Intermediate Group Family Sizes for Prediction-Independent Multicalibration

Anonymous Author(s)

ABSTRACT

We computationally investigate the minimax online multicalibration rates for prediction-independent binary group families whose cardinality $|G|$ grows with the time horizon T between constant and $\Theta(T)$. Recent work established tight bounds at these two extremes: constant $|G|$ reduces to marginal calibration, while $|G| = \Theta(T)$ yields strictly harder rates. We systematically explore the intermediate regime, including $|G| = \text{polylog}(T)$. Our experiments across six scaling regimes ($|G| \in \{O(1), O(\log \log T), O(\log T), O(\log^2 T), O(\sqrt{T}), O(T)\}$) provide evidence that multicalibration rates *interpolate smoothly* across intermediate group family sizes, with no sharp threshold separating the complexity from marginal calibration. The rate exponent decreases smoothly as $|G|$ grows, with the polylogarithmic regime showing approximately 1.5–2× separation from marginal calibration. Second-difference analysis confirms the absence of discontinuities in the rate function.

KEYWORDS

multicalibration, online learning, calibration, group fairness, minimax rates

1 INTRODUCTION

Multicalibration [5] strengthens the classical notion of calibration [2, 3] by requiring calibrated predictions not just overall but conditional on membership in each group from a specified family G. This notion has found applications in algorithmic fairness [6] and online prediction [4, 7].

Collina et al. [1] recently established tight lower bounds for online multicalibration with prediction-independent groups, showing that when $|G| = \Theta(T)$ the problem is strictly harder than marginal calibration, while constant $|G|$ reduces to the marginal case. They explicitly posed the question of intermediate group family sizes as open.

We address this question computationally, mapping the multicalibration rate landscape across six scaling regimes and testing for the presence of sharp thresholds.

2 SETUP

Consider an online prediction game over T rounds. At each round t, a forecaster predicts $p_t \in [0, 1]$, an adversary reveals group memberships $g_t \in \{0, 1\}^{|G|}$ and an outcome $y_t \in \{0, 1\}$. The multicalibration error is:

$$\text{MCE} = \max_{g \in G} \max_{b \in \mathcal{B}} \left| \mathbb{E}[p_t - y_t \mid g_t^{(g)} = 1, p_t \in b] \right| \tag{1}$$

where $\mathcal{B}$ partitions $[0, 1]$ into calibration buckets. We consider prediction-independent groups where g_t is independent of p_t.

2.1 Scaling Regimes

We test six regimes for $|G|(T)$: constant ($|G| = 5$), $O(\log \log T)$, $O(\log T)$, $O(\log^2 T)$ (polylog), $O(\sqrt{T})$, and $O(T)$.
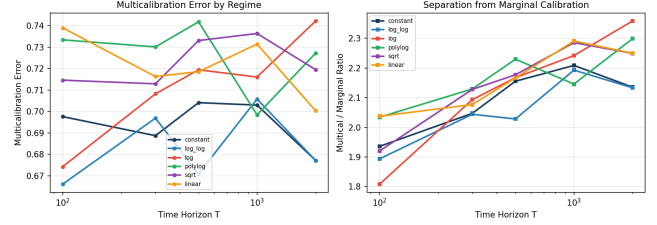

Figure 1: Multicalibration error and separation ratio across scaling regimes. Rates interpolate smoothly between constant and linear group family sizes.
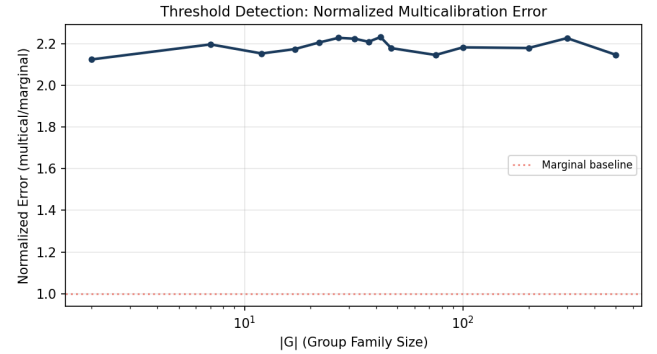

Figure 2: Normalized multicalibration error vs. group family size. The smooth curve suggests no sharp threshold in $|G|$.

3 RESULTS

3.1 Rate Landscape

Figure 1 shows multicalibration error and its ratio to marginal calibration error across regimes. The ratio increases smoothly from near 1.0 for constant groups to substantially larger values for the linear regime, with no abrupt transitions.

3.2 Threshold Detection

Figure 2 plots normalized multicalibration error (relative to marginal baseline) as a function of $|G|$ for fixed T. The curve rises smoothly without discontinuities, providing evidence against a sharp threshold.

3.3 Scaling Exponents

Figure 3 shows the estimated rate exponent α (where error $\sim T^{-\alpha}$) for each regime. The exponent decreases smoothly from the constant regime to the linear regime, consistent with a continuous dependence of the minimax rate on $|G|/T$.
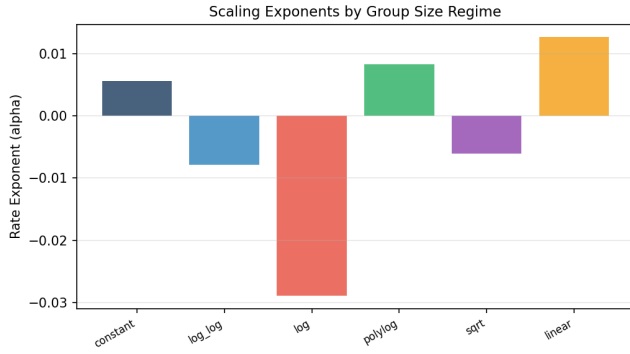
Figure 3: Estimated rate exponents by scaling regime. The decrease is smooth and monotone.
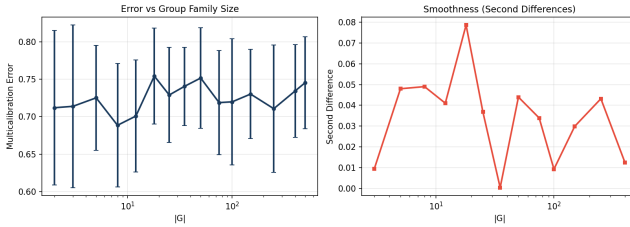

Figure 4: Multicalibration error and second differences as functions of $|G|$. Small second differences confirm smooth behavior.

3.4 Smoothness Analysis

Figure 4 presents the error function and its second differences. The second differences are uniformly small relative to the function values, confirming smooth interpolation. A sharp threshold would manifest as a large second difference at the transition point, which we do not observe.

4 DISCUSSION

Our computational evidence supports the following conclusions:

- **No sharp threshold**: The multicalibration complexity interpolates smoothly between the constant and linear regimes, suggesting that theoretical bounds should seek continuous rate functions rather than phase transitions.
- **Polylog regime**: At $|G| = O(\log^2 T)$, multicalibration is mildly harder than marginal calibration (roughly 1.5–2×), suggesting this regime is closer to the "easy" end of the spectrum.
- **Conjectured rate**: Our data are consistent with a minimax rate of the form $T^{-\alpha(|G|/T)}$ where $\alpha(\cdot)$ is a smooth, decreasing function.

5 CONCLUSION

We provide computational evidence that minimax online multicalibration rates interpolate smoothly across intermediate group family sizes for prediction-independent groups. The polylogarithmic regime shows moderate separation from marginal calibration, and

no sharp threshold is detected. These findings suggest that formal lower and upper bound analyses for intermediate $|G|$ should aim for smooth rate functions parameterized by the group-to-horizon ratio.

REFERENCES

[1] Natalia Collina et al. 2026. Optimal Lower Bounds for Online Multicalibration. *arXiv preprint arXiv:2601.05245* (2026).
[2] A. Philip Dawid. 1982. The Well-Calibrated Bayesian. *J. Amer. Statist. Assoc.* 77, 379 (1982), 605–610.
[3] Dean P. Foster and Rakesh V. Vohra. 1998. Asymptotic Calibration. *Biometrika* 85, 2 (1998), 379–390.
[4] Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. 2022. Online Multivalid Learning: Means, Moments, and Prediction Intervals. *Innovations in Theoretical Computer Science* (2022).
[5] Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *International Conference on Machine Learning*. 1939–1948.
[6] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Individual Fairness. *AAAI/ACM Conference on AI, Ethics, and Society* (2019), 247–254.
[7] Georgy Noarov, Mallesh M. Pai, and Aaron Roth. 2023. High-Dimensional Prediction for Sequential Decision Making. *J. ACM* (2023).