \in \{\pm 1\}$ gives

$$\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi_S;\mathcal{G}) = \max_{g\in\mathcal{G}}\max_{s\in\Sigma}\mathcal{L}_{g,s}(\Pi_S).$$

Its empirical counterpart is

$$\widehat{\mathcal{L}}_{g,s}(S) := \frac{1}{T}\sum_{t=1}^{T} g(X_t)\sum_{p\in\mathcal{P}_Q}\pi_t(p\mid X_t)\sum_{j=1}^{k} s_{p,j}R_j(p_{<j},p_j,Y_t).$$

Then

$$\widehat{\mathrm{MCErr}}_T^{\Gamma} = \max_{g,s}\widehat{\mathcal{L}}_{g,s}(S).$$

Therefore

$$\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi_S;\mathcal{G}) \le \widehat{\mathrm{MCErr}}_T^{\Gamma} + \max_{g,s}\left(\mathcal{L}_{g,s}(\Pi_S) - \widehat{\mathcal{L}}_{g,s}(S)\right).$$

Fix (g,s). For each t, let

$$Z_t^{g,s} := g(X_t)\sum_{p\in\mathcal{P}_Q}\pi_t(p\mid X_t)\sum_{j=1}^{k} s_{p,j}R_j(p_{<j},p_j,Y_t).$$

Define

$$\bar{Z}_t^{g,s} := \mathbb{E}_{(X,Y)\sim\mathsf{D}}\left[g(X)\sum_{p\in\mathcal{P}_Q}\pi_t(p\mid X)\sum_{j=1}^{k} s_{p,j}R_j(p_{<j},p_j,Y)\middle|\mathcal{F}_{t-1}\right].$$

Conditionally on $\mathcal{F}_{t-1}$, the rule π_t is fixed. The linearity of $\mathcal{L}_{g,s}$ in the predictor and the definition of Π_S therefore give

$$\mathcal{L}_{g,s}(\Pi_S) = \frac{1}{T}\sum_{t=1}^{T}\bar{Z}_t^{g,s}.$$

The empirical quantity satisfies

$$\widehat{\mathcal{L}}_{g,s}(S) = \frac{1}{T}\sum_{t=1}^{T} Z_t^{g,s}.$$

Because the sample is i.i.d., $\bar{Z}_t^{g,s}$ is the conditional expectation of $Z_t^{g,s}$ given $\mathcal{F}_{t-1}$. Hence the differences

$$M_t^{g,s} := \bar{Z}_t^{g,s} - Z_t^{g,s}$$

are martingale differences with respect to the filtration generated by the transcript and the forecaster's internal randomization. Since group functions take values in $[0,1]$, we have $|Z_t^{g,s}| \le kR_{\max}$, and hence $|M_t^{g,s}| \le 2kR_{\max}$. Subtracting the two equalities gives

$$\mathcal{L}_{g,s}(\Pi_S) - \widehat{\mathcal{L}}_{g,s}(S) = \frac{1}{T}\sum_{t=1}^{T} M_t^{g,s}.$$

Lemma 4.6, applied to the $|\mathcal{G}||\Sigma| = |\mathcal{G}|2^{k|\mathcal{P}_Q|}$ martingale-difference sequences, gives

$$\mathbb{E}\max_{g,s}\sum_{t=1}^{T} M_t^{g,s} \le C_k R_{\max}\sqrt{T\big(\log|\mathcal{G}| + k|\mathcal{P}_Q|\big)}.$$

Dividing by T proves the lemma. □

4.5.2 Sample Complexity

Combining the empirical guarantee for the online forecaster with the online-to-batch transfer gives the sample-complexity bound.

Theorem 4.8 (Sample Complexity under Approximate Minimax Optimization)**.** *Suppose Assumption 4.1 holds. Let* D *be any* $\mathcal{M}$*-compatible distribution on* $\mathcal{X}\times\mathcal{Y}$*, and let* $\mathcal{G}$ *be any finite group family. Fix a sufficiently small target error* $\varepsilon > 0$ *and a minimax accuracy*

$$0 \le \rho < \frac{\varepsilon}{3}.$$

There is a randomized learner using minimax accuracy ρ *and a grid with*

$$Q = \Theta\big((\varepsilon - 3\rho)^{-1}\big)$$

which, from

$$n \ge C\left((\varepsilon-3\rho)^{-(k+2)} + (\varepsilon-3\rho)^{-2}\log|\mathcal{G}|\right)$$

i.i.d. samples from D*, outputs a finite-support predictor* $\Pi_S : \mathcal{X} \to \Delta(\mathcal{P}_Q)$ *satisfying*

$$\mathbb{P}\left[\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi_S;\mathcal{G}) \le \varepsilon\right] \ge \frac{2}{3}.$$

The probability is over the sample and any internal randomization of the learner, and C *depends only on* k*,* $R_{\max}$*, and* C_{grid}*.*

Proof of Theorem 4.8. Run the online forecaster for $T = n$ rounds. Because the sample is i.i.d., the conditional law of Y_t given $\mathcal{F}_{t-1}$ and X_t is $\mathsf{D}_{Y|X=X_t}$. The $\mathcal{M}$-compatibility of D therefore ensures that the hypothesis of Theorem 4.3 holds. Theorem 4.3 and Lemma 4.7 now give

$$\mathbb{E}\,\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi_S;\mathcal{G}) \le \rho + C'\left(\frac{1}{Q} + R_{\max}\sqrt{\frac{Q^k + \log|\mathcal{G}|}{n}}\right).$$

Because $\varepsilon - 3\rho > 0$, take

$$Q = \left\lceil \frac{6C'}{\varepsilon - 3\rho} \right\rceil.$$

If

$$n \geq C\left((\varepsilon - 3\rho)^{-(k+2)} + (\varepsilon - 3\rho)^{-2} \log |\mathcal{G}|\right)$$

for a sufficiently large constant C, then the expectation is at most

$$\rho + \frac{\varepsilon - 3\rho}{3} = \frac{\varepsilon}{3}.$$

Markov's inequality therefore gives

$$\mathbb{P}\left[\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi_S; \mathcal{G}) > \varepsilon\right] \leq \frac{1}{3}.$$

□

If $\rho \leq \varepsilon/6$, then $\varepsilon - 3\rho \geq \varepsilon/2$, so the theorem immediately gives the following bound.

Corollary 4.9 (Upper Bound on Sample Complexity)**.** *Under the conditions of Theorem 4.8, suppose each inner minimax problem is solved to additive accuracy*

$$\rho \leq \frac{\varepsilon}{6}.$$

Then there is a randomized learner using $Q = \Theta(1/\varepsilon)$ and

$$n \geq C\left(\varepsilon^{-(k+2)} + \varepsilon^{-2} \log |\mathcal{G}|\right)$$

i.i.d. samples from D whose output satisfies

$$\mathbb{P}\left[\mathrm{MCErr}_{\mathsf{D}}^{\Gamma}(\Pi_S; \mathcal{G}) \leq \varepsilon\right] \geq \frac{2}{3}.$$

In particular, if $|\mathcal{G}| \leq \lceil \varepsilon^{-\kappa} \rceil$ for a fixed $\kappa > 0$, then

$$n = O(\varepsilon^{-(k+2)}).$$

Consequently,

$$\mathrm{SC}_{\Gamma,\mathcal{M}}^{(\kappa)}(\varepsilon) = O(\varepsilon^{-(k+2)}).$$

Binary groups are included among the group functions allowed in the minimax definition. Thus, if Assumption 3.1 also holds, Theorem 3.2 and Corollary 4.9 give

$$\mathrm{SC}_{\Gamma,\mathcal{M}}^{(\kappa)}(\varepsilon) = \widetilde{\Theta}\left(\varepsilon^{-(k+2)}\right).$$

4.6 Polynomial-Time Implementation

The exponential weights procedure analyzed above is information-theoretic and maintains weights over $|\mathcal{G}|2^{k|\mathcal{P}_Q|}$ experts indexed by sign patterns. This section gives a polynomial-time implementation under an explicit linear optimization oracle over $\mathcal{M}$.

All running-time statements in this paper use a unit-cost real-arithmetic model. Arithmetic operations, comparisons, evaluations of the elementary functions used in the implementation, and evaluations of the supplied group and residual functions are exact and have unit cost. The running time also includes the operations performed by the linear optimization oracle, as specified in Assumption 4.10.

There are two computational tasks. We first compute the exponential weights distribution without enumerating all sign arrays. We then solve the minimax problem using linear optimization over $\mathcal{M}$.

4.6.1 Computing Exponential Weights Implicitly

Let

$$\mathrm{CumRes}_{g,p,j}^{t-1} := \sum_{\tau<t} g(X_\tau)\pi_\tau(p \mid X_\tau)R_j(p_{<j}, p_j, Y_\tau).$$

The definition of the gain gives

$$C_{t-1}(g,s) = \sum_{\tau<t} h_\tau(g,s) = \sum_{p\in\mathcal{P}_Q}\sum_{j=1}^{k} s_{p,j}\,\mathrm{CumRes}_{g,p,j}^{t-1}\,.$$

Hence the exponential weights distribution over $(g,s) \in \mathcal{G}\times\Sigma$, with learning rate $\eta_{\mathrm{EW}} > 0$, has weight proportional to

$$\exp\left(\eta_{\mathrm{EW}}\sum_{p\in\mathcal{P}_Q}\sum_{j=1}^{k} s_{p,j}\,\mathrm{CumRes}_{g,p,j}^{t-1}\right).$$

For fixed g, the exponent is a sum of terms that each depend on only one sign coordinate, so the conditional distribution of the signs factorizes. Using

$$\sum_{\sigma\in\{\pm1\}} e^{\eta\sigma a} = 2\cosh(\eta a),$$

the group marginal is

$$\omega_t(g) = \frac{\prod_{p,j} 2\cosh\bigl(\eta_{\mathrm{EW}}\,\mathrm{CumRes}_{g,p,j}^{t-1}\bigr)}{\sum_{g'\in\mathcal{G}}\prod_{p,j} 2\cosh\bigl(\eta_{\mathrm{EW}}\,\mathrm{CumRes}_{g',p,j}^{t-1}\bigr)},$$

and, conditional on g, each sign has mean

$$m_{t,g,p,j} := \mathbb{E}[s_{p,j} \mid g] = \tanh\bigl(\eta_{\mathrm{EW}}\,\mathrm{CumRes}_{g,p,j}^{t-1}\bigr).$$

Conditional on g, the objective in (4) is linear in the signs, so its average over the signs depends only on the means $m_{t,g,p,j}$. Averaging also over the group marginal ω_t, the objective becomes

$$\sum_{p\in\mathcal{P}_Q}\pi_p\sum_{j=1}^{k} a_{t,p,j}(x)R_j(p_{<j}, p_j, \mu),$$

where

$$a_{t,p,j}(x) := \sum_{g\in\mathcal{G}}\omega_t(g)g(x)m_{t,g,p,j}.$$

Thus it suffices to maintain the $O(k|\mathcal{G}||\mathcal{P}_Q|)$ values $\mathrm{CumRes}_{g,p,j}^{t-1}$, rather than enumerate $2^{k|\mathcal{P}_Q|}$ sign patterns.

4.6.2 The Optimization Oracle

Once the coefficients $a_{t,p,j}(x)$ have been computed, it remains to solve the minimax problem. Noarov et al. [38] solve an analogous problem for high-dimensional unbiased prediction by expressing it as a finite-dimensional linear program with continuously many constraints and using an approximate separation oracle. Here the required approximate separation step reduces to linear optimization over $\mathcal{M}$.

Assumption 4.10 (Linear Optimization Oracle over $\mathcal{M}$)**.** *There is an oracle* $\mathsf{OPT}_{\mathcal{M}}$ *which, for every requested accuracy* $\xi \in (0,1)$ *and given coefficients* $c_{p,j} \in \mathbb{R}$*, returns a law* $\widehat{\mu} \in \mathcal{M}$ *satisfying*

$$\sum_{p\in\mathcal{P}_Q}\sum_{j=1}^{k} c_{p,j}R_j(p_{<j},p_j,\widehat{\mu}) \geq \sup_{\mu\in\mathcal{M}}\sum_{p\in\mathcal{P}_Q}\sum_{j=1}^{k} c_{p,j}R_j(p_{<j},p_j,\mu) \;-\xi$$

using a number of operations polynomial in $|\mathcal{P}_Q|$ *and* $\log(1/\xi)$*. Within the same bound, the objective value and the residual expectations at* $\widehat{\mu}$ *can be evaluated to additive accuracy* ξ*.*

For the computational implementation, fix one deterministic, Borel-measurable realization of this oracle, including deterministic tie-breaking whenever more than one admissible output is available.

Given the coefficients $a_{t,p,j}(x)$, the inner minimax problem is the semi-infinite linear program

$$\begin{aligned}\min_{\pi,\zeta}\quad & \zeta \\ \text{s.t.}\quad & \pi \in \Delta(\mathcal{P}_Q), \\ & \sum_{p,j}\pi_p a_{t,p,j}(x)R_j(p_{<j},p_j,\mu) \leq \zeta \qquad \forall \mu\in\mathcal{M}.\end{aligned}$$

For a candidate (π,ζ), a violated constraint is found by calling $\mathsf{OPT}_{\mathcal{M}}$ with coefficients

$$c_{p,j} = \pi_p a_{t,p,j}(x).$$

To see how the two optimization accuracies enter, call the oracle with internal accuracy ξ, let $\widehat{\mu}$ be the returned law, and evaluate its objective to the same accuracy. If the resulting estimate exceeds $\zeta+\xi$, then the constraint associated with $\widehat{\mu}$ is violated. Otherwise, the oracle and evaluation guarantees certify that every constraint holds with additive slack at most 3ξ. Since $|a_{t,p,j}(x)| \leq 1$ and the residuals are bounded, ζ may be restricted to a bounded interval. Thus the linear program has a bounded finite-dimensional feasible region, and the oracle supplies approximate separation for its continuously many constraints. Given a target minimax accuracy $\rho \in (0,1)$, the weak ellipsoid method [26] chooses its internal tolerances, including ξ, with $\log(1/\xi)$ polynomial in $|\mathcal{P}_Q|$ and $\log(1/\rho)$. It returns a distribution $\pi_t(\cdot \mid x)$ satisfying (4) using a number of operations and oracle calls polynomial in the same quantities.

Fix a deterministic implementation of the weak ellipsoid method, including deterministic tie-breaking, and combine it with the fixed oracle from Assumption 4.10. Denote the resulting solver by

$$\mathsf{Solve}_\rho(H,x) \in \Delta(\mathcal{P}_Q),$$

where $H = (\mathrm{CumRes}_{g,p,j})_{g,p,j}$ is the array of cumulative residuals available before the round. The internal tolerances are chosen so that $\mathsf{Solve}_\rho(H,x)$ satisfies (4) with additive error at most ρ. In the computational implementation, the rule selected at round t is defined for every context by

$$\pi_t(\cdot \mid x) := \mathsf{Solve}_\rho\Big(\big(\mathrm{CumRes}^{t-1}_{g,p,j}\big)_{g,p,j}, x\Big).$$

Thus the rule is fixed by the pre-round state and the context, and the same rule can be reproduced after training.

We can therefore compute the exponential weights distribution and solve the minimax problem in polynomial time; the following theorem records the resulting runtime guarantee.

Theorem 4.11 (Polynomial-Time Implementation)**.** *Assume Assumptions 4.1 and 4.10. Then, for every minimax accuracy $\rho \in (0,1)$, the learner can be implemented with training time polynomial in*

$$n, \quad Q^k, \quad |\mathcal{G}|, \quad \log(1/\rho).$$

Choosing $\rho = \varepsilon/6$ and the remaining parameters as in Corollary 4.9 implements its learner. In particular, for every fixed $\kappa > 0$, if $|\mathcal{G}| \le \varepsilon^{-\kappa}$, then the learner runs in time polynomial in $1/\varepsilon$.

The output predictor has a representation of polynomial size, and for any queried context x, either its full prediction distribution or an exact sample from that distribution can be computed in time polynomial in the same parameters.

Proof. The formulas for $\omega_t(g)$ and $m_{t,g,p,j}$ represent the exact exponential weights distribution over the sign-pattern experts without enumerating them, while the formula for $a_{t,p,j}(x)$ computes the resulting coefficients in the minimax objective. These calculations require storing only the $O(k|\mathcal{G}||\mathcal{P}_Q|)$ values $\mathrm{CumRes}^{t-1}_{g,p,j}$. At each round, the coefficients $a_{t,p,j}(x)$ can be computed in time polynomial in $k|\mathcal{G}||\mathcal{P}_Q|$ for any queried context x. During training, they need only be computed at the observed context X_t. The inner minimax problem has $|\mathcal{P}_Q|+1$ variables, and the approximate-separation argument in the preceding subsection computes a ρ-approximate solution in the stated running time. The ρ-approximate minimax solution contributes the additive ρ term in (5). Finally, choose $\rho = \varepsilon/6$, $Q = \Theta(1/\varepsilon)$, and $n = O(\varepsilon^{-(k+2)} + \varepsilon^{-2}\log|\mathcal{G}|)$ as in Corollary 4.9. These choices give the claimed polynomial runtime in $1/\varepsilon$.

To represent the output, store the fixed solver specification together with the T pre-round arrays

$$H_{t-1} := \left(\mathrm{CumRes}^{t-1}_{g,p,j}\right)_{g,p,j}, \qquad t \in [T].$$

This uses $O(Tk|\mathcal{G}||\mathcal{P}_Q|)$ real numbers. For a queried context x, rerunning the same deterministic map $\mathsf{Solve}_\rho(H_{t-1}, x)$ reproduces the rule $\pi_t(\cdot \mid x)$ used by the learner. The full averaged distribution

$$\Pi_S(\cdot \mid x) = \frac{1}{T}\sum_{t=1}^{T} \mathsf{Solve}_\rho(H_{t-1}, x)$$

is obtained by T solver calls. Alternatively, an exact sample from $\Pi_S(\cdot \mid x)$ is obtained by drawing $t \sim \mathrm{Unif}\{1,\dots,T\}$, computing $\mathsf{Solve}_\rho(H_{t-1}, x)$, and sampling from that distribution. Both the representation size and these evaluation procedures are polynomial in n, Q^k, $|\mathcal{G}|$, and $\log(1/\rho)$. □

5 Canonical Instantiations

We now show that the general lower and upper bounds apply to three multilevel properties: mean and mean absolute deviation; mean, variance, and skewness; and quantile and CVaR. In each case, they give matching sample-complexity bounds up to logarithmic factors. The lower bound uses a local family of outcome distributions, while the upper bound may hold over a larger class and uses an explicit optimization oracle.

5.1 Mean, Mean Absolute Deviation

Let $\mathcal{M}$ be the class of all probability laws on $[0,1]$, and use the prediction space

$$\mathcal{P} = [0,1]^2.$$

For $\mu \in \mathcal{M}$, define

$$m(\mu) := \mathbb{E}_\mu Y, \qquad d(\mu) := \mathbb{E}_\mu |Y - m(\mu)|.$$

The two-level property is

$$\Gamma(\mu) = (m(\mu), d(\mu)).$$

It is sequentially conditionally identifiable with residual functions

$$R_1(\varnothing, m, y) = m - y, \qquad R_2(m, d, y) = d - |y - m|.$$

The second coordinate is the expected ℓ_1 loss evaluated at the first coordinate. It is therefore analogous to variance, but uses absolute loss rather than squared loss. It is not a bilevel Bayes pair, since absolute loss elicits medians rather than means.

We first construct a local witness family on the three-point support

$$\mathcal{Y}_{\mathrm{MAD}} = \{0, 1/2, 1\}.$$

Let $v^\circ = (7/20, 7/20)$, and let $\mathcal{R}_{\mathrm{MAD}}$ be a sufficiently small closed rectangle around v°, contained in $(0, 1/2) \times (0, 1)$. For $v = (v_1, v_2) \in \mathcal{R}_{\mathrm{MAD}}$, define μ_v by

$$\mathbb{P}_{Y \sim \mu_v}(Y = 0) = \frac{v_2}{2v_1},$$

$$\mathbb{P}_{Y \sim \mu_v}(Y = 1/2) = 2(1 - v_1) - \frac{v_2}{v_1},$$

$$\mathbb{P}_{Y \sim \mu_v}(Y = 1) = 2v_1 - 1 + \frac{v_2}{2v_1}.$$

At v°, these probabilities are $1/2, 3/10, 1/5$. Since they depend continuously on v, $\mathcal{R}_{\mathrm{MAD}}$ can be chosen small enough that each is uniformly bounded below by a positive constant throughout the rectangle. Since $v_1 < 1/2$ on $\mathcal{R}_{\mathrm{MAD}}$, direct calculation gives

$$\mathbb{E}_{\mu_v} Y = v_1, \qquad \mathbb{E}_{\mu_v} |Y - v_1| = v_2.$$

Proposition 5.1 (Mean, Mean Absolute Deviation: Witness)**.** *The family* $\{\mu_v : v \in \mathcal{R}_{\mathrm{MAD}}\}$ *satisfies Assumption 3.1.*

Proof. We verified above that $\Gamma(\mu_v) = v$, as required by Assumption 3.1(i). The residuals evaluated at the true prefix are

$$R_1(v_{<1}, q, \mu_v) = q - v_1, \qquad R_2(v_1, q, \mu_v) = q - v_2.$$

Therefore conditions (ii) and (iii) in Assumption 3.1 hold.

It remains to verify Assumption 3.1(iv). The first coordinate has no preceding predictions. For the second coordinate,

$$\begin{aligned} &|R_2(p_1, p_2, \mu_v) - R_2(v_1, p_2, \mu_v)| \\ &\quad = \left|\mathbb{E}_{\mu_v}\left[|Y - v_1| - |Y - p_1|\right]\right| \le |p_1 - v_1|, \end{aligned}$$

where the last step uses the reverse triangle inequality. This verifies Assumption 3.1(iv).

Finally, the probability vector of μ_v is a smooth function of $v = (v_1, v_2)$ on $\mathcal{R}_{\mathrm{MAD}}$, and all three coordinates are uniformly bounded below. Hence, for some constant $C < \infty$,

$$D_{\mathrm{KL}}(\mu_v \,\|\, \mu_{v'}) \le C\|v - v'\|_2^2.$$

This verifies Assumption 3.1(v). □

Corollary 5.2 (Mean, Mean Absolute Deviation: Lower Bound)**.** *Fix $\kappa > 0$. There exist constants $c, C, \varepsilon_0 > 0$, depending only on κ, such that the following holds for every $0 < \varepsilon \le \varepsilon_0$. For the property*

$$\Gamma = (\mathbb{E}Y, \mathbb{E}|Y - \mathbb{E}Y|),$$

one can construct a finite context space, a binary group family of size at most $\varepsilon^{-\kappa}$, and a finite collection of data distributions whose conditional outcome laws are supported on three points in $[0,1]$, such that every learner achieving Γ-ECE at most ε with probability at least $2/3$ must use

$$n \ge c\frac{\varepsilon^{-4}}{\log^4(C/\varepsilon)}.$$

Equivalently,

$$\mathrm{SC}^{(\kappa)}_{\Gamma,\mathcal{M}}(\varepsilon) = \widetilde{\Omega}(\varepsilon^{-4}).$$

Proof. Apply Theorem 3.2 using Proposition 5.1. □

We next verify the upper-bound conditions over $\mathcal{M}$ and give an exact optimization oracle.

Proposition 5.3 (Mean, Mean Absolute Deviation: Conditions for the Upper Bound)**.** *The regularity conditions in Assumption 4.1 hold over $\mathcal{M}$. Moreover, the linear optimization oracle in Assumption 4.10 has an exact polynomial-time implementation.*

Proof. The law class $\mathcal{M}$ is convex and compact in the weak topology. For each fixed prediction (m, d), the maps

$$\mu \mapsto \mathbb{E}_\mu[m - Y], \qquad \mu \mapsto \mathbb{E}_\mu[d - |Y - m|]$$

are affine and continuous. These observations verify Assumption 4.1(i). Condition (ii) also holds because both coordinates of the prediction and the outcome lie in $[0,1]$.

For two predictions $(m, d), (m', d')$,

$$\big|R_1(\varnothing, m, \mu) - R_1(\varnothing, m', \mu)\big| = |m - m'|,$$

and

$$\big|R_2(m, d, \mu) - R_2(m', d', \mu)\big| \le |d - d'| + |m - m'|.$$

Thus $R(\cdot, \mu)$ is uniformly Lipschitz. A rectangular grid with mesh $O(1/Q)$ therefore satisfies the required approximation bound and has $O(Q^2)$ points, verifying Assumption 4.1(iii).

For the oracle, given coefficients $c_{p,j}$, write each grid point as $p = (m_p, d_p)$, and define

$$\phi(y) := \sum_{p \in \mathcal{P}_Q} \left[c_{p,1}(m_p - y) + c_{p,2}\big(d_p - |y - m_p|\big)\right].$$

Maximizing $\mathbb{E}_\mu\phi(Y)$ over all probability laws on $[0,1]$ is the same as maximizing $\phi(y)$ over $y \in [0,1]$, because an optimizer may be taken to be a point mass at a maximizer. The function ϕ is piecewise affine with breakpoints among the grid values m_p, so a maximizer lies in $\{0,1\} \cup \{m_p : p \in \mathcal{P}_Q\}$. Checking these points gives an exact polynomial-time oracle. □

Corollary 5.4 (Mean, Mean Absolute Deviation: Upper Bound)**.** *Fix $\kappa > 0$. There exist constants $C, \varepsilon_0 > 0$, depending only on κ, such that the following holds for every $0 < \varepsilon \le \varepsilon_0$. Let*

$$\Gamma = (\mathbb{E}Y, \mathbb{E}|Y - \mathbb{E}Y|)$$

be the property on $[0,1]$. *For every data distribution* D *on* $\mathcal{X} \times [0,1]$ *and every finite group family satisfying* $|\mathcal{G}| \le \lceil \varepsilon^{-\kappa} \rceil$, *there is a randomized learner using at most* $C\varepsilon^{-4}$ *samples whose output* Π_S *satisfies*

$$\mathbb{P}\left[\mathrm{MCErr}^{\Gamma}_{\mathsf{D}}(\Pi_S; \mathcal{G}) \le \varepsilon\right] \ge \frac{2}{3}.$$

The exact oracle in Proposition 5.3 gives training time polynomial in $1/\varepsilon$.

Proof. Apply Corollary 4.9 and Theorem 4.11 using Proposition 5.3. □

5.2 Mean, Variance, Skewness

Let

$$\mathcal{Y}_{\mathrm{MVS}} = \{0, 1/3, 2/3, 1\}.$$

Fix $h \in (0, 1/4)$, and let $\mathcal{M}_h$ be the set of probability laws on $\mathcal{Y}_{\mathrm{MVS}}$ that assign mass at least h to every atom. This class is convex and compact. Moreover, every $\mu \in \mathcal{M}_h$ has variance bounded below:

$$\mathrm{Var}_\mu(Y) = \frac{1}{2} \sum_{y,y' \in \mathcal{Y}_{\mathrm{MVS}}} \mu(y)\mu(y')(y - y')^2 \ge \mu(0)\mu(1) \ge h^2.$$

For $\mu \in \mathcal{M}_h$, define

$$m(\mu) := \mathbb{E}_\mu Y, \qquad \sigma^2(\mu) := \mathbb{E}_\mu (Y - m(\mu))^2,$$

and the standardized skewness

$$\gamma(\mu) := \frac{\mathbb{E}_\mu (Y - m(\mu))^3}{\left(\sigma^2(\mu)\right)^{3/2}}.$$

The three-level property is

$$\Gamma(\mu) = (m(\mu), \sigma^2(\mu), \gamma(\mu)).$$

Since $|Y - m(\mu)| \le 1$ and $\sigma^2(\mu) \ge h^2$, its range lies in

$$\mathcal{P} := [0,1] \times [h^2, 1/4] \times [-h^{-3}, h^{-3}].$$

It is sequentially conditionally identifiable with residual functions

$$R_1(\varnothing, m, y) = m - y,$$

$$R_2(m, \sigma^2, y) = \sigma^2 - (y - m)^2,$$

$$R_3((m, \sigma^2), \gamma, y) = \gamma\,(\sigma^2)^{3/2} - (y - m)^3.$$

The third residual depends on both preceding coordinates and identifies skewness only after the mean and variance have been fixed.

We now construct a local three-dimensional witness family. Let μ° be the uniform distribution on $\mathcal{Y}_{\mathrm{MVS}}$, and let

$$v^\circ := \Gamma(\mu^\circ) = \left(\frac{1}{2}, \frac{5}{36}, 0\right).$$

For a target vector $v = (v_1, v_2, v_3)$ with $v_2 > 0$, define the corresponding raw moments

$$r_1(v) := v_1, \qquad r_2(v) := v_1^2 + v_2, \qquad r_3(v) := v_1^3 + 3v_1v_2 + v_3 v_2^{3/2}.$$

Define weights on $\mathcal{Y}_{\mathrm{MVS}}$ by

$$\begin{aligned}
\mu_v(0) &:= 1 - \frac{11}{2} r_1(v) + 9 r_2(v) - \frac{9}{2} r_3(v), \\
\mu_v(1/3) &:= 9 r_1(v) - \frac{45}{2} r_2(v) + \frac{27}{2} r_3(v), \\
\mu_v(2/3) &:= -\frac{9}{2} r_1(v) + 18 r_2(v) - \frac{27}{2} r_3(v), \\
\mu_v(1) &:= r_1(v) - \frac{9}{2} r_2(v) + \frac{9}{2} r_3(v).
\end{aligned}$$

A direct calculation gives

$$\sum_y \mu_v(y) = 1, \qquad \sum_y \mu_v(y) y^j = r_j(v) \quad \text{for } j \in \{1, 2, 3\}.$$

At $v = v^\circ$, all four weights equal $1/4$. Since $h < 1/4$, continuity gives a nondegenerate closed box

$$\mathcal{R}_{\mathrm{MVS}} \subset \mathrm{int}(\mathcal{P})$$

around v°, small enough that $\mu_v(y) \geq h$ for every $v \in \mathcal{R}_{\mathrm{MVS}}$ and $y \in \mathcal{Y}_{\mathrm{MVS}}$. Thus $\mu_v \in \mathcal{M}_h$ throughout this box. Expanding the centered moments gives

$$\mathbb{E}_{\mu_v}(Y - v_1)^2 = v_2, \qquad \mathbb{E}_{\mu_v}(Y - v_1)^3 = v_3 v_2^{3/2},$$

and hence

$$\Gamma(\mu_v) = v \qquad \forall v \in \mathcal{R}_{\mathrm{MVS}}.$$

Proposition 5.5 (Mean, Variance, Skewness: Witness)**.** *The family* $\{\mu_v : v \in \mathcal{R}_{\mathrm{MVS}}\}$ *satisfies Assumption 3.1.*

Proof. The identity $\Gamma(\mu_v) = v$ holds by construction, as required by Assumption 3.1(i). The residuals evaluated at the true prefix are

$$R_1(v_{<1}, q, \mu_v) = q - v_1, \qquad R_2(v_1, q, \mu_v) = q - v_2,$$

and

$$R_3((v_1, v_2), q, \mu_v) = v_2^{3/2}(q - v_3).$$

Since $\mathcal{R}_{\mathrm{MVS}}$ is compact and $v_2 \geq h^2$, the coefficients 1 and $v_2^{3/2}$ are uniformly bounded above and away from zero. Therefore conditions (ii) and (iii) in Assumption 3.1 hold.

It remains to verify Assumption 3.1(iv). The first coordinate has no preceding predictions. For the second coordinate,

$$R_2(p_1, p_2, \mu_v) - R_2(v_1, p_2, \mu_v) = \mathbb{E}_{\mu_v}\left[(v_1 - Y)^2 - (p_1 - Y)^2\right].$$

Since $p_1, v_1, Y \in [0, 1]$, the above equation gives

$$|R_2(p_1, p_2, \mu_v) - R_2(v_1, p_2, \mu_v)| \leq 2|p_1 - v_1|.$$

For the third coordinate,

$$\begin{aligned}
&|R_3((p_1, p_2), p_3, \mu_v) - R_3((v_1, v_2), p_3, \mu_v)| \\
&\quad \leq |p_3|\,|p_2^{3/2} - v_2^{3/2}| + \mathbb{E}_{\mu_v}\left|(Y - p_1)^3 - (Y - v_1)^3\right|.
\end{aligned}$$

On $\mathcal{P}$, $|p_3| \le h^{-3}$, the map $u \mapsto u^{3/2}$ is Lipschitz on $[h^2, 1/4]$, and $(y, a) \mapsto (y-a)^3$ is Lipschitz for $y, a \in [0,1]$. Therefore

$$|R_3((p_1,p_2),p_3,\mu_v) - R_3((v_1,v_2),p_3,\mu_v)| \le C\big(|p_1 - v_1| + |p_2 - v_2|\big),$$

with C depending only on h. This verifies Assumption 3.1(iv).

Finally, the explicit formulas above show that $v \mapsto \mu_v$ is Lipschitz on $\mathcal{R}_{\mathrm{MVS}}$. Since $\mu_{v'}(y) \ge h$ for every atom, the inequality $\log u \le u - 1$ gives

$$\begin{aligned} D_{\mathrm{KL}}(\mu_v \,\|\, \mu_{v'}) &\le \sum_{y \in \mathcal{Y}_{\mathrm{MVS}}} \frac{\big(\mu_v(y) - \mu_{v'}(y)\big)^2}{\mu_{v'}(y)} \\ &\le \frac{1}{h} \sum_{y \in \mathcal{Y}_{\mathrm{MVS}}} \big(\mu_v(y) - \mu_{v'}(y)\big)^2 \\ &\le C\|v - v'\|_2^2. \end{aligned}$$

This verifies Assumption 3.1(v). □

Corollary 5.6 (Mean, Variance, Skewness: Lower Bound)**.** *Fix $h \in (0, 1/4)$ and $\kappa > 0$. There exist constants $c, C, \varepsilon_0 > 0$, depending only on h and κ, such that the following holds for every $0 < \varepsilon \le \varepsilon_0$. For the property*

$$\Gamma = (\textit{mean}, \textit{variance}, \textit{skewness}),$$

one can construct a finite context space, a binary group family of size at most $\varepsilon^{-\kappa}$, and a finite collection of data distributions whose conditional outcome laws belong to $\mathcal{M}_h$, such that every learner achieving Γ-ECE at most ε with probability at least $2/3$ must use

$$n \ge c \frac{\varepsilon^{-5}}{\log^5(C/\varepsilon)}.$$

Equivalently,

$$\mathrm{SC}^{(\kappa)}_{\Gamma, \mathcal{M}_h}(\varepsilon) = \widetilde{\Omega}(\varepsilon^{-5}).$$

Proof. Apply Theorem 3.2 using Proposition 5.5. □

We next verify the upper-bound conditions over $\mathcal{M}_h$ and give an exact optimization oracle.

Proposition 5.7 (Mean, Variance, Skewness: Conditions for the Upper Bound)**.** *The regularity conditions in Assumption 4.1 hold over $\mathcal{M}_h$. Moreover, the linear optimization oracle in Assumption 4.10 has an exact polynomial-time implementation.*

Proof. The law class $\mathcal{M}_h$ is a convex, compact subset of the four-dimensional probability simplex. For each fixed prediction $p = (m, \sigma^2, \gamma)$, the maps

$$\mu \mapsto \mathbb{E}_\mu[m - Y],$$

$$\mu \mapsto \mathbb{E}_\mu\big[\sigma^2 - (Y-m)^2\big],$$

$$\mu \mapsto \mathbb{E}_\mu\Big[\gamma\,(\sigma^2)^{3/2} - (Y-m)^3\Big]$$

are affine and continuous. These observations verify Assumption 4.1(i). Condition (ii) also holds because $\mathcal{P}$ and $\mathcal{Y}_{\mathrm{MVS}}$ are compact.

It remains to verify Assumption 4.1(iii). For two predictions $p = (m, \sigma^2, \gamma)$ and $p' = (m', \sigma'^2, \gamma')$ in $\mathcal{P}$,

$$\big|R_1(\varnothing, m, \mu) - R_1(\varnothing, m', \mu)\big| = |m - m'|,$$

$$\Big|R_2(m, \sigma^2, \mu) - R_2(m', \sigma'^2, \mu)\Big| \le |\sigma^2 - \sigma'^2| + 2|m - m'|,$$

and

$$\Big|R_3((m, \sigma^2), \gamma, \mu) - R_3((m', \sigma'^2), \gamma', \mu)\Big| \le C_h\big(|m - m'| + |\sigma^2 - \sigma'^2| + |\gamma - \gamma'|\big),$$

because γ is bounded, $u \mapsto u^{3/2}$ is Lipschitz on $[h^2, 1/4]$, and $(y, m) \mapsto (y - m)^3$ is Lipschitz on $\mathcal{Y}_{\mathrm{MVS}} \times [0, 1]$. Thus $R(\cdot, \mu)$ is uniformly Lipschitz on $\mathcal{P}$. A rectangular grid with mesh $O(1/Q)$ therefore satisfies the required approximation bound and has $O(Q^3)$ points.

For the oracle, given coefficients $c_{p,j}$, write each grid point as $p = (m_p, \sigma_p^2, \gamma_p)$, and define

$$\phi(y) := \sum_{p \in \mathcal{P}_Q} \Big[c_{p,1}(m_p - y) + c_{p,2}\big(\sigma_p^2 - (y - m_p)^2\big) + c_{p,3}\big(\gamma_p\,(\sigma_p^2)^{3/2} - (y - m_p)^3\big)\Big].$$

The oracle problem is

$$\sup_{\mu \in \mathcal{M}_h} \sum_{y \in \mathcal{Y}_{\mathrm{MVS}}} \mu(y)\phi(y).$$

This is a linear optimization problem over the simplex with lower bounds $\mu(y) \ge h$. An optimizer assigns mass h to each atom and the remaining mass $1 - 4h$ to an atom maximizing $\phi(y)$. Thus the oracle is exact and polynomial time. □

Corollary 5.8 (Mean, Variance, Skewness: Upper Bound)**.** *Fix $h \in (0, 1/4)$ and $\kappa > 0$. There exist constants $C, \varepsilon_0 > 0$, depending only on h and κ, such that the following holds for every $0 < \varepsilon \le \varepsilon_0$. Let*

$$\Gamma = (\textit{mean}, \textit{variance}, \textit{skewness})$$

be the property over $\mathcal{M}_h$. For every $\mathcal{M}_h$-compatible data distribution D and every finite group family satisfying $|\mathcal{G}| \le \lceil \varepsilon^{-\kappa} \rceil$, there is a randomized learner using at most $C\varepsilon^{-5}$ samples whose output Π_S satisfies

$$\mathbb{P}\left[\mathrm{MCErr}^{\Gamma}_{\mathsf{D}}(\Pi_S; \mathcal{G}) \le \varepsilon\right] \ge \frac{2}{3}.$$

The exact oracle in Proposition 5.7 gives training time polynomial in $1/\varepsilon$.

Proof. Apply Corollary 4.9 and Theorem 4.11 using Proposition 5.7. □

5.3 Quantile, CVaR

Fix $\tau \in (0, 1)$ and constants $0 < h < 1 < H < \infty$. Let $\mathcal{M}_{h,H}$ be the class of probability laws on $[0, 1]$ with densities f satisfying

$$h \le f(y) \le H \qquad \text{for almost every } y \in [0, 1].$$

We use the prediction space

$$\mathcal{P} = [0, 1]^2.$$

For $\mu \in \mathcal{M}_{h,H}$, define the τ-quantile

$$q_\tau(\mu) := \inf\{q : \mu(Y \le q) \ge \tau\}.$$

Use the upper-tail convention

$$\mathrm{CVaR}_\tau(\mu) := q_\tau(\mu) + \frac{\mathbb{E}_\mu(Y - q_\tau(\mu))_+}{1-\tau}.$$

Consider the scoring function

$$L_\tau(q, y) := q + \frac{(y-q)_+}{1-\tau}.$$

Over $\mathcal{M}_{h,H}$, L_τ is strictly consistent for q_τ, and its Bayes risk is CVaR_τ [37]. Thus

$$(q_\tau, \mathrm{CVaR}_\tau)$$

is a Bayes pair. This is the $k = 2$ case of the sequential framework with residual functions

$$R_1(\varnothing, q, y) = \mathbf{1}\{y \le q\} - \tau, \qquad R_2(q, r, y) = r - L_\tau(q, y).$$

We now construct a two-dimensional local witness. For $\lambda = (\lambda_1, \lambda_2)$ in a small neighborhood of $0 \in \mathbb{R}^2$, define the density

$$f_\lambda(y) = \exp(\lambda_1 y + \lambda_2 y^2 - A(\lambda)), \qquad 0 \le y \le 1,$$

where $A(\lambda)$ is the log normalizer. Let μ_λ be the corresponding distribution. Since $f_0 \equiv 1$, we may choose the neighborhood of 0 small enough that $\mu_\lambda \in \mathcal{M}_{h,H}$ throughout it.

Define

$$\Psi(\lambda) := \left(q_\tau(\mu_\lambda), \mathrm{CVaR}_\tau(\mu_\lambda)\right).$$

At $\lambda = 0$, $\mu_0 = \mathrm{Unif}[0, 1]$, so

$$\Psi(0) = \left(\tau, \frac{1+\tau}{2}\right).$$

Lemma 5.9 (Local Two-Dimensional Parameterization)**.** *The Jacobian $D\Psi(0)$ is nonsingular. Consequently, there exists a nondegenerate rectangle*

$$\mathcal{R}_{\mathrm{QC}} \subset \mathrm{range}(\Psi)$$

around $\big(\tau, (1+\tau)/2\big)$ and a smooth inverse map

$$\lambda = \lambda(v), \qquad v = (v_1, v_2) \in \mathcal{R}_{\mathrm{QC}},$$

such that

$$q_\tau(\mu_{\lambda(v)}) = v_1, \qquad \mathrm{CVaR}_\tau(\mu_{\lambda(v)}) = v_2.$$

Proof. Since $f_\lambda(y)$ is smooth in (λ, y) and strictly positive, the implicit equation $\int_0^{q_\tau(\mu_\lambda)} f_\lambda(y)\, dy = \tau$ and the implicit function theorem show that $\lambda \mapsto q_\tau(\mu_\lambda)$ is smooth near 0. The defining integral for $\mathrm{CVaR}_\tau(\mu_\lambda)$ is then smooth as well, so Ψ is smooth near 0.

At $\lambda = 0$, direct differentiation gives

$$D\Psi(0) = \begin{pmatrix} \dfrac{\tau(1-\tau)}{2} & \dfrac{\tau(1-\tau^2)}{3} \\[2ex] \dfrac{(1-\tau)(2\tau+1)}{12} & \dfrac{(1-\tau)(\tau+1)^2}{12} \end{pmatrix}.$$

Its determinant is

$$\det D\Psi(0) = \frac{\tau(1-\tau)^3(1+\tau)}{72} > 0.$$

The inverse function theorem gives the claim. □

For $v \in \mathcal{R}_{\mathrm{QC}}$, define

$$\mu_v := \mu_{\lambda(v)}.$$

It remains to check that the local parameterization from Lemma 5.9 has the residual and KL properties required by Assumption 3.1.

Proposition 5.10 (Quantile, CVaR: Witness)**.** *For $\mathcal{R}_{\mathrm{QC}}$ sufficiently small, the family*

$$\{\mu_v : v \in \mathcal{R}_{\mathrm{QC}}\}$$

is contained in $\mathcal{M}_{h,H}$ and satisfies Assumption 3.1 for the property

$$(q_\tau, \mathrm{CVaR}_\tau).$$

Proof. By Lemma 5.9, $\Gamma(\mu_v) = v$, as required by Assumption 3.1(i).

By the choice of the parameter neighborhood, $\mu_v \in \mathcal{M}_{h,H}$ for every $v \in \mathcal{R}_{\mathrm{QC}}$. Let F_v be the CDF of μ_v. Since $F_v(v_1) = \tau$,

$$R_1(v_{<1}, q, \mu_v) = F_v(q) - \tau = F_v(q) - F_v(v_1), \qquad R_2(v_1, r, \mu_v) = r - v_2.$$

Therefore,

$$\operatorname{sgn}(q - v_1)\,(R_1(v_{<1}, q, \mu_v) - R_1(v_{<1}, v_1, \mu_v)) \geq h|q - v_1|.$$

The second residual evaluated at the true prefix satisfies

$$\operatorname{sgn}(r - v_2)\,(R_2(v_1, r, \mu_v) - R_2(v_1, v_2, \mu_v)) = |r - v_2|.$$

Thus Assumption 3.1(ii) holds. The bound in condition (iii) follows from

$$|R_1(v_{<1}, q, \mu_v) - R_1(v_{<1}, v_1, \mu_v)| = |F_v(q) - F_v(v_1)| \leq H|q - v_1|$$

and

$$|R_2(v_1, r, \mu_v) - R_2(v_1, v_2, \mu_v)| = |r - v_2|.$$

It remains to verify Assumption 3.1(iv). The first coordinate has no preceding predictions. For the second coordinate, fix $y \in [0, 1]$. The map

$$q \mapsto L_\tau(q, y) = q + \frac{(y - q)_+}{1 - \tau}$$

is Lipschitz with constant

$$L_\tau^* := \max\left\{1, \frac{\tau}{1 - \tau}\right\}.$$

Hence

$$|\mathbb{E}_{\mu_v} L_\tau(p_1, Y) - \mathbb{E}_{\mu_v} L_\tau(v_1, Y)| \leq L_\tau^* |p_1 - v_1|.$$

Since

$$R_2(p_1, p_2, \mu_v) - R_2(v_1, p_2, \mu_v) = -\left(\mathbb{E}_{\mu_v} L_\tau(p_1, Y) - \mathbb{E}_{\mu_v} L_\tau(v_1, Y)\right),$$

this verifies Assumption 3.1(iv).

Finally, the family $\{f_\lambda\}$ is a regular two-parameter exponential family with bounded sufficient statistics y and y^2. On a compact neighborhood of 0,

$$D_{\mathrm{KL}}(\mu_\lambda \,\|\, \mu_{\lambda'}) \leq C\|\lambda - \lambda'\|_2^2.$$

Because $v \mapsto \lambda(v)$ is smooth on $\mathcal{R}_{\mathrm{QC}}$,

$$D_{\mathrm{KL}}(\mu_v \,\|\, \mu_{v'}) \leq C'\|v - v'\|_2^2.$$

Thus Assumption 3.1(v) holds. □

Corollary 5.11 (Quantile, CVaR: Lower Bound)**.** *Fix $\tau \in (0,1)$, constants $0 < h < 1 < H < \infty$, and $\kappa > 0$. There exist constants $c, C, \varepsilon_0 > 0$, depending only on τ, h, H, and κ, such that the following holds for every $0 < \varepsilon \le \varepsilon_0$. Let*

$$\Gamma = (q_\tau, \mathrm{CVaR}_\tau)$$

be the property over $\mathcal{M}_{h,H}$. One can construct a finite context space, a binary group family of size at most $\varepsilon^{-\kappa}$, and a finite collection of $\mathcal{M}_{h,H}$-compatible data distributions, such that every learner achieving Γ-ECE at most ε with probability at least $2/3$ must use

$$n \ge c \frac{\varepsilon^{-4}}{\log^4(C/\varepsilon)}.$$

Equivalently,

$$\mathrm{SC}^{(\kappa)}_{\Gamma, \mathcal{M}_{h,H}}(\varepsilon) = \widetilde{\Omega}(\varepsilon^{-4}).$$

Proof. Apply Theorem 3.2 using Proposition 5.10. □

For the upper bound, we verify the required conditions over the same class $\mathcal{M}_{h,H}$.

Proposition 5.12 (Quantile, CVaR: Conditions for the Upper Bound)**.** *The regularity conditions in Assumption 4.1 hold over $\mathcal{M}_{h,H}$. Moreover, the linear optimization oracle in Assumption 4.10 has a polynomial-time implementation to any additive accuracy $\xi > 0$.*

Proof. Recall that

$$R((q,r), y) = \left(\mathbf{1}\{y \le q\} - \tau,\ r - L_\tau(q,y)\right).$$

The class $\mathcal{M}_{h,H}$ is convex and compact in the weak topology. For each fixed (q,r), the map

$$\mu \mapsto F_\mu(q) - \tau$$

is affine and continuous on $\mathcal{M}_{h,H}$, because every law in this class has no atom at q. The map

$$\mu \mapsto \mathbb{E}_\mu[r - L_\tau(q, Y)]$$

is affine and continuous because $L_\tau(q,\cdot)$ is bounded and continuous. These observations verify Assumption 4.1(i). Condition (ii) also holds because $q, r, y \in [0,1]$ and $\tau \in (0,1)$.

To verify Assumption 4.1(iii), let F_μ be the CDF of μ. For $q, q' \in [0,1]$,

$$\left|R_1(\varnothing, q, \mu) - R_1(\varnothing, q', \mu)\right| = |F_\mu(q) - F_\mu(q')| \le H|q - q'|.$$

Also, the map $q \mapsto L_\tau(q,y)$ is Lipschitz uniformly in y, with a constant depending only on τ. Therefore $R(\cdot, \mu)$ is uniformly Lipschitz in (q,r). A rectangular grid with mesh $O(1/Q)$ satisfies the required approximation bound and has $O(Q^2)$ points.

It remains to implement the oracle. Given coefficients $c_{p,j}$, write each grid point as $p = (q_p, r_p)$. The integrand

$$\phi(y) = \sum_{p \in \mathcal{P}_Q} \left[c_{p,1}(\mathbf{1}\{y \le q_p\} - \tau) + c_{p,2}\big(r_p - q_p - (y - q_p)_+/(1-\tau)\big)\right]$$

is affine on each open interval between consecutive distinct grid values q_p, and it may have jump discontinuities at those values. Since there are only finitely many such points, their values do not affect the oracle objective. The oracle must maximize

$$\int_0^1 \phi(y) f(y)\, dy \qquad \text{subject to} \qquad h \le f \le H, \quad \int_0^1 f = 1.$$

Write $f = h + (H - h)u$, where $0 \le u \le 1$ and

$$\int_0^1 u(y)\,dy = \frac{1-h}{H-h} =: m_0.$$

The objective differs by the constant $h\int\phi$ from maximizing $\int \phi u$ over $0 \le u \le 1$ with mass m_0. For a threshold t, let

$$A(t) := \int_0^1 \mathbf{1}\{\phi(y) > t\}\,dy, \qquad B(t) := \int_0^1 \mathbf{1}\{\phi(y) \ge t\}\,dy.$$

Choose t such that $A(t) \le m_0 \le B(t)$, and set

$$\gamma := \begin{cases} \dfrac{m_0 - A(t)}{B(t) - A(t)}, & B(t) > A(t), \\ 0, & B(t) = A(t). \end{cases}$$

Then

$$u_t(y) := \mathbf{1}\{\phi(y) > t\} + \gamma \mathbf{1}\{\phi(y) = t\}$$

has mass m_0. It is optimal by the bathtub principle, which places the available mass where ϕ is largest. This includes the case in which ϕ is constant on an interval: γ supplies exactly the required fraction of that level set.

Because ϕ has $O(|\mathcal{P}_Q|)$ affine pieces, $A(t)$ is affine on each interval between consecutive critical values. These critical values are the values of ϕ at the endpoints of its affine pieces, together with the values on any constant pieces. Sorting them and solving one linear equation locates an admissible t and γ in polynomial time. The resulting density $f_t = h + (H - h)u_t$ is piecewise constant on $O(|\mathcal{P}_Q|)$ intervals, which gives a polynomial-size representation. Integrating the residuals interval by interval computes the objective and all residual expectations to additive accuracy ξ in time polynomial in $|\mathcal{P}_Q|$ and $\log(1/\xi)$. This implements Assumption 4.10. □

Corollary 5.13 (Quantile, CVaR: Upper Bound)**.** *Fix $\tau \in (0,1)$, constants $0 < h < 1 < H < \infty$, and $\kappa > 0$. There exist constants $C, \varepsilon_0 > 0$, depending only on τ, h, H, and κ, such that the following holds for every $0 < \varepsilon \le \varepsilon_0$. Let*

$$\Gamma = (q_\tau, \mathrm{CVaR}_\tau)$$

be the property over $\mathcal{M}_{h,H}$. For every $\mathcal{M}_{h,H}$-compatible data distribution D and every finite group family satisfying $|\mathcal{G}| \le \lceil \varepsilon^{-\kappa} \rceil$, there is a randomized learner using at most $C\varepsilon^{-4}$ samples whose output Π_S satisfies

$$\mathbb{P}\left[\mathrm{MCErr}^{\Gamma}_{\mathsf{D}}(\Pi_S; \mathcal{G}) \le \varepsilon\right] \ge \frac{2}{3}.$$

The oracle in Proposition 5.12 gives training time polynomial in $1/\varepsilon$.

Proof. Apply Corollary 4.9 and Theorem 4.11 using Proposition 5.12. □

Statement of AI Use

GPT-5.6 Sol Pro was used during manuscript preparation for brainstorming, drafting, editing, and formatting assistance, including preliminary drafts of proof arguments. The LLM-generated material was not used without author review. All proof arguments were independently checked, corrected, and finalized by the authors, who take full responsibility for the correctness, originality, and presentation of the paper.

References

[1] Krishnakumar Balasubramanian, Saeed Ghadimi, and Anthony Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

[2] Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*, volume 35, pages 29362–29373. Curran Associates, Inc., 2022.

[3] Sílvia 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. Association for Computing Machinery, 2024.

[4] Sílvia Casacuberta, Parikshit Gopalan, Varun Kanade, and Omer Reingold. How global calibration strengthens multiaccuracy. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1198–1227. IEEE, 2025.

[5] Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69: 4937–4948, 2021.

[6] 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. Association for Computing Machinery, 2025.

[7] 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)*, pages 4712–4798. Society for Industrial and Applied Mathematics, 2026.

[8] Natalie Collina, Jiuyao Lu, Georgy Noarov, and Aaron Roth. Optimal lower bounds for online multicalibration. In *2026 IEEE 67th Annual Symposium on Foundations of Computer Science (FOCS)*, 2026. To appear.

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

[10] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006.

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

[12] Zhun Deng, Cynthia Dwork, and Linjun Zhang. HappyMap: A generalized multicalibration method. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 41:1–41:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

[13] Darinka Dentcheva, Spiridon Penev, and Andrzej Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, 2017.

[14] Cynthia Dwork and Pranay Tankala. Supersimulators. *arXiv preprint arXiv:2509.17994*, 2025.

[15] Yuri M. Ermoliev and Vladimir I. Norkin. Sample average approximation method for compound stochastic optimization problems. *SIAM Journal on Optimization*, 23(4):2231–2263, 2013.

[16] Maxwell Fishelson, Noah Golowich, Mehryar Mohri, and Jon Schneider. High-dimensional calibration from swap regret. In *Advances in Neural Information Processing Systems*, volume 38, pages 50433–50465. Curran Associates, Inc., 2025.

[17] 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. Society for Industrial and Applied Mathematics, 2024.

[18] Rohan Ghuge, Vidya Muthukumar, and Sahil Singla. Improved and oracle-efficient online ℓ_1-multicalibration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19437–19457. PMLR, 2025.

[19] Isaac Gibbs and Ryan J. Tibshirani. Sample-efficient omniprediction for proper losses. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 2679–2719. PMLR, 2026.

[20] Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11459–11492. PMLR, 2023.

[21] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

[22] Parikshit Gopalan, Michael P. Kim, Mihir A. Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3193–3234. PMLR, 2022.

[23] 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)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 60:1–60:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

[24] Parikshit Gopalan, Michael Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In *Advances in Neural Information Processing Systems*, volume 36, pages 39936–39956. Curran Associates, Inc., 2023.

[25] Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1983–2026. PMLR, 2024.

[26] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer, 1988.

[27] 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)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 82:1–82:24. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

[28] Nika Haghtalab, Michael I. Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 72464–72506. Curran Associates, Inc., 2023.

[29] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 2018.

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

[31] Lunjia Hu, Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Efficient swap multicalibration of elicitable properties. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 3314–3348. PMLR, 2026.

[32] Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2634–2678. PMLR, 2021.

[33] Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *The Eleventh International Conference on Learning Representations*, 2023.

[34] Daniel D. Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. In *Advances in Neural Information Processing Systems*, volume 35, pages 29051–29063. Curran Associates, Inc., 2022.

[35] Haipeng Luo, Spandan Senapati, and Vatsal Sharan. Improved bounds for swap multicalibration and swap omniprediction. In *Advances in Neural Information Processing Systems*, volume 38, pages 20425–20467. Curran Associates, Inc., 2025.

[36] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2901–2907. AAAI Press, 2015.

[37] Georgy Noarov and Aaron Roth. The statistical scope of multicalibration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26283–26310. PMLR, 2023.

[38] Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 46762–46783. PMLR, 2025.

[39] Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-optimal algorithms for omniprediction. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1595–1609. IEEE, 2025.

[40] Binghui Peng. High dimensional online calibration in polynomial time. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2025.

[41] Harrison Rosenberg, Robi Bhattacharjee, Kassem Fawaz, and Somesh Jha. An exploration of multicalibration uniform convergence bounds. *arXiv preprint arXiv:2202.04530*, 2022.

[42] Ron M. Roth. *Introduction to Coding Theory*. Cambridge University Press, 2006.

[43] Andrzej Ruszczyński. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021.

[44] Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. In *Advances in Neural Information Processing Systems*, volume 33, pages 13331–13340. Curran Associates, Inc., 2020.

[45] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[46] Amnon Ta-Shma. Explicit, almost optimal, epsilon-balanced codes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 238–251. Association for Computing Machinery, 2017.

A Comparison with Multi-Level Stochastic Optimization

Here, we discuss the difference between k-level multicalibration problem and the well-studied k-level stochastic composite optimization problem. The key distinction is that the levels are *sequential* in compositional optimization but *Cartesian* in multicalibration.

A k-level composition optimization problem involves solving

$$F(x) = f_1 \circ f_2 \circ \cdots \circ f_k(x),$$

given access to stochastic evaluations of the gradients and function values of functions f_j. In this setup, typically, the quantity of interest is a fixed point x that determines only one chain of intermediate values,

$$y_k = x, \qquad y_{j-1} = f_j(y_j), \qquad F(x) = f_1(y_1).$$

The optimization algorithm must estimate and track the quantities along this chain, but its statistical goal is still to find a single point $\widehat{x}_N$ (where N is the sample or iteration complexity) satisfying a first-order condition, such as

$$\mathbb{E}\big[\|\nabla F(\widehat{x}_N)\|\big] \le \varepsilon.$$

It is not required to certify stationarity separately for every possible vector of intermediate values $(y_1, \ldots, y_{k-1})$. Smoothness allows the estimation errors arising at the different levels to be propagated along the chain and controlled on the same stochastic time scale. For the stationarity criterion considered in the multi-level optimization literature, the resulting bound has the schematic form

$$\mathbb{E}\big[\|\nabla F(\widehat{x}_N)\|^2\big] \lesssim N^{-1/2}, \qquad \text{and hence} \qquad \mathbb{E}\big[\|\nabla F(\widehat{x}_N)\|\big] \lesssim N^{-1/4}.$$

Thus $N \gtrsim \varepsilon^{-4}$ suffices, independently of the number of compositional levels in the exponent of ε. Additional levels make the gradient estimator more involved, but they do not create additional ε-scale locations at which the desired guarantee must hold. This sequential propagation of estimation error also underlies level-independent accuracy rates in statistical and sample-average analyses of compound objectives [13, 15], as well as in stochastic algorithms for nested composition problems [1, 5, 43].

Multicalibration has a different geometry. The predictor outputs a prediction vector

$$p = (p_1, \dots, p_k) \in \mathcal{P}_1 \times \cdots \times \mathcal{P}_k,$$

and calibration is imposed after conditioning on the *entire* prediction vector. Although the residuals are sequentially conditionally identifiable, the collection of prediction values over which calibration must be controlled is a Cartesian product. Discretizing each coordinate at resolution Q^{-1} therefore produces

$$|\mathcal{P}_Q| \asymp Q^k$$

prediction values. Correspondingly, the statistical error has the form

$$\mathrm{MCErr}_D^{\Gamma}(\Pi; \mathcal{G}) \lesssim \frac{1}{Q} + \sqrt{\frac{Q^k + \log|\mathcal{G}|}{n}}.$$

The difference from ordinary mean estimation can be seen directly from the second term. Suppose, for intuition, that the $M = Q^k$ prediction buckets have comparable probability. A bucket then contains about n/M observations. Its conditional residual is estimated with error of order $\sqrt{M/n}$; after multiplying by the bucket probability $1/M$, its contribution to the calibration error is of order $1/\sqrt{Mn}$. Since the calibration error sums absolute residual contributions over all M buckets, their total stochastic contribution is

$$M \cdot \frac{1}{\sqrt{Mn}} = \sqrt{\frac{M}{n}} = \sqrt{\frac{Q^k}{n}}.$$

To make the discretization error at most ε, one takes $Q \asymp \varepsilon^{-1}$. Requiring the stochastic term to be at most ε then gives

$$\sqrt{\frac{Q^k}{n}} \lesssim \varepsilon \qquad \Longrightarrow \qquad n \gtrsim \frac{Q^k}{\varepsilon^2} \asymp \varepsilon^{-(k+2)}.$$

Thus the factor ε^{-2} is the usual cost of estimating expectations to accuracy ε, while the additional factor ε^{-k} is the number of distinguishable prediction values at that resolution. Each additional calibrated level adds another prediction coordinate, multiplies the size of the joint grid by ε^{-1}, and therefore increases the exponent by one.

The lower bound confirms that this difference is intrinsic rather than an artifact of discretization. At resolution Q^{-1}, one can encode independent perturbations of size $\Theta(Q^{-1})$ across $\Theta(Q^k)$ grid points. The logarithm of the number of distinguishable alternatives is then $\Omega(Q^k)$, whereas one sample contributes only $O(Q^{-2})$ Kullback–Leibler information. Fano's inequality consequently requires

$$nQ^{-2} \gtrsim Q^k, \qquad \text{or equivalently} \qquad n \gtrsim Q^{k+2}.$$

The reduction from calibration error to prediction accuracy loses a factor of $\log Q$, so a construction at resolution Q^{-1} applies when $\varepsilon \asymp 1/(Q \log Q)$. Thus

$$Q \asymp \frac{1}{\varepsilon \log(1/\varepsilon)}, \qquad n \gtrsim \frac{\varepsilon^{-(k+2)}}{\log^{k+2}(1/\varepsilon)} = \widetilde{\Omega}\left(\varepsilon^{-(k+2)}\right).$$

In summary, stochastic composition optimization follows one nested chain and asks for stationarity at one output point, whereas multicalibration must resolve and control a k-dimensional Cartesian grid of prediction values.

B Measurability of the Online Forecaster

We justify that the prediction rules in Algorithm 1 can be chosen jointly measurably. For the fixed grid $\mathcal{P}_Q$, define the finite-dimensional residual image

$$\mathcal{K}_Q := \left\{ \big(R_j(p_{<j}, p_j, \mu)\big)_{(p,j)\in\mathcal{P}_Q\times[k]} : \mu \in \mathcal{M} \right\} \subseteq \mathbb{R}^{k|\mathcal{P}_Q|}.$$

Assumption 4.1(i) implies that $\mathcal{K}_Q$ is compact, because it is the continuous image of the compact set $\mathcal{M}$. For $w \in \Delta(\mathcal{G}\times\Sigma)$, $z = (z_g)_{g\in\mathcal{G}} \in [0,1]^{\mathcal{G}}$, $\pi \in \Delta(\mathcal{P}_Q)$, and $r = (r_{p,j})_{p,j} \in \mathcal{K}_Q$, set

$$F(\pi, r; w, z) := \sum_{p\in\mathcal{P}_Q} \pi_p \sum_{g\in\mathcal{G}} \sum_{s\in\Sigma} w(g,s) z_g \sum_{j=1}^{k} s_{p,j} r_{p,j}.$$

This function is jointly continuous on finite-dimensional compact spaces. Consequently,

$$\Phi(\pi; w, z) := \max_{r\in\mathcal{K}_Q} F(\pi, r; w, z), \qquad V(w, z) := \min_{\pi\in\Delta(\mathcal{P}_Q)} \Phi(\pi; w, z),$$

are continuous by the maximum theorem. For every $\rho \geq 0$, the correspondence

$$\mathcal{A}_\rho(w, z) := \{\pi \in \Delta(\mathcal{P}_Q) : \Phi(\pi; w, z) \leq V(w, z) + \rho\}$$

has nonempty compact values and a measurable graph. The measurable selection theorem therefore provides a Borel map $a_\rho(w, z) \in \mathcal{A}_\rho(w, z)$.

At round t, take

$$z(x) := (g(x))_{g\in\mathcal{G}}, \qquad \pi_t(\cdot \mid x) := a_\rho(w_t, z(x)).$$

The map $x \mapsto z(x)$ is measurable, and w_t is a measurable function of the cumulative gains through round $t-1$. Hence $(\text{history}, x) \mapsto \pi_t(\cdot \mid x)$ is jointly measurable and satisfies (4).