Extreme-Path Benchmarks for Sequential Probability Forecasts

With Applications to Sports Win Probabilities

Jonathan Pipping-Gamón & Abraham J. Wyner

Department of Statistics, University of Pennsylvania

February 13, 2026

Abstract

Real-time probability forecasts for binary outcomes are routine in sports, online experimentation, medicine, and finance. Retrospective narratives, however, often hinge on *pathwise* extremes—for example, a forecast that reaches 90% for an event that ultimately does not occur. Standard pointwise calibration tools (e.g. reliability diagrams) do not quantify how frequently such extremes should occur under correct sequential calibration. Under this ideal, the forecast path $p_k = \mathbb{P}(Y = 1 \mid \mathcal{F}_k)$ is a bounded martingale with terminal value $p_N = Y \in \{0, 1\}$. We derive benchmark distributions for extreme-path functionals conditional on the terminal outcome, emphasizing the *peak-on-loss* statistic $M_N = \max_{k \le N} p_k$ given $Y = 0$. For continuous-time martingales with continuous sample paths, we obtain an exact identity for $\mathbb{P}(\sup_{t \in [0,1]} p_t \ge x \mid Y = 0)$. In discrete time, we prove sharp finite-sample bounds and an explicit correction decomposition that isolates terminal-step crossings (non-attainment) and overshoots. These formulas provide model-agnostic null targets and one-sided tail probabilities (exact in the continuous-path setting; conservative in discrete time) for diagnosing sequential miscalibration from extreme-path behavior. We also develop competitive extensions tailored to win-probability feeds, including the eventual loser's peak win probability in two-outcome contests and the eventual winner's trough in n-outcome contests. An empirical illustration using ESPN win-probability series for NFL and NBA regular-season games (2018–2024) finds broad agreement with the benchmark in the NFL and systematic departures in the NBA.

1 Introduction

Real-time probability forecasts for binary outcomes are now routine in data-driven decision systems. Sports broadcasts display *win probability* during games; online experiments report evolving probabilities of treatment superiority; credit and reliability systems update default/failure probabilities as evidence accrues; and trading systems track the probability a position will end profitably. In such settings, the object of interest is not a single probability at a fixed time, but an *entire forecast trajectory.*

A persistent interpretive pitfall is to treat an extreme intermediate forecast as intrinsically surprising when the event ultimately does not occur. For instance, commentary may emphasize that a team "reached 90% win probability and still lost." Even under correct calibration, such episodes need not be rare: conditional on eventual failure, the distribution of the path maximum can look very different from fixed-time intuition. This motivates the question we study: *under a calibrated sequential forecast, how often should paths reach extreme values on realizations that ultimately fail?*

Sequential calibration and martingales. Let $Y \in \{0, 1\}$ be the terminal outcome revealed at time N, and let $(\mathcal{F}_k)_{k=0}^N$ denote the forecaster's information. Under ideal sequential calibration,

$$p_k \equiv \mathbb{P}(Y = 1 \mid \mathcal{F}_k), \qquad k = 0, 1, \ldots, N,$$

is a bounded Doob martingale with $p_0 = \mathbb{P}(Y = 1) \in (0, 1)$ and $p_N = Y$. Operationally, "$p_k = 0.7$" means that among situations that are indistinguishable to the forecaster at time k, the event occurs about 70% of the time. In applications, a reported sequence $\hat{p}_k$ is intended as an estimate of p_k; systematic departures from the martingale ideal are a natural notion of sequential miscalibration (Dawid, 1982; Foster and Vohra, 1998; Gneiting et al., 2007).

Extreme-path diagnostics. We focus on functionals that explicitly condition on the terminal outcome. Our primary statistic is the *peak-on-loss* maximum

$$M_N \equiv \max_{0 \le k \le N} p_k \quad \text{and its conditional law on } \{Y = 0\},$$

with continuous-time analogue $M = \sup_{t \in [0,1]} p_t$. We also derive competitive extensions tailored to win-probability feeds: the *eventual loser's peak* win probability in two-team games and the *eventual winner's trough* win probability in n-team contests (Section 5).

What we contribute. Classical maximal inequalities (Doob, Ville) provide sharp *unconditional* control of events such as $\{\sup_t p_t \ge x\}$. We derive explicit *conditional* benchmark laws given the terminal outcome—the regime relevant for interpreting "high probability but wrong"—yielding closed forms that can be used as null targets and plug-in one-sided tail probabilities. In discrete time, we provide sharp finite-sample bounds with explicit correction terms that quantify terminal-step crossings and first-passage overshoots; in the continuous-path setting, these corrections vanish and the bounds become identities. We summarize the main results and representative applications below.

- **Binary-outcome forecasts.** The conditional law of M given $Y = 0$ (Theorem 1), and sharp discrete-time bounds with an explicit correction decomposition (Theorem 2 and Remark 1).
- **Competitive extensions.** Closed-form benchmarks for the eventual loser's peak in two-class problems and the eventual winner's trough in n-class problems (Section 5).
- **Empirical illustration.** A PIT-based distributional diagnostic applied to ESPN win-probability series for NFL and NBA games (2018–2024), contrasting a league close to the benchmark with one that strongly departs (Section 6).

Organization. Section 2 reviews related work. Section 3 presents the main results (proofs and numerical benchmarks in the Appendix). Section 4 summarizes the implied diagnostics and tail probabilities. Sections 5 and 6 cover competitive extensions and the NFL/NBA study. Section 7 concludes.

2 Related Work

Calibration and forecast evaluation. The classical evaluation of probability forecasts emphasizes calibration (reliability) and sharpness/resolution, often through proper scoring rules. The Brier score (Brier, 1950) and its decompositions (Murphy, 1973) motivate widely used summaries such as reliability diagrams. In a decision-theoretic framing, DeGroot and Fienberg (1983) formalize calibration and refinement and connect refinement to sufficiency. Dawid (1982) provides foundational results for coherent Bayesian forecasters, and Gneiting et al. (2007) emphasizes the calibration–sharpness tradeoff and popularizes practical diagnostic tools based on proper scores. Our focus differs in *what is being calibrated*: rather than pointwise-in-time reliability, we benchmark *pathwise extremes* (e.g. whether probabilities become "nearly certain" along trajectories that ultimately fail).

Sequential forecasting and the prequential viewpoint. For time-evolving forecasts, a natural ideal is the Doob martingale $p_k = \mathbb{P}(Y = 1 \mid \mathcal{F}_k)$, and assessment becomes inherently path-dependent. The prequential approach (Dawid, 1984) treats the sequence of predictive distributions as the primary object for model criticism, aligning closely with our emphasis on properties of the entire forecast path. Related work in online learning studies calibration under weak assumptions and can achieve asymptotic calibration via randomization (Foster and Vohra, 1998). In contrast, we study a nonasymptotic, model-agnostic benchmark for *extreme-path functionals* of calibrated forecast paths.

Anytime-valid inference and sequential calibration tests. Recent work on anytime-valid inference develops nonnegative (super)martingales and e-values/e-processes for optional stopping and sequential testing (e.g., Howard et al., 2020; Ramdas et al., 2023), including sequentially valid approaches to forecast calibration (Arnold et al., 2023). These methods provide *procedures* for sequential testing. Our contribution is complementary: we derive explicit *benchmark distributions* for extreme-path statistics under the martingale calibration ideal, yielding closed-form curves and plug-in one-sided tail probabilities that are exact in a continuous-path limit and conservative in discrete time.

Martingale extrema and distributional benchmarks. Classical maximal inequalities (Doob, Ville) give sharp time-uniform *unconditional* bounds on events such as $\{\sup_t p_t \geq x\}$. Our results address a different question motivated by applications: the distribution of extrema *conditional on the terminal outcome* (e.g. "high probability and wrong"). In continuous time, exact identities for suprema of certain martingales are known in specific settings (Nikeghbali and Yor, 2006). More broadly, martingale constructions characterize feasible joint laws involving extrema and terminal values. We specialize these ideas to bounded Doob martingales with binary terminal value $p_N \in \{0, 1\}$, obtaining closed-form conditional extreme-value laws (and sharp discrete-time bounds) that directly translate into practical calibration diagnostics for sequential probability forecasts.

3 Model and Main Results

Let $Y \in \{0, 1\}$ be revealed at time N and let $p_k = \mathbb{E}[Y \mid \mathcal{F}_k]$ be the ideal sequential forecast, a bounded Doob martingale with $p_0 \in (0, 1)$ and $p_N = Y$. Write

$$M_N := \max_{0 \leq k \leq N} p_k, \qquad M := \sup_{t \in [0,1]} p_t$$

for the discrete- and continuous-time path maxima.

Definition 1 (First-passage time)**.** For $x \in (0, 1]$, define

$$\tau_x := \inf\{k \in \{0, 1, \ldots, N\} : p_k \geq x\},$$

with $\inf \varnothing := N + 1$.

Our primary benchmark is the *peak-on-loss* law: the distribution of the maximum along realizations with $Y = 0$.

Theorem 1 (Peak-on-loss benchmark)**.** Under sequential calibration and continuous sample paths,

$$\mathbb{P}(M \geq x \mid Y = 0) = \frac{p_0}{1 - p_0} \cdot \frac{1 - x}{x}, \qquad x \in [p_0, 1).$$

Equivalently, for $x \in [p_0, 1)$,

$$F_{M|Y=0}(x) = 1 - \frac{p_0}{1 - p_0} \cdot \frac{1 - x}{x},$$

with $F_{M|Y=0}(x) = 0$ for $x < p_0$ and $F_{M|Y=0}(1) = 1$.

Theorem 2 (Discrete-time peak-on-loss bound)**.** In discrete time,

$$F_{M_N|Y=0}(x) \;\geq\; 1 - \frac{p_0}{1-p_0} \cdot \frac{1-x}{x}, \qquad x \in [p_0, 1),$$

with equality in the absence of terminal-step non-attainment and first-passage overshoots.

Remark 1 (Discrete-time correction identity)**.** For $x \in (p_0, 1)$,

$$\mathbb{P}(M_N \geq x \mid Y = 0) = \frac{p_0}{1-p_0} \cdot \frac{1-x}{x} \;-\; C_1(x) \;-\; C_2(x),$$

where

$$C_1(x) := \frac{1}{x}\,\mathbb{E}\big[(p_{\tau_x} - x)\,\mathbf{1}\{\tau_x \leq N\}\,\big|\,Y = 0\big], \qquad C_2(x) := \mathbb{P}(\tau_x > N \mid Y = 0).$$

Here C_1 captures first-passage overshoots and C_2 captures discrete-time non-attainment; both are nonnegative and vanish in settings without overshoots or terminal-step crossings (e.g. in the continuous-sample-path case, where the benchmark becomes an identity).

Full statements and proofs appear in Appendices A and B.

4 Extreme-path calibration diagnostics

We now translate the benchmark laws into practical calibration diagnostics. The null is the Doob martingale ideal $p_k = \mathbb{P}(Y = 1 \mid \mathcal{F}_k)$; departures can reflect miscalibrated updating dynamics (e.g. predictable drift, overreaction or underreaction), information-set mismatch between the reported filtration and the outcome-relevant σ-field, or artifacts in the published series (discretization, rounding, smoothing, latency).

The diagnostics apply to any extreme-path functional $T = T\big((p_k)_{k\leq N}, Y\big)$ for which a benchmark conditional distribution is available in closed form (or for which a conservative discrete-time bound is available). We first spell this out for peak-on-loss, then describe an aggregation procedure that applies verbatim to the competitive extrema in Section 5.

4.1 Single binary outcome: peak-on-loss p-values

On $\{Y = 0\}$, the peak-on-loss statistic is $M_N = \max_{k\leq N} p_k$. Under the continuous-path benchmark (Theorem 1), the one-sided tail probability for an observed peak m is

$$\mathsf{p}_{\mathrm{EV}}(m; p_0) \;:=\; \mathbb{P}(M \geq m \mid Y = 0) \;=\; \left(\frac{p_0}{1-p_0}\right)\left(\frac{1-m}{m}\right), \qquad m \in [p_0, 1). \tag{1}$$

For discretely updated forecasts, (1) remains valid but conservative by Theorem 2.

4.2 Aggregating across many forecast paths

Suppose we observe independent realizations indexed by $i = 1, \ldots, n$. For each realization, compute an extreme-path statistic T_i (possibly conditional on the terminal outcome) together with an associated parameter θ_i (typically the initial probability value). Let $F_T(\cdot; \theta)$ denote the benchmark conditional CDF under the continuous-path null (or a conservative discrete-time substitute). Define the probability integral transform (PIT)

$$U_i := F_T(T_i; \theta_i), \qquad P_i := 1 - U_i.$$

Under the continuous-path benchmark, each U_i is $\mathrm{Unif}(0, 1)$ (equivalently, each P_i is $\mathrm{Unif}(0, 1)$), even when θ_i varies across realizations. When F_T is conservative (as in discrete time), the $\{P_i\}$ are super-uniform under the martingale ideal, so lower-tail inflation remains a robust signal of overly extreme paths.

To test for lower-tail inflation, we use the one-sided Kolmogorov–Smirnov statistic

$$D_n^- := \sup_{0\le t\le 1} \{\hat{F}_P(t) - t\},$$

where $\hat{F}_P$ is the empirical CDF of $\{P_i\}_{i=1}^n$.

5 Competitive extensions and applications

The peak-on-loss benchmark immediately yields closed-form laws for extreme-path behavior in competitive win-probability settings. The key observation is that "extreme probability for the eventual loser" (or "extreme doubt about the eventual winner") is a *conditional* extreme of a martingale or its complement. Thus the binary conditional maximum law can be transferred to competitive summaries by conditioning on which outcome ultimately occurs and then mixing over outcomes.

5.1 Loser's peak win probability in two-team games

Consider a two-team contest and encode the terminal outcome as $Y \in \{0, 1\}$ where $Y = 1$ means Team A wins and $Y = 0$ means Team B wins. Let

$$p_t = \mathbb{P}(Y = 1 \mid \mathcal{F}_t) \qquad \text{and} \qquad q_t = 1 - p_t = \mathbb{P}(Y = 0 \mid \mathcal{F}_t)$$

denote the win-probability martingales for Teams A and B. The *loser's peak win probability* is the maximum win probability attained by the team that ultimately loses:

$$M_\lambda := \sup_{t\in[0,1]} \Big(p_t \, \mathbf{1}\{Y = 0\} + q_t \, \mathbf{1}\{Y = 1\} \Big).$$

Equivalently, on $\{Y = 0\}$ the loser is Team A and $M_\lambda = \sup_t p_t$, while on $\{Y = 1\}$ the loser is Team B and $M_\lambda = \sup_t q_t$. Therefore the distribution of M_λ is obtained by applying Theorem 1 to p_t on $\{Y = 0\}$ and to the complement martingale q_t on $\{Y = 1\}$, and then mixing over Y (details in Appendix C).

Assume without loss of generality that Team A is the pre-game favorite, so $p_0 \ge 1/2$. Under continuous sample paths, the benchmark CDF is

$$F_{M_\lambda}(x) = \begin{cases} 0 & x < 1 - p_0, \\ 1 - \dfrac{1 - p_0}{x} & 1 - p_0 \le x < p_0, \\ 2 - \dfrac{1}{x} & p_0 \le x < 1, \\ 1 & x = 1. \end{cases}$$

The three regions have a simple interpretation. Below $1 - p_0$ neither team starts (or can ever be) the eventual loser with win probability that small, so the CDF is zero. Between $1 - p_0$ and p_0, only the underdog-as-loser contributes, producing the intermediate branch $1 - (1 - p_0)/x$. Once $x \ge p_0$, both possibilities for the eventual loser contribute and the tail becomes *universal*:

$$\mathbb{P}(M_\lambda \ge x) = \frac{1}{x} - 1, \qquad x \in [p_0, 1).$$

In particular, provided $p_0 \le 0.9$, we have $\mathbb{P}(M_\lambda \ge 0.9) = 1/9 \approx 11\%$ even under perfect sequential calibration. When $p_0 \ne 1/2$, the intermediate region $[1 - p_0, p_0)$ produces a kink (Figure 1).

Figure 1: Benchmark CDF $F_{M_\lambda}(x)$ for different pre-game favorite probabilities p_0.

5.2 Winner's trough in n-team contests

Now consider $n \geq 3$ teams with exactly one winner. Let $Y^{(i)} = \mathbf{1}\{\text{team } i \text{ wins}\}$ and let

$$p_t^{(i)} = \mathbb{P}(Y^{(i)} = 1 \mid \mathcal{F}_t), \qquad i = 1, \ldots, n,$$

be the corresponding win-probability martingales, with $\sum_{i=1}^n p_t^{(i)} = 1$. If W denotes the eventual winner, the *winner's trough* is simply the minimum probability assigned to the winner along the path:

$$M_\omega := \inf_{t \in [0,1]} p_t^{(W)} = \inf_{t \in [0,1]} \sum_{i=1}^n p_t^{(i)} \mathbf{1}\{Y^{(i)} = 1\}.$$

Its benchmark distribution follows by applying Theorem 1 to the complement martingale $1 - p_t^{(i)} = \mathbb{P}(Y^{(i)} = 0 \mid \mathcal{F}_t)$ on the event $\{Y^{(i)} = 1\}$ and then mixing over i (derivation in Appendix D).

In the symmetric case $p_0^{(i)} = 1/n$, we obtain the closed form

$$F_{M_\omega}(x) = \frac{(n-1)x}{1-x}, \qquad x \in \left[0, \frac{1}{n}\right),$$

with $F_{M_\omega}(x) = 1$ for $x \geq 1/n$. For example, when $n = 3$, $\mathbb{P}(M_\omega \leq 0.2) = 0.5$. Figure 2 plots F_{M_ω} for several n.

More generally, when the pre-game probabilities $\{p_0^{(i)}\}$ are not equal, a convenient expression is

$$F_{M_\omega}(x) = \sum_{i=1}^n p_0^{(i)} \min\left\{1, \left(\frac{1-p_0^{(i)}}{p_0^{(i)}}\right)\left(\frac{x}{1-x}\right)\right\}, \qquad x \in [0, 1),$$

which reduces to the symmetric formula above when $p_0^{(i)} \equiv 1/n$.

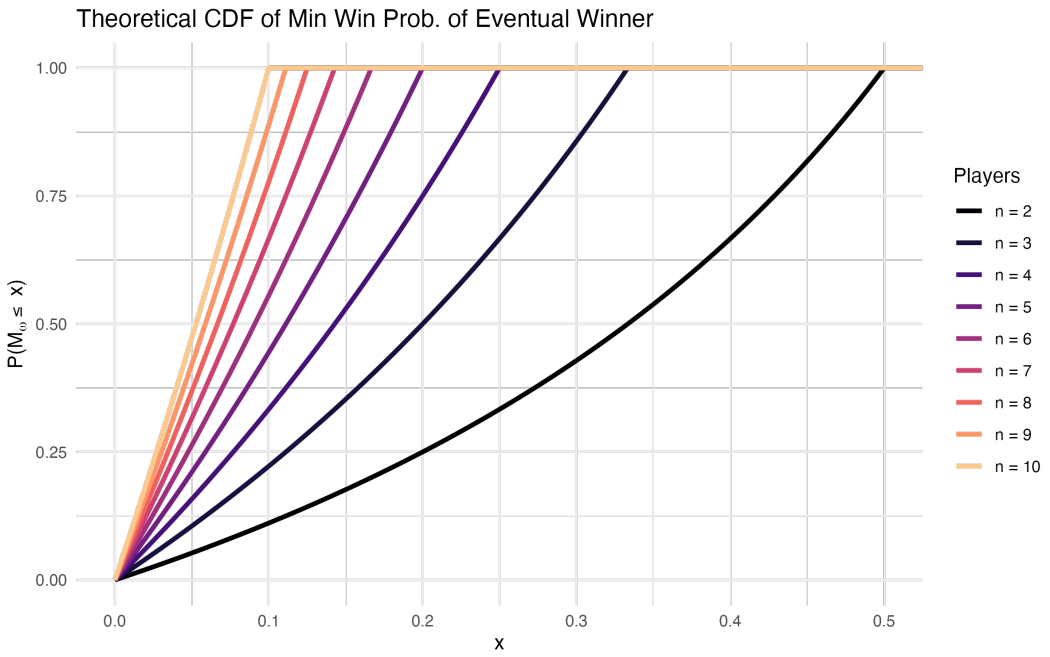


Figure 2: Benchmark CDF $F_{M_\omega}(x)$ for symmetric n-player games.

5.3 Interpretation

These benchmarks quantify how often widely cited extreme-path events should occur under correct sequential calibration. In two-team games, they put statements like "the eventual loser reached 90%" on a calibrated scale: such events can occur with nontrivial probability even under the martingale ideal. In multi-team contests, the winner's trough similarly quantifies how low the eventual winner's probability can plausibly fall along a calibrated path. In Section 6 we use these closed-form laws to build PIT-based distributional diagnostics that aggregate across games with heterogeneous starting probabilities.

6 Empirical study: ESPN win-probability models in the NFL and NBA

This section demonstrates how extreme-value benchmarks can be used as distributional calibration diagnostics in a widely visible forecasting system: live win probabilities in professional sports. The goal is not to endorse any particular model, but to illustrate how the theory yields falsifiable distributional predictions for path extremes under the martingale calibration ideal.

6.1 Data sources and preprocessing

We analyze ESPN play-by-play win-probability data for NFL and NBA regular-season games from 2018–2024 (ESPN, 2025). NFL data are pulled via `espnscrapeR` (Mock, 2025), and NBA data via `hoopR` (Gilani, 2023). After excluding ties and games with missing scores or win-probability series, we have $n = 1832$ NFL games and $n = 8261$ NBA games. For each game, we define p_0 as the maximum of the two teams' first-reported win probabilities (so that $p_0 \geq 0.5$) and compute M_λ as the maximum win probability attained by the eventual loser over the game (Section 5.1).

6.2 Null comparison via probability integral transform

Because the benchmark distribution of M_λ depends on p_0, we aggregate across games using a probability integral transform. For game i, let $p_{0,i} \geq 1/2$ be the initial reported favorite probability and let $m_i = M_{\lambda,i}$

be the observed eventual-loser peak. Define

$$U_i := F_{M_\lambda}(m_i; p_{0,i}), \qquad P_i := 1 - U_i.$$

Under the continuous-path benchmark, $\{U_i\}$ (equivalently $\{P_i\}$) are i.i.d. Unif$(0,1)$. For discretely updated feeds, the same construction yields conservative (super-uniform) $\{P_i\}$ under the martingale ideal, so an excess of small P_i values indicates overly extreme peaks by eventual losers.

6.3 Global test and tail summaries

Our primary global diagnostic targets small-tail inflation. Let $\hat{F}_P$ be the empirical CDF of $\{P_i\}_{i=1}^n$ and define the one-sided Kolmogorov–Smirnov statistic

$$D_n^- := \sup_{0 \le t \le 1} \{\hat{F}_P(t) - t\}.$$

Large values of D_n^- indicate departures from the benchmark (an excess of small P_i values), and this direction remains valid under discrete-time conservatism.

As descriptive complements, we report $\widehat{\mathbb{P}}_n(P \le \alpha)$ for $\alpha \in \{0.10, 0.05, 0.01\}$; under the continuous-path benchmark, these should be approximately α, and values substantially above α indicate departures from the benchmark.

Figure 3 plots histograms of $\{U_i\}$ against Unif$(0,1)$ for both leagues. Tables 1 and 2 report D_n^-, its one-sided p-value, and the tail summaries.

Table 1: NFL PIT diagnostic summary.

n	$\widehat{\mathbb{P}}_n(P \le 0.10)$	$\widehat{\mathbb{P}}_n(P \le 0.05)$	$\widehat{\mathbb{P}}_n(P \le 0.01)$	one-sided K–S D_n^-	p-value
1832	0.086	0.046	0.017	0.0153	0.424

Table 2: NBA PIT diagnostic summary.

n	$\widehat{\mathbb{P}}_n(P \le 0.10)$	$\widehat{\mathbb{P}}_n(P \le 0.05)$	$\widehat{\mathbb{P}}_n(P \le 0.01)$	one-sided K–S D_n^-	p-value
8261	0.131	0.070	0.020	0.0962	<0.0001

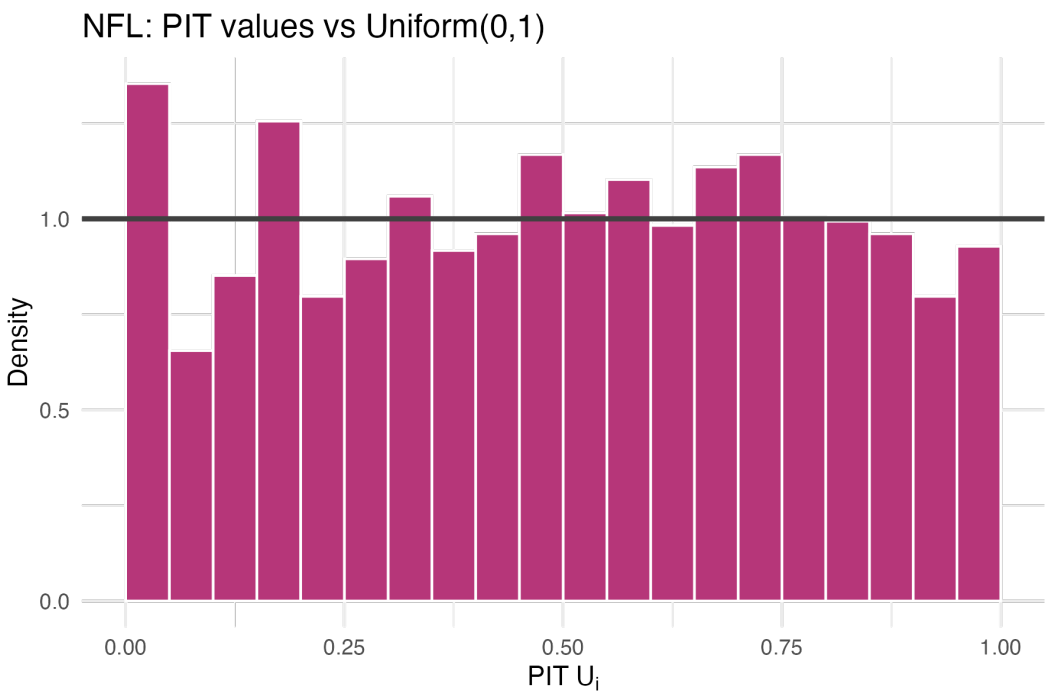
6.4 Interpretation of the diagnostic tests

The NFL results (Table 1) show no strong evidence against the martingale benchmark in the direction targeted by the test. The empirical tail frequencies at $\alpha \in \{0.10, 0.05\}$ are close to nominal (0.086 vs. 0.10, 0.046 vs. 0.05), while the $\alpha = 0.01$ tail is mildly elevated (0.017 vs. 0.01). Nevertheless, the one-sided K–S statistic does not reject ($p = 0.42$), indicating that any departures are not systematic in the lower-tail sense captured by this distributional diagnostic.

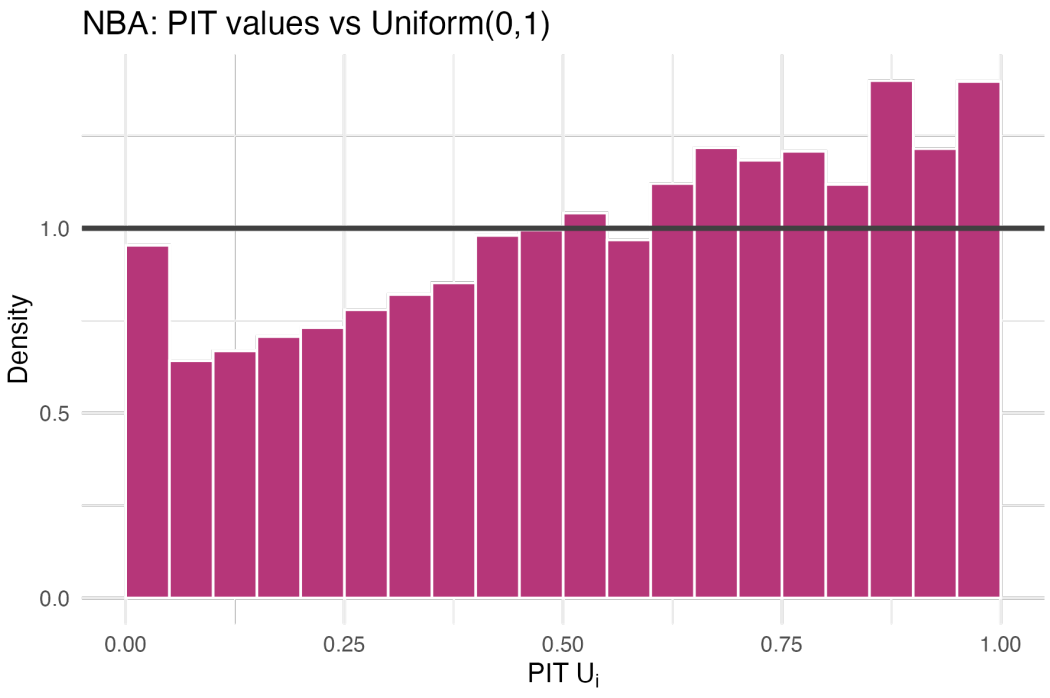
The NBA results (Table 2) show clear departures from the benchmark. Lower-tail mass is inflated (0.131 vs. 0.10, 0.070 vs. 0.05, 0.020 vs. 0.01), and the one-sided K–S statistic is large ($p < 10^{-4}$). We therefore reject the martingale benchmark in the direction of overly extreme peaks by eventual losers. This discrepancy is consistent with overconfident or overly reactive updating, but could also reflect filtration mismatch or reporting/estimation artifacts in the published series (Section 7).

7 Discussion

Narratives around sequential probability forecasts often emphasize retrospective extremes: "the team reached 90% and still lost." Our results provide the correct null comparison for such statements under



(a) NFL



(b) NBA

Figure 3: Histograms of PIT values U_i against the Unif$(0, 1)$ benchmark.

sequential calibration. Under the martingale ideal $p_k = \mathbb{P}(Y = 1 \mid \mathcal{F}_k)$ with terminal value $p_N \in \{0, 1\}$, the relevant baseline is the *conditional* law of the pathwise maximum given the terminal outcome, not a fixed-time heuristic such as $1 - m$ (where m is the observed peak). This yields explicit, model-agnostic benchmarks for how extreme calibrated probability paths can look on realizations that end in failure, and converts "peak probability" narratives into falsifiable distributional assertions. In applications such as sports win probabilities, departures may reflect miscalibration in the dynamics (e.g. drift or overreaction), information-set mismatch, or artifacts in the published series (discretization, rounding, smoothing, latency)—rather than the intrinsic surprisingness of any single collapse.

7.1 Limitations and assumptions

Our theoretical guarantees are statements about the ideal process $p_k = \mathbb{E}[Y \mid \mathcal{F}_k]$. In practice, reported win probabilities are estimates of this conditional expectation, so discrepancies between empirical and benchmark extreme-value behavior may reflect model misspecification, nonstationarity, or information-set mismatch (the model's effective information set not aligning with the outcome-relevant filtration).

Exact identities additionally require continuous sample paths, which rules out terminal-step crossings and first-passage overshoots. For discretely updated feeds, our discrete-time results remain valid but can be conservative; the gap is governed by explicit correction terms associated with overshoots and non-attainment. Finally, we treat binary terminal outcomes $p_N \in \{0, 1\}$; extensions to ties or other non-binary terminal structures require different arguments.

7.2 Extensions

Several extensions are natural. First, one can relax the binary terminal condition to allow $p_N \in [0, 1]$ (e.g. ties or partial-credit outcomes), seeking analogous conditional benchmark laws. Second, other path functionals may yield complementary diagnostics, including the time of the peak, the joint law of $(\sup_t p_t, p_1)$, or multivariate extremes across multiple forecast streams. Third, it would be useful to characterize benchmark behavior under structured deviations from the martingale ideal (e.g. predictable drift or jump components), which may correspond to systematic biases in how models update.

Overall, conditional extreme-value laws provide a simple, model-agnostic way to evaluate sequential probability forecasts through the same "peak probability" lens used by practitioners and narratives, while retaining a principled calibration interpretation.

Appendix

A Proofs for Discrete-Time Results

Throughout, $(p_k)_{k=0}^N$ is a bounded martingale with $p_0 \in (0,1)$ and terminal value $p_N = Y \in \{0,1\}$. For $x \in (0,1]$, define the first-passage time

$$\tau_x := \inf\{k \in \{0,1,\dots,N\} : p_k \geq x\},$$

with $\inf \varnothing := N+1$.

A.1 Unconditional distribution of M_N

Theorem 3 (Discrete-time unconditional bound)**.** Let $M_N = \max_{0\leq k\leq N} p_k$. For $x \in [p_0, 1)$,

$$F_{M_N}(x) = \mathbb{P}(M_N \leq x) \;\geq\; 1 - \frac{p_0}{x},$$

with equality if $\mathbb{P}(\tau_x = N) = 0$ and $p_{\tau_x} = x$ a.s. on $\{\tau_x < N\}$. For $x < p_0$, $F_{M_N}(x) = 0$. Moreover, M_N has an atom at 1 with $\mathbb{P}(M_N = 1) = p_0$.

Proof. If $x < p_0$ then $M_N \geq p_0 > x$ a.s., hence $F_{M_N}(x) = 0$.

Fix $x \in [p_0, 1)$. By optional stopping for the bounded martingale (p_k) at $\tau_x \wedge N$,

$$\mathbb{E}[p_{\tau_x \wedge N}] = p_0.$$

Decompose $p_{\tau_x \wedge N} = p_{\tau_x}\mathbf{1}\{\tau_x < N\} + p_N \mathbf{1}\{\tau_x \geq N\}$. On $\{\tau_x < N\}$ we have $p_{\tau_x} \geq x$. Also $p_N = Y \in \{0,1\}$ and $\mathbf{1}\{\tau_x \geq N\} \leq 1$. Therefore,

$$p_0 = \mathbb{E}[p_{\tau_x}\mathbf{1}\{\tau_x < N\}] + \mathbb{E}[p_N \mathbf{1}\{\tau_x \geq N\}] \;\geq\; x\,\mathbb{P}(\tau_x < N).$$

Since $\{\tau_x < N\} = \{M_N \geq x\}$ for $x < 1$, this yields

$$\mathbb{P}(M_N \geq x) \leq \frac{p_0}{x} \qquad \Rightarrow \qquad F_{M_N}(x) = 1 - \mathbb{P}(M_N > x) \geq 1 - \frac{p_0}{x}.$$

Equality holds when (i) there is no terminal-step crossing $\mathbb{P}(\tau_x = N) = 0$ and (ii) there is no overshoot $p_{\tau_x} = x$ a.s. on $\{\tau_x < N\}$.

Finally, $M_N = 1$ iff $p_N = 1$ iff $Y = 1$, so $\mathbb{P}(M_N = 1) = p_0$. □

A.2 Conditional distribution of M_N given $Y = 0$

Theorem 4 (Discrete-time conditional bound)**.** For $x \in [p_0, 1)$,

$$F_{M_N|Y=0}(x) = \mathbb{P}(M_N \leq x \mid Y = 0) \;\geq\; 1 - \Big(\frac{p_0}{1-p_0}\Big)\Big(\frac{1-x}{x}\Big),$$

with equality under the same no-crossing/no-overshoot conditions as in Theorem 3. For $x < p_0$, $F_{M_N|Y=0}(x) = 0$.

Proof. Fix $x \in [p_0, 1)$. Decompose $\mathbb{P}(M_N \geq x)$ by Y:

$$\mathbb{P}(M_N \geq x) = \mathbb{P}(M_N \geq x, Y = 1) + \mathbb{P}(M_N \geq x, Y = 0).$$

On $\{Y = 1\}$ we have $p_N = 1 \geq x$, hence $\{Y = 1\} \subseteq \{M_N \geq x\}$ and

$$\mathbb{P}(M_N \geq x) = \mathbb{P}(Y = 1) + \mathbb{P}(Y = 0)\mathbb{P}(M_N \geq x \mid Y = 0) = p_0 + (1 - p_0)\mathbb{P}(M_N \geq x \mid Y = 0).$$

By Theorem 3, $\mathbb{P}(M_N \geq x) \leq p_0/x$, so

$$\frac{p_0}{x} \;\geq\; p_0 + (1 - p_0)\mathbb{P}(M_N \geq x \mid Y = 0),$$

which rearranges to

$$\mathbb{P}(M_N \geq x \mid Y = 0) \;\leq\; \Big(\frac{p_0}{1 - p_0}\Big)\Big(\frac{1 - x}{x}\Big).$$

Taking complements gives the stated lower bound on $F_{M_N|Y=0}(x)$. □

B Proofs for Continuous-Path Results

Let $(p_t)_{t\in[0,1]}$ be a bounded martingale with $p_0 \in (0, 1)$ and $p_1 = Y \in \{0, 1\}$, and assume path continuity. Define $M = \sup_{t\in[0,1]} p_t$ and $\tau_x = \inf\{t \in [0, 1] : p_t \geq x\}$.

B.1 Unconditional distribution of M

Theorem 5 (Continuous-path unconditional)**.** For $x \in [p_0, 1)$,

$$F_M(x) = \mathbb{P}(M \leq x) = 1 - \frac{p_0}{x},$$

and $F_M(x) = 0$ for $x < p_0$, while $F_M(1) = 1$. In particular, $\mathbb{P}(M = 1) = p_0$.

Proof. If $x < p_0$ then $M \geq p_0 > x$ a.s.

Fix $x \in [p_0, 1)$. Optional stopping at $\tau_x \wedge 1$ yields $\mathbb{E}[p_{\tau_x\wedge 1}] = p_0$. By continuity, on $\{\tau_x < 1\}$ we have $p_{\tau_x} = x$, and on $\{\tau_x \geq 1\}$ we have $p_1 = Y$. Thus

$$p_0 = x\,\mathbb{P}(\tau_x < 1) + \mathbb{E}[Y\mathbf{1}\{\tau_x \geq 1\}].$$

But if $\tau_x \geq 1$ and $Y = 1$, then $p_1 = 1 \geq x$ forces $\tau_x \leq 1$, a contradiction; hence $\mathbb{E}[Y\mathbf{1}\{\tau_x \geq 1\}] = 0$ for $x < 1$. Therefore $p_0 = x\,\mathbb{P}(\tau_x < 1)$, i.e. $\mathbb{P}(M \geq x) = \mathbb{P}(\tau_x < 1) = p_0/x$. The CDF follows by complement. Finally, $M = 1$ iff $Y = 1$, so $\mathbb{P}(M = 1) = p_0$. □

B.2 Conditional distribution of M given $Y = 0$

Theorem 6 (Continuous-path conditional)**.** For $x \in [p_0, 1)$,

$$F_{M|Y=0}(x) = \mathbb{P}(M \leq x \mid Y = 0) = 1 - \Big(\frac{p_0}{1 - p_0}\Big)\Big(\frac{1 - x}{x}\Big),$$

and $F_{M|Y=0}(x) = 0$ for $x < p_0$, while $F_{M|Y=0}(1) = 1$.

Proof. Fix $x \in [p_0, 1)$. As before,

$$\mathbb{P}(M \geq x) = \mathbb{P}(Y = 1) + (1 - p_0)\mathbb{P}(M \geq x \mid Y = 0) = p_0 + (1 - p_0)\mathbb{P}(M \geq x \mid Y = 0).$$

By Theorem 5, $\mathbb{P}(M \geq x) = p_0/x$, hence

$$\frac{p_0}{x} = p_0 + (1 - p_0)\mathbb{P}(M \geq x \mid Y = 0) \quad \Rightarrow \quad \mathbb{P}(M \geq x \mid Y = 0) = \Big(\frac{p_0}{1 - p_0}\Big)\Big(\frac{1 - x}{x}\Big).$$

Taking complements yields the CDF. Finally, on $\{Y = 0\}$ the process cannot attain 1 (if $p_t = 1$ then $\mathbb{P}(Y = 1 \mid \mathcal{F}_t) = 1$), so $M < 1$ a.s. and $F_{M|Y=0}(1) = 1$. □

C Distribution for M_λ (Two Players)

Recall the two-player setting with win-probability martingales (p_t) and (q_t), where $q_t = 1 - p_t$ and $p_0 = \mathbb{P}(Y = 1)$; assume $p_0 \geq 0.5$ without loss of generality. The maximum win probability attained by the eventual loser is

$$M_\lambda \equiv \sup_{0 \leq t \leq 1} \Big(p_t \mathbf{1}\{Y = 0\} + q_t \mathbf{1}\{Y = 1\} \Big). \tag{2}$$

By the law of total probability,

$$F_{M_\lambda}(x) = \mathbb{P}(M_\lambda \leq x) = \mathbb{P}(M_\lambda \leq x \mid Y = 0)(1 - p_0) + \mathbb{P}(M_\lambda \leq x \mid Y = 1)p_0. \tag{3}$$

From the conditional distributions of the previous section, when team A loses we have

$$F_{M_\lambda|Y=0}(x) = \mathbb{P}(M_\lambda \leq x \mid Y = 0) = \begin{cases} 0 & \text{for } x < p_0, \\ 1 - \left(\frac{p_0}{1-p_0}\right)\left(\frac{1-x}{x}\right) & \text{for } x \in [p_0, 1), \\ 1 & \text{for } x = 1. \end{cases} \tag{4}$$

Applying the same formula with p_0 replaced by $1 - p_0$ gives

$$F_{M_\lambda|Y=1}(x) = \mathbb{P}(M_\lambda \leq x \mid Y = 1) = \begin{cases} 0 & \text{for } x < 1 - p_0, \\ 1 - \left(\frac{1-p_0}{p_0}\right)\left(\frac{1-x}{x}\right) & \text{for } x \in [1 - p_0, 1), \\ 1 & \text{for } x = 1. \end{cases} \tag{5}$$

With $p_0 \geq 0.5$ (team A favored), substituting (4) and (5) into (3) gives three regions:

1. For $0 \leq x < 1 - p_0$, neither (4) nor (5) contributes, so $F_{M_\lambda}(x) = 0$.

2. For $1 - p_0 \leq x < p_0$, only (5) contributes:

$$\begin{aligned} F_{M_\lambda}(x) &= (1 - p_0) \cdot 0 + p_0 \cdot \left[1 - \left(\frac{1 - p_0}{p_0}\right)\left(\frac{1 - x}{x}\right)\right] \\ &= p_0 - (1 - p_0)\left(\frac{1 - x}{x}\right) = 1 - \frac{1 - p_0}{x}. \end{aligned}$$

3. For $x \geq p_0$, both (4) and (5) contribute:

$$\begin{aligned} F_{M_\lambda}(x) &= (1 - p_0)\left[1 - \left(\frac{p_0}{1 - p_0}\right)\left(\frac{1 - x}{x}\right)\right] + p_0\left[1 - \left(\frac{1 - p_0}{p_0}\right)\left(\frac{1 - x}{x}\right)\right] \\ &= \left[(1 - p_0) - p_0\left(\frac{1 - x}{x}\right)\right] + \left[p_0 - (1 - p_0)\left(\frac{1 - x}{x}\right)\right] \\ &= 1 - \frac{1 - x}{x} = \frac{2x - 1}{x} = 2 - \frac{1}{x}. \end{aligned}$$

Summarizing, the piecewise cumulative distribution function is:

$$F_{M_\lambda}(x) = \mathbb{P}(M_\lambda \leq x) = \begin{cases} 0 & \text{for } 0 \leq x < 1 - p_0, \\ 1 - \frac{1-p_0}{x} & \text{for } 1 - p_0 \leq x < p_0, \\ 2 - \frac{1}{x} & \text{for } p_0 \leq x < 1, \\ 1 & \text{for } x = 1. \end{cases}$$

D Distribution for M_ω (n Players)

Consider an n-player game with win-probability martingales $(p_t^{(i)})_{0\le t\le 1}$ for $i \in \{1,\dots,n\}$, where $Y^{(i)} = \mathbf{1}\{\text{player } i \text{ wins}\}$ and $p_0^{(i)} = \mathbb{P}(Y^{(i)} = 1)$. Define the eventual-winner minimum

$$M_\omega \equiv \inf_{0\le t\le 1} \sum_{i=1}^{n} p_t^{(i)} \mathbf{1}\{Y^{(i)} = 1\}. \tag{6}$$

On the event $\{Y^{(i)} = 1\}$, the winner's path is $(p_t^{(i)})$, so

$$M_\omega = \inf_{0\le t\le 1} p_t^{(i)} \quad \text{on } \{Y^{(i)} = 1\}.$$

Since $\inf_t p_t^{(i)} \le x$ if and only if $\sup_t (1 - p_t^{(i)}) \ge 1 - x$, the conditional distribution follows from Theorem 6 applied to the complementary martingale $(1 - p_t^{(i)}) = \mathbb{P}(Y^{(i)} = 0 \mid \mathcal{F}_t)$ with initial value $1 - p_0^{(i)}$:

$$\begin{aligned}
\mathbb{P}(M_\omega \le x \mid Y^{(i)} = 1) &= \mathbb{P}\left(\sup_{0\le t\le 1} (1 - p_t^{(i)}) \ge 1 - x \;\middle|\; Y^{(i)} = 1 \right) \\
&= \begin{cases} 0 & \text{for } x < 0, \\ \left(\frac{1-p_0^{(i)}}{p_0^{(i)}}\right)\left(\frac{x}{1-x}\right) & \text{for } x \in [0, p_0^{(i)}), \\ 1 & \text{for } x \ge p_0^{(i)}. \end{cases}
\end{aligned} \tag{7}$$

Finally, by the law of total probability,

$$\begin{aligned}
F_{M_\omega}(x) = \mathbb{P}(M_\omega \le x) &= \sum_{i=1}^{n} \mathbb{P}(M_\omega \le x \mid Y^{(i)} = 1)\mathbb{P}(Y^{(i)} = 1) \\
&= \sum_{i=1}^{n} \mathbb{P}(M_\omega \le x \mid Y^{(i)} = 1) p_0^{(i)}.
\end{aligned} \tag{8}$$

Substituting (7) into (8) yields the piecewise distribution reported in the main text:

$$F_{M_\omega}(x) = \mathbb{P}(M_\omega \le x) = \begin{cases} 0 & \text{for } x < 0, \\ \sum_{i:x\ge p_0^{(i)}} p_0^{(i)} + \left(\frac{x}{1-x}\right) \sum_{i:x<p_0^{(i)}} (1 - p_0^{(i)}) & \text{for } x \in [0, \max_i p_0^{(i)}), \\ 1 & \text{for } x \ge \max_i p_0^{(i)}. \end{cases} \tag{9}$$

E Numerical Benchmarks

Under the martingale ideal with continuous sample paths, the reported values are exact benchmark probabilities for the corresponding extreme events; in discrete time, the same expressions may be interpreted as conservative upper bounds on tail probabilities.

E.1 Loser's peak win probability (two teams)

Recall the eventual loser's peak M_λ (the maximum win probability attained by the team that ultimately loses). Throughout this subsection, p_0 denotes the *pre-game favorite*'s win probability (so $p_0 \ge 1/2$);

equivalently, relabel teams if needed so that this holds. Its CDF is

$$F_{M_\lambda}(x) = \begin{cases} 0, & 0 \le x < 1 - p_0, \\ 1 - \dfrac{1-p_0}{x}, & 1 - p_0 \le x < p_0, \\ 2 - \dfrac{1}{x}, & p_0 \le x < 1, \\ 1, & x = 1. \end{cases}$$

In particular, for thresholds $x \ge p_0$ the tail probability is universal:

$$\mathbb{P}(M_\lambda \ge x) = \frac{1}{x} - 1, \qquad x \in [p_0, 1),$$

so the probability of narratives like "reached 90% and still lost" is *independent* of pre-game strength whenever the peak exceeds the favorite's initial probability.

Symmetric case ($p_0 = 0.5$). Here $F_{M_\lambda}(x) = 2 - \frac{1}{x}$ on $[0.5, 1)$, and

$$\mathbb{P}(M_\lambda \ge 2/3) = \frac{1}{2}, \qquad \mathbb{P}(M_\lambda \ge 3/4) = \frac{1}{3}, \qquad \mathbb{P}(M_\lambda \ge 0.9) = \frac{1}{9} \approx 0.111.$$

Asymmetric case ($p_0 = 0.75$). The CDF has an intermediate region on $[0.25, 0.75)$:

$$F_{M_\lambda}(x) = \begin{cases} 0, & 0 \le x < 0.25, \\ 1 - \dfrac{0.25}{x}, & 0.25 \le x < 0.75, \\ 2 - \dfrac{1}{x}, & 0.75 \le x < 1, \\ 1, & x = 1. \end{cases}$$

For thresholds $x \ge 0.75$, the same universal tail applies:

$$\mathbb{P}(M_\lambda \ge 3/4) = \frac{1}{3}, \qquad \mathbb{P}(M_\lambda \ge 0.9) = \frac{1}{9} \approx 0.111.$$

(For a sub-favorite threshold such as $x = 0.5 < p_0$, the tail depends on p_0; e.g. $\mathbb{P}(M_\lambda \ge 0.5) = 0.5$ here.)

E.2 Winner's trough (n teams)

Recall the eventual winner's trough M_ω (the minimum win probability attained by the eventual winner). In the symmetric n-team case with $p_0^{(i)} = 1/n$, the CDF is

$$F_{M_\omega}(x) = \begin{cases} 0, & x < 0, \\ \dfrac{(n-1)x}{1-x}, & 0 \le x < 1/n, \\ 1, & x \ge 1/n. \end{cases}$$

Thus M_ω is supported on $[0, 1/n]$ and becomes more concentrated near 0 as n grows.

Symmetric three-team case ($n = 3$). For $x \in [0, 1/3)$, $F_{M_\omega}(x) = \frac{2x}{1-x}$, so

$$\mathbb{P}(M_\omega \le 0.2) = \frac{1}{2}, \qquad \mathbb{P}(M_\omega \le 0.1) = \frac{2}{9} \approx 0.222, \qquad \mathbb{P}(M_\omega \le 0.05) = \frac{2}{19} \approx 0.105.$$

Asymmetric three-team case $(p_0^{(1)}, p_0^{(2)}, p_0^{(3)}) = (1/6, 1/3, 1/2)$**.** Using the general mixture formula for asymmetric n-team games, the CDF is

$$F_{M_\omega}(x) = \begin{cases} 2 \cdot \dfrac{x}{1-x}, & 0 \le x < 1/6, \\ \dfrac{1}{6} + \dfrac{7}{6} \cdot \dfrac{x}{1-x}, & 1/6 \le x < 1/3, \\ \dfrac{1}{2} + \dfrac{1}{2} \cdot \dfrac{x}{1-x}, & 1/3 \le x < 1/2, \\ 1, & x \ge 1/2, \end{cases}$$

so, for example,

$$\mathbb{P}(M_\omega \le 0.1) = \frac{2}{9} \approx 0.222, \qquad \mathbb{P}(M_\omega \le 0.25) = \frac{5}{9} \approx 0.556, \qquad \mathbb{P}(M_\omega \le 0.4) = \frac{5}{6} \approx 0.833.$$

Acknowledgments

The authors would like to thank Professor Jiaoyang Huang, Dr. Ryan Brill, and Dr. Paul Sabin for helpful discussions and feedback on this work.

References

Arnold, D., Henzi, A., and Ziegel, J. F. (2023). Sequentially valid tests for forecast calibration. *The Annals of Applied Statistics*, 17(3):1909–1935.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Dawid, A. P. (1982). The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–613.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–292.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

ESPN (2025). ESPN Play-by-Play and Win Probability Data. Accessed 2025-12-21.

Foster, D. P. and Vohra, R. V. (1998). Asymptotic Calibration. *Biometrika*, 85(2):379–390.

Gilani, S. (2023). *hoopR: Access Men's Basketball Play by Play Data.* R package version 2.1.0.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). Time-Uniform Chernoff Bounds via Nonnegative Supermartingales. *Probability Surveys*, 17:257–317.

Mock, T. (2025). *espnscrapeR: Scrapes Or Collects NFL Data From ESPN.* R package version 0.8.0.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600.

Nikeghbali, A. and Yor, M. (2006). Doob's Maximal Identity, Multiplicative Decompositions and Enlargements of Filtrations. *Illinois Journal of Mathematics*, 50(1–4):791–814.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–597.