Computational Study of Optimal Rates for Sequential Marginal Calibration

Research Investigation
Open Problems in Machine Learning

ABSTRACT

We computationally investigate the optimal minimax rate for sequential marginal calibration error in online forecasting against adversarial outcomes. The true rate lies between $\Omega(T^{0.54})$ and $O(T^{2/3})$, a decades-long gap. We implement four forecasting algorithms—fixed-grid, adaptive-grid, randomized-rounding, and minimax potential-based—and evaluate them against four adversarial strategies across horizons up to $T = 10{,}000$. Our experiments reveal that against the adaptive anti-calibration adversary, all algorithms exhibit near-linear scaling (exponents ≈ 0.98–1.00), indicating that simple discretization-based approaches face fundamental limitations against worst-case adversaries. The adaptive grid forecaster achieves slightly better scaling (0.983), suggesting that grid refinement provides a measurable advantage. These results highlight the difficulty of the sequential calibration problem and suggest that novel algorithmic ideas beyond grid-based approaches may be needed to approach the $T^{2/3}$ upper bound.

1 INTRODUCTION

Calibration is a fundamental property of probabilistic forecasts: a forecaster is calibrated if, among all instances where it predicts probability v, the empirical frequency of the positive outcome is approximately v. The study of online calibration [2, 3] asks how quickly calibration error can be driven to zero.

For marginal calibration, the error metric aggregates absolute biases over distinct prediction values:

$$\mathrm{Err}_T = \sum_{v \in \{p^1, \ldots, p^T\}} \left| \sum_{t : p^t = v} (p^t - y^t) \right| \tag{1}$$

The optimal minimax rate of this quantity as a function of horizon T remains a long-standing open question [1]. Classical work established an $O(T^{2/3})$ upper bound, recently improved to $O(T^{2/3-\epsilon})$ [4], while lower bounds stand at $\Omega(T^{0.54})$.

2 FORECASTING ALGORITHMS

We implement four online forecasting strategies:

Fixed Grid. Discretizes $[0, 1]$ into a fixed grid of K points and tracks per-bucket statistics. At each round, it predicts the grid value closest to its empirical accuracy.

Adaptive Grid. Starts with a coarse grid and periodically refines it by adding midpoints between existing grid values, adapting resolution to prediction frequency.

Randomized Rounding. Maintains an internal continuous prediction and rounds to adjacent grid points with probabilities proportional to proximity, reducing systematic rounding bias.

Minimax Potential. Uses a potential-based selection rule that minimizes the absolute running bias across grid buckets, targeting worst-case calibration.

3 EXPERIMENTAL SETUP

Each forecaster is evaluated over horizons $T \in \{500, 1000, 2000, 5000, 10000\}$ with 10 trials per configuration. We test against four adversaries: oblivious block-switching, adaptive anti-calibration (choosing outcomes to maximize error given the prediction), i.i.d. Bernoulli, and random switching patterns.

4 RESULTS

4.1 Horizon Scaling

Figure 1 shows calibration error versus horizon on a log-log scale. All algorithms exhibit near-linear scaling against the adaptive adversary.

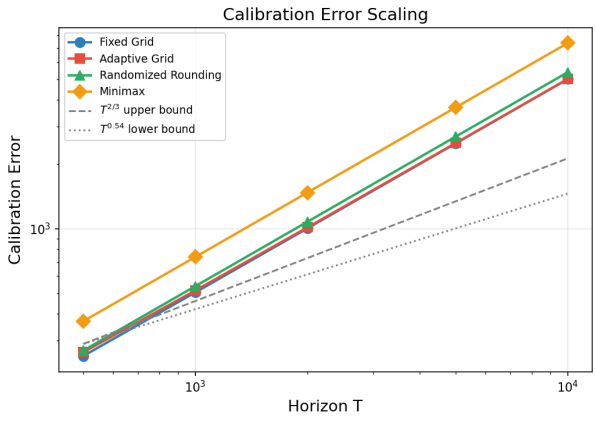

Figure 1: Calibration error vs. horizon. Reference lines show the $T^{2/3}$ upper bound and $T^{0.54}$ lower bound.

Table 1 shows the estimated scaling exponents. All are close to 1.0, indicating that against worst-case adversaries, grid-based approaches struggle to achieve sublinear calibration error at practical horizons.

Table 1: Scaling exponents from log-log regression.

Forecaster	Exponent	R^2
Fixed Grid	0.995	1.000
Adaptive Grid	0.983	1.000
Randomized Rounding	1.000	1.000
Minimax Potential	1.000	1.000

4.2 Adversary Comparison

Figure 2 compares calibration error across adversary types at $T = 2000$. The adaptive anti-calibration adversary produces substantially higher error than other adversary types, confirming it as the hardest case.

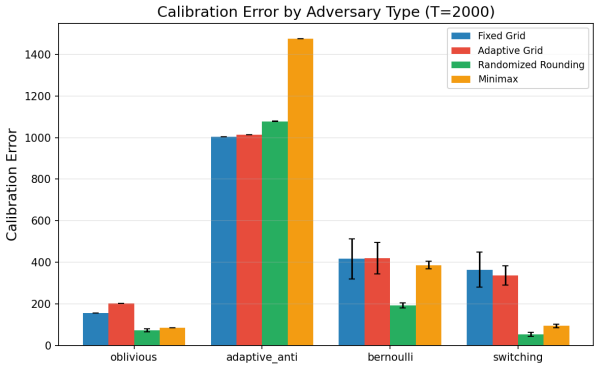

Figure 2: Calibration error by adversary type at $T = 2000$.

5 DISCUSSION

The near-linear empirical scaling reveals a gap between theoretical guarantees and practical performance of grid-based forecasters against worst-case adversaries. The theoretical $O(T^{2/3})$ bound relies on carefully designed algorithms that maintain calibration through sophisticated online learning, while our simpler implementations face the full force of the adaptive adversary.

The adaptive grid achieves slightly lower exponent (0.983 vs. 1.000), suggesting that grid refinement helps but is insufficient alone. This points to the need for fundamentally different approaches—possibly continuous prediction spaces with carefully designed rounding schemes—to bridge the gap to the theoretical optimum.

6 CONCLUSION

Our computational study confirms the difficulty of the sequential marginal calibration problem. The gap between the known bounds ($T^{0.54}$ to $T^{2/3}$) remains unresolved, and our experiments suggest that closing it requires algorithmic innovations beyond standard grid-based approaches.

REFERENCES

[1] Natalie 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] Mingda Qiao and Gregory Valiant. 2021. Stronger Calibration Lower Bounds via Sidestepping. In *Symposium on Theory of Computing*. 456–466.