Asymptotically Log-Optimal Bayes-Assisted Confidence Sequences for Bounded Means

Valentin Kilian[*]
Department of Statistics,
University of Oxford
kilian@stats.ox.ac.uk

Stefano Cortinovis[*]
Department of Statistics,
University of Oxford
cortinovis@stats.ox.ac.uk

François Caron
Department of Statistics,
University of Oxford
caron@stats.ox.ac.uk

Abstract

Confidence sequences based on test martingales provide time-uniform uncertainty quantification for the mean of bounded i.i.d. observations without parametric distributional assumptions. Their practical efficiency, however, depends strongly on the choice of martingale updates, and many existing constructions do not exploit prior information about plausible data-generating distributions or mean values. We propose a Bayes-assisted framework that uses a Bayesian working predictive model to adaptively construct confidence sequences. For each candidate mean and time point, the predictive distribution selects, among valid one-step martingale factors, the update maximising predictive expected log-growth; validity is therefore preserved even when the prior or working model is misspecified. We prove that if the predictive distribution is Wasserstein-consistent, the resulting procedure is asymptotically log-optimal, matching the per-sample log-growth of an oracle procedure with access to the true distribution. We instantiate the framework using robust predictives based on Dirichlet-process mixtures and Bayesian exponentially tilted empirical likelihood. Experiments on synthetic data, sequential best-arm identification for LLM evaluation, and prediction-powered inference show that informative priors can substantially reduce confidence-sequence width and sampling effort while retaining anytime-valid coverage.

1 Introduction

Modern data analysis is often sequential: online experiments are monitored during data collection, adaptive benchmarks allocate samples to promising systems, and bandit algorithms act before a fixed terminal sample size is known. Repeated inspection and data-dependent stopping can invalidate ordinary fixed-time confidence intervals. The appropriate object is a *confidence sequence*, a sequence $(C_{\alpha,n})_{n\geq 1}$ satisfying

$$\mathbb{P}^*\left(\mu^* \in C_{\alpha,n} \text{ for all } n \geq 1\right) \geq 1-\alpha, \tag{1}$$

where μ^* is the mean of the data-generating distribution $\mathbb{P}^*$. We study the nonparametric bounded-mean problem: $X_1, X_2, \ldots$ are i.i.d. in $[0,1]$, $\mu^* = \mathbb{E}_{\mathbb{P}^*}[X_1]$, and no parametric form is assumed.

Confidence sequences can be constructed by sequentially testing each candidate mean μ. For each μ, one builds a nonnegative evidence process that is controlled if $\mu = \mu^*$ and tends to grow otherwise; values not yet rejected form the confidence sequence. Test martingales and e-processes justify this construction through Ville's inequality [1, 2, 3, 4, 5, 6]. For bounded observations, this can be done with predictable linear factors $1+\lambda_i(X_i-\mu)$, constrained to be nonnegative on $[0,1]$. The constraints give validity; the choice of λ_i determines efficiency.

[*]Equal contribution. Order decided by the outcome of 10 judo fights.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The key practical question is how to choose these predictable coefficients. The Growth Rate Adaptive to the Particular Alternative (GRAPA) strategy [7] uses the empirical distribution of past observations as a forecast for the next one. We follow the same principle, but replace the purely empirical forecast by a Bayesian or pseudo-Bayesian predictive distribution, allowing prior information from previous experiments, expert knowledge, simulations, or structure to guide the evidence process.

This differs from fully parametric Bayesian sequential testing. Under a specified likelihood, mixture likelihood ratios or Bayes factors yield valid sequential tests under that model [8, 9, 10, 11]. Here, the Bayesian model is used only to choose the next coefficient. The one-step factor itself remains in a distribution-free bounded-mean martingale family. Hence validity holds for any predictable working predictive distribution, even under model or prior misspecification. Poor prior information may widen the confidence sequence, but cannot by itself invalidate (1).

We propose *Bayes-assisted confidence sequences* for bounded means. At each time and candidate mean μ, a working predictive distribution is computed from the past and used to choose the next linear coefficient. This separates *validity*, which follows from martingales and Ville's inequality, from *efficiency*, which depends on predictive quality. We show that if the working predictive learns the true distribution in Wasserstein distance, then the method is asymptotically log-optimal: for each false candidate mean, it attains the same limiting per-sample growth rate as an oracle method that knew $\mathbb{P}^*$.

We instantiate the framework with predictives that use prior information early while becoming data-driven over time. A parametric Bayesian predictive is effective when its working model is accurate. For robustness, we also study a mixture Dirichlet-process predictive, interpolating between a parametric Bayesian predictive and the empirical distribution [12, 13, 14]; without the prior component, it recovers the empirical plug-in idea underlying GRAPA. Finally, we develop Bayesian exponentially tilted empirical likelihood and regularized variants, which place prior information on the mean parameter of interest without specifying a full likelihood [15, 16, 17, 18].

Our contributions are as follows.

- We introduce a Bayes-assisted framework in which Bayesian or pseudo-Bayesian predictives guide bounded-mean confidence sequences.
- We prove nonasymptotic time-uniform coverage for any predictable working predictive distribution, including misspecified models and inaccurate priors.
- We show asymptotic log-optimality when the predictive distribution learns the true data-generating distribution.
- We propose asymptotically log-optimal constructions based on mixture Dirichlet-process, BETEL, and RETEL predictives.
- We demonstrate empirically that informative priors can reduce confidence-sequence width and sampling effort, while misspecification affects efficiency rather than validity.

The remainder of the paper is organised as follows. Section 2 discusses related work. Section 3 reviews bounded-mean martingales and confidence sequences by test inversion. Section 4 introduces the generic predictive-assisted construction and efficiency theory. Section 5 develops the parametric, mixture Dirichlet-process, BETEL, and RETEL predictives. Section 6 studies synthetic examples, best-arm identification for LLM evaluation, and prediction-powered inference. Sections 7 and 8 discuss extensions and limitations. The appendices contain proofs, implementation details, prior specifications, and additional experiments.

2 Related work

Prior-assisted fixed-time inference. Fixed-time confidence intervals for bounded means are often obtained by inverting Hoeffding, Bennett–Bernstein, Bentkus, or empirical-Bernstein inequalities [19, 20, 21, 22, 23, 24]. A separate line asks how prior or side information can shorten frequentist intervals while preserving coverage, from Pratt's expected-length-optimal intervals to modern "frequentist, assisted by Bayes" (FAB) procedures [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35]. These methods are primarily parametric and fixed-sample-size; our target is a nonparametric, anytime-valid confidence sequence for bounded means.

Confidence sequences, e-processes, and betting. Confidence sequences originate in Darling and Robbins [2], Robbins [36], Robbins and Siegmund [37], Lai [3], building on martingale and sequential-testing foundations [1, 38]. Modern treatments use nonnegative supermartingales, e-values, and e-processes to obtain validity under optional stopping and continuation [4, 5, 39, 40, 41, 6, 42]. For bounded means, betting martingales are especially relevant: Waudby-Smith and Ramdas [7] introduced betting CIs and CSs, with GRAPA choosing stakes by empirical log-growth optimization, while Orabona and Jun [43] obtained PRECiSE sequences via universal-portfolio regret bounds [44, 45]. Near-optimality and optimality of betting confidence sets have since been studied by Shekhar and Ramdas [46], Clerico [47], and closed-form empirical-Bernstein sequences give strong variance-adaptive baselines [48]. Our method follows the Kelly log-growth principle [49], but replaces GRAPA's empirical predictive with a Bayesian working predictive; validity still follows solely from the bounded-mean martingale and Ville's inequality.

Bayesian and empirical-likelihood predictives. Bayesian sequential tests, mixture likelihood ratios, and Bayes factors yield martingales when a probabilistic model is fully specified, and have been used to encode prior information in parametric or model-based sequential inference [8, 9, 10, 11]. The nonparametric Bayes-assisted confidence sequences of Kilian et al. [50] are related, but their guarantees are asymptotic, whereas ours are nonasymptotic for bounded observations. We instantiate the betting rule using Dirichlet-process posterior predictives [12, 13, 14] and Bayesian empirical or exponentially tilted empirical likelihood [15, 16, 17, 51, 52, 53, 54, 18]. These Bayesian objects are used only to select predictable stakes, not to assert Bayesian posterior coverage.

3 Background on bounded-mean test martingales and confidence sequences

This section gives the general construction used throughout the paper, see e.g. [6]. Let $X_1, X_2, \ldots$ be i.i.d. random variables in $[0,1]$ from $\mathbb{P}^*$ with mean $\mu^* \in (0,1)$. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. For each candidate mean $\mu \in (0,1)$, consider the composite null hypothesis

$$H_\mu : \mu^* = \mu. \tag{2}$$

Bounded-mean test martingales. For $\mu \in (0,1)$ and a parameter $c \in (0,1)$, let

$$I_{\mu,c} := \left[-\frac{c}{1-\mu}, \frac{c}{\mu} \right]. \tag{3}$$

For any predictable (that is, $\mathcal{F}_{i-1}$–measurable) sequence of coefficients $\widehat{\lambda}_i(X_{1:i-1}, \mu) \in I_{\mu,c}$, define

$$W_n(X_{1:n}, \mu) = \prod_{i=1}^{n} \left(1 + \widehat{\lambda}_i(X_{1:i-1}, \mu)(X_i - \mu) \right), \qquad W_0(\mu) = 1. \tag{4}$$

The interval $I_{\mu,c}$ is chosen so that each one-step factor in (4) is nonnegative for all possible observations $x \in [0,1]$. In the betting interpretation, W_n is the bettor's wealth and $\widehat{\lambda}_i(X_{1:i-1}, \mu)$ is the predictable betting coefficient, chosen before observing X_i, on the centered payoff $X_i - \mu$; such bet is fair under H_μ. Such nonnegative multiplicative wealth processes are the standard linear betting martingales used for bounded-mean confidence sequences [7, 47].

Proposition 1 (Test-martingale property)**.** *Fix $\mu \in (0,1)$ and $c \in (0,1)$. If $\widehat{\lambda}_i(X_{1:i-1}, \mu)$ is predictable and belongs to $I_{\mu,c}$ for every $i \geq 1$, then $(W_n(X_{1:n}, \mu))_{n \geq 0}$ is nonnegative. Moreover, under the null hypothesis H_μ in* (2)*, $(W_n(X_{1:n}, \mu))_{n \geq 0}$ is a martingale with initial value one.*

Confidence sequence and anytime-valid coverage. For each $\mu \in (0,1)$ and each $\alpha \in (0,1)$, the process $W_n(X_{1:n}, \mu)$ can be thresholded at level $1/\alpha$ to obtain an anytime-valid sequential test for H_μ with type-I error α. Inversion of this collection of tests yields the confidence sequence

$$C_{\alpha,n}(X_{1:n}) = \left\{ \mu \in (0,1) : \; W_n(X_{1:n}, \mu) \leq \frac{1}{\alpha} \right\} \subseteq [0,1]. \tag{5}$$

Proposition 2 (Anytime-valid coverage)**.** *For any $\alpha \in (0,1)$, $(C_{\alpha,n}(X_{1:n}))_{n \geq 1}$ in Equation* (5) *is a $1-\alpha$ confidence sequence for the mean parameter μ^*. That is, it satisfies Equation* (1)*.*

Oracle log-optimality. The anytime-coverage result above holds for every predictable choice of the coefficients. The remaining question is how to choose them so that the sequential tests have high power against false candidate means. For a fixed candidate $\mu \in (0,1)$, define the expected log-evidence of a constant coefficient $\lambda \in I_{\mu,c}$ under the true distribution $\mathbb{P}^*$ by

$$M(\lambda, \mu) := \mathbb{E}_{X_1 \sim \mathbb{P}^*} \left[\log \left(1 + \lambda (X_1 - \mu) \right) \right]. \tag{6}$$

If $\mathbb{P}^*$ were known, the oracle coefficient would be

$$\lambda^*(\mu) = \arg\max_{\lambda \in I_{\mu,c}} M(\lambda, \mu), \tag{7}$$

and the corresponding oracle test martingale would be

$$W_n^*(X_{1:n}, \mu) = \prod_{i=1}^{n} \left(1 + \lambda^*(\mu)(X_i - \mu) \right). \tag{8}$$

If $\mu = \mu^*$, then $\lambda^*(\mu) = 0$ and the oracle log-growth is zero. For a candidate mean $\mu \neq \mu^*$, $M(\lambda^*(\mu), \mu)$ is the oracle per-observation log-evidence against H_μ within the class (4). Since $\mathbb{P}^*$ is unknown, $\lambda^*(\mu)$ is unavailable. We call a data-dependent construction asymptotically log-optimal if it matches this oracle log-evidence asymptotically, in the sense that

$$\frac{1}{n} \mathbb{E}_{\mathbb{P}^*} \left[\log W_n(X_{1:n}, \mu) \right] \to M(\lambda^*(\mu), \mu). \tag{9}$$

4 Predictive-Assisted Test Martingales for Bounded Means

We now describe our generic predictive-based approach to choose the predictable betting coefficient $(\widehat{\lambda}_i)$ in Equation (4) and state a simple sufficient condition for the associated test martingales to be asymptotically log-optimal.

Consider a sequence of $\mathcal{F}_{n-1}$–measurable predictive distributions $\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1})$, $n \geq 1$, on $[0,1]$. For a fixed candidate mean μ, the nth predictive distribution is used to choose the next betting coefficient by maximising the conditional expected logarithm of the nth factor of W_n in Equation (4):

$$\widehat{\lambda}_n(X_{1:n-1}, \mu) \in \arg\max_{\lambda \in I_{\mu,c}} \left\{ \mathbb{E}_{\widetilde{X}_n \sim \mathbb{Q}_{n|n-1}(\cdot | X_{1:n-1})} \left[\log \left(1 + \lambda (\widetilde{X}_n - \mu) \right) \right] \right\}. \tag{10}$$

When the maximiser is not unique, we fix an arbitrary measurable selection. A process of the form (4), with coefficients chosen by (10), is called a *predictive-assisted test martingale*. The role of $\mathbb{Q}_{n|n-1}$ in (10) is purely algorithmic: it guides the choice of a one-step test factor. The predictive model is not assumed to be correctly specified for the purpose of coverage: anytime coverage holds for any choice of predictive sequence.

The next theorem shows that the predictive-assisted construction is asymptotically log-optimal whenever the sequence of predictive distributions consistently estimates $\mathbb{P}^*$ in Wasserstein distance. Let $\mathcal{W}_1$ denote the Wasserstein-1 distance on probability distributions on $[0,1]$.

Theorem 1 (Asymptotic log-optimality)**.** *Let $(X_n)_{n \geq 1}$ be i.i.d. in $[0,1]$ from a non-degenerate distribution $\mathbb{P}^*$, and let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Let $(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}))_{n \geq 1}$ be an $\mathcal{F}_{n-1}$-measurable sequence of predictive distributions on $[0,1]$ such that*

$$\mathcal{W}_1(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}), \mathbb{P}^*) \to 0 \qquad \textit{in probability}. \tag{11}$$

Fix $\mu \in (0,1)$ and $c \in (0,1)$. Define

$$M(\lambda, \mu) := \mathbb{E}_{X_1 \sim \mathbb{P}^*} \left[\log \left(1 + \lambda (X_1 - \mu) \right) \right], \tag{12}$$

$$M_n(\lambda, X_{1:n-1}, \mu) := \mathbb{E}_{X'_n \sim \mathbb{Q}_{n|n-1}} \left[\log \left(1 + \lambda (X'_n - \mu) \right) \mid X_{1:n-1} \right]. \tag{13}$$

Let $\widehat{\lambda}_n(X_{1:n-1}, \mu)$ be an $\mathcal{F}_{n-1}$-measurable maximiser of $\lambda \mapsto M_n(\lambda, X_{1:n-1}, \mu)$ on $I_{\mu,c}$. Then $\lambda \mapsto M(\lambda, \mu)$ admits a unique maximiser $\lambda^*(\mu)$ over $I_{\mu,c}$, and

$$\widehat{\lambda}_n(X_{1:n-1}, \mu) \to \lambda^*(\mu) \qquad \textit{in probability}. \tag{14}$$

Moreover, for $W_n(X_{1:n}, \mu) = \prod_{i=1}^{n} (1 + \widehat{\lambda}_i(X_{1:i-1}, \mu)(X_i - \mu))$,

$$\frac{1}{n} \log W_n(X_{1:n}, \mu) \to M(\lambda^*(\mu), \mu) \qquad \textit{in probability}, \tag{15}$$

$$\frac{1}{n} \mathbb{E}_{\mathbb{P}^*} \left[\log W_n(X_{1:n}, \mu) \right] \to M(\lambda^*(\mu), \mu). \tag{16}$$

Proposition 2 gives anytime-valid coverage for any sequence of predictive distributions. In contrast, Theorem 1 states that, when the predictive distributions converge to $\mathbb{P}^*$, the induced sequential tests attain the same first-order log-evidence as the oracle construction that knows $\mathbb{P}^*$.

The following finite-sample comparison makes explicit how the loss relative to the oracle is controlled by the predictive error.

Theorem 2 (Predictive error and expected log-evidence)**.** *With the notation of Theorem 1,*

$$M(\lambda^*(\mu),\mu) - \frac{1}{n}\mathbb{E}_{\mathbb{P}^*}\left[\log W_n(X_{1:n},\mu)\right] \le 2L_{\mu,c} \times \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbb{P}^*}\left[\mathcal{W}_1\left(\mathbb{Q}_{i|i-1}(\cdot \mid X_{1:i-1}),\mathbb{P}^*\right)\right], \tag{17}$$

where $L_{\mu,c} = \frac{c}{(1-c)\min(\mu,1-\mu)} < \infty$.

The proposed method can therefore use any sequence of predictive distributions satisfying (11). In the next section, we describe constructions of $\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1})$ based on Bayesian nonparametric and empirical-likelihood ideas.

5 Bayesian and pseudo-Bayesian predictive constructions

Proposition 2 implies that anytime coverage is obtained for any predictable sequence of predictive distributions $\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1})$ on $[0,1]$. The predictive model is used only to choose the one-step betting coefficient in Equation (10). Theorem 1 then shows that the same construction is asymptotically log-optimal whenever Equation (11) holds. This section gives several choices of $\mathbb{Q}_{n|n-1}$. A first natural approach is to take a fully Bayesian approach to this problem, by specifying a parametric or nonparametric prior distribution on the unknown $\mathbb{P}^*$. $\mathbb{Q}_{n|n-1}$ is then the posterior predictive distribution under this Bayesian model. A parametric version is presented in Section 5.1, and a log-optimal nonparametric approach in Section 5.2.

In many applications, prior information may be available for the mean parameter of interest μ^*, but the elicitation of a prior for the unknown distribution $\mathbb{P}^*$ is challenging. To this end, we consider a Bayesian empirical likelihood in Section 5.3, which bypasses the need to fully specify a Bayesian model. This approach is also shown to be asymptotically log-optimal.

5.1 Parametric Bayesian predictive

The most direct construction starts from a parametric working model. Let $G_0(\cdot;\phi)$ be a probability distribution on $[0,1]$ indexed by $\phi \in \Phi$, and let $\pi(d\phi)$ be a prior on Φ. The posterior predictive distribution is

$$\mathbb{Q}_{n+1|n}(dx \mid X_{1:n}) = \int_\Phi G_0(dx;\phi)\,\pi(d\phi \mid X_{1:n}), \tag{18}$$

with $\mathbb{Q}_{1|0}(dx) = \int_\Phi G_0(dx;\phi)\,\pi(d\phi)$. While such approach has anytime coverage, and can provide shorter confidence sequences for small sample sizes, it need not be asymptotically log-optimal if misspecified. This motivates the use of a Bayesian nonparametric approach, as introduced in the next section.

5.2 Mixture Dirichlet process-style predictive

Let $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ denote the empirical distribution, and

$$\mathbb{H}_{n+1|n}(dx \mid X_{1:n}) = \int_\Phi G_0(dx;\phi)\,\pi(d\phi \mid X_{1:n})$$

be a parametric posterior predictive distribution of the form (18). For $\kappa > 0$, define

$$\mathbb{Q}_{n+1|n}(dx \mid x_{1:n}) = \frac{\kappa}{\kappa+n}\,\mathbb{H}_{n+1|n}(dx \mid X_{1:n}) + \frac{n}{\kappa+n}\,\mathbb{P}_n(dx), \qquad n \ge 1. \tag{19}$$

The form in (19) is inspired by the Pólya-urn posterior predictive of a Dirichlet process [12, 13]. When G_0 is non-atomic and all observations are distinct, it corresponds to the posterior predictive in Antoniak's mixture of Dirichlet processes [14]. We call (19) a mixture-DP-style (MDP) predictive.

It can also be viewed as a robust interpolation between the parametric predictive and the empirical distribution. For $\kappa = 0$, one recovers GRAPA Waudby-Smith and Ramdas [7].

The parameter κ controls how quickly the construction moves from the prior-driven component to the data-driven component. Large κ gives more weight to the parametric predictive at early sample sizes, while small κ moves quickly toward the empirical distribution. For every fixed κ, the empirical component has weight tending to one. Thus the construction can exploit informative prior information early, while recovering predictive consistency even if the parametric likelihood is misspecified. This type of prior-data interpolation is closely related to robust Bayesian nonparametric and general Bayesian ideas for mitigating model misspecification [55, 56, 57].

Proposition 3 (Predictive consistency of the MDP predictive)**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. from a distribution $\mathbb{P}^*$ supported on $[0,1]$. Let $\mathbb{Q}_{n+1|n}$ be defined by* (19)*, with the parametric component $\mathbb{H}_{n+1|n}$ also supported on $[0,1]$. Then*

$$\mathbb{E}_{\mathbb{P}^*}\left[\mathcal{W}_1\big(\mathbb{Q}_{n+1|n}(\cdot \mid X_{1:n}), \mathbb{P}^*\big)\right] = O(n^{-1/2}). \tag{20}$$

Consequently, $\mathcal{W}_1\big(\mathbb{Q}_{n+1|n}(\cdot \mid X_{1:n}), \mathbb{P}^*\big) \to 0$ in $\mathbb{P}^*$-probability.

By Theorem 1, the Bayes-assisted confidence sequence based on the MDP predictive is therefore asymptotically log-optimal.

5.3 Bayesian empirical likelihood predictive

The MDP predictive in Equation (19) is robust to likelihood misspecification, but still requires a prior model for the full data-generating distribution. When prior information is available only for the mean, we instead use a Bayesian empirical likelihood: the prior is placed on μ, while the likelihood is estimated from the moment restriction $\mathbb{E}_{\mathbb{P}^*}[X - \mu] = 0$. We consider Bayesian exponentially tilted empirical likelihood (BETEL) [16, 17] and a regularized version (RETEL), related to adjusted and regularized empirical likelihood methods [51, 52, 54, 53, 18].

Let π be a prior on $[0,1]$. For each $\mu \in (0,1)$, let $\mathbb{H}_\mu$ be a regularizing distribution on $[0,1]$, and let $\tau_n \geq 0$ with $\tau_n = o(n)$. Recall that $\mathbb{P}_n$ is the empirical distribution, and define $\mathbb{P}_{n,\mu} = \frac{n}{n+\tau_n}\mathbb{P}_n + \frac{\tau_n}{n+\tau_n}\mathbb{H}_\mu$. For a candidate mean μ, consider as an empirical likelihood distribution

$$\widetilde{\mathbb{P}}_{n,\mu} = \arg\min_{P \ll \mathbb{P}_{n,\mu}} \left\{ \mathrm{KL}(P \,\|\, \mathbb{P}_{n,\mu}) : \int (z-\mu)\, dP(z) = 0 \right\}. \tag{21}$$

Thus $\widetilde{\mathbb{P}}_{n,\mu}$ is the closest tilt of the regularized empirical measure whose mean is exactly μ. Writing $\psi_{n,\mu}(\gamma) = \log \int \exp\{\gamma(z-\mu)\}\, d\mathbb{P}_{n,\mu}(z)$, the solution is

$$\frac{d\widetilde{\mathbb{P}}_{n,\mu}}{d\mathbb{P}_{n,\mu}}(z) = \exp\{\gamma_{n,\mu}(z-\mu) - \psi_{n,\mu}(\gamma_{n,\mu})\}, \qquad \psi'_{n,\mu}(\gamma_{n,\mu}) = 0. \tag{22}$$

The case $\tau_n = 0$ gives BETEL, with likelihood set to zero outside the interior of the empirical convex hull. If $\tau_n > 0$ and $\mathrm{conv}(\operatorname{supp} \mathbb{H}_\mu) = [0,1]$, the construction is defined for every $\mu \in (0,1)$. Let

$$M_{\mathbb{H}_\mu}(\gamma) = \int e^{\gamma(z-\mu)}\, d\mathbb{H}_\mu(z), \qquad Z_{n,\mu} = \sum_{j=1}^{n} e^{\gamma_{n,\mu}(X_j - \mu)} + \tau_n M_{\mathbb{H}_\mu}(\gamma_{n,\mu}),$$

and define the empirical-likelihood weights $w_{n,i}(\mu) = e^{\gamma_{n,\mu}(X_i-\mu)}/Z_{n,\mu}$, $i = 1, \ldots, n$. For the two-point regularizer used below, the corresponding empirical-likelihood contribution is, up to a μ-free factor, $L_n(\mu) = \prod_{i=1}^n w_{n,i}(\mu)$. The pseudo-posterior and pseudo-predictive are

$$\Pi_n(d\mu) = \frac{L_n(\mu)\pi(d\mu)}{\int_0^1 L_n(u)\pi(du)}, \qquad \mathbb{Q}_{n+1|n}(A \mid X_{1:n}) = \int_0^1 \widetilde{\mathbb{P}}_{n,\mu}(A)\, \Pi_n(d\mu). \tag{23}$$

The next proposition verifies the Wasserstein consistency condition needed for Theorem 1; its proof is given in Appendix A.6.

Proposition 4 (Consistency of the RETEL predictive)**.** *Assume that $X_1, X_2, \ldots$ are i.i.d. from a non-degenerate distribution $\mathbb{P}^*$ supported on $[0,1]$, with mean $\mu^* \in (0,1)$ and variance $\sigma^2 > 0$. Assume*

that π is absolutely continuous on $[0,1]$, with density bounded away from zero in a neighbourhood of μ^*, that $\tau_n = o(n)$, and that each $\mathbb{H}_\mu$ is supported on $[0,1]$ with $\mathrm{conv}(\mathrm{supp}\,\mathbb{H}_\mu) = [0,1]$. Then, for every $\epsilon > 0$,

$$\Pi_n\big(|\mu - \mu^*| > \epsilon\big) \to 0, \quad \textit{and} \quad \Pi_n\big(|\gamma_{n,\mu}| > \epsilon\big) \to 0, \tag{24}$$

in probability. Moreover, in probability, $\mathcal{W}_1\big(\mathbb{Q}_{n+1|n}(\cdot \mid X_{1:n}), \mathbb{P}^*\big) \to 0$.

By Theorem 1, the Bayes-assisted confidence sequence based on the RETEL predictive is asymptotically log-optimal. In the experiments we use two cases: unregularized BETEL, $\tau_n = 0$, and the two-point RETEL regularizer $\tau_n = 1$, $\mathbb{H}_\mu = \mathbb{H} = \frac{1}{2}(\delta_0 + \delta_1)$. The latter acts as two endpoint half-pseudo-observations, stabilizing the empirical likelihood while remaining asymptotically data-dominated. Computational details are given in Appendix C.

6 Experiments

We evaluate our Bayes-assisted CS methods on synthetic data and two real-world applications, showing that informative working predictives can shorten confidence sequences and reduce stopping times while preserving anytime validity under misspecification. The Bayes-assisted methods use the robust predictives from Section 5: the mixture Dirichlet process (MDP) predictive (19) with a beta base distribution, BETEL (23) with $\tau_n = 0$, and RETEL (23) with $\tau_n = 1$ and the two-point regulariser. We compare against the empirical predictive and a parametric beta predictive within the same framework, and against prior-free baselines outside it: Hedged-Capital (HC) [7], PRECiSE-CO96 [43], and the full Cover–Ordentlich universal portfolio (UP) method [45]. The main text reports the proposed methods together with the empirical predictive and UP, which we find to be the strongest prior-free baselines in our experiments; full comparisons with HC, PRECiSE-CO96, and parametric Bayes are deferred to Appendix G. We report CS width and, in decision-making experiments, stopping time; in synthetic experiments we also report empirical cumulative miscoverage and widths normalised by the oracle CS whose coefficient maximises expected log-growth under the true distribution. Unless otherwise specified, we set $\alpha = 0.1$, invert tests on the 499 interior points of the $1/500$ grid on $[0,1]$, and use $c = 0.95$ for the empirical and Bayes-assisted methods; method- and baseline-specific details are provided in Appendix D.

Synthetic experiments. We first consider synthetic data, where $\mathbb{P}^*$ is known and the oracle coefficient $\lambda^*(\mu)$ in Equation (7), and hence the oracle confidence sequence $C^\star_{\alpha,n}$, can be computed. This lets us assess how close each method comes to the log-optimal benchmark suggested by (8). We generate i.i.d. observations on $[0,1]$ from six distributions: $\mathrm{Bernoulli}(p)$ with $p \in \{0.1, 0.5\}$, $\mathrm{Beta}(a,b)$ with $(a,b) \in \{(0.5,0.5), (1,1), (10,30)\}$, and the asymmetric mixture $0.25\,\mathrm{Beta}(5,15) + 0.75\,\mathrm{Beta}(15,5)$. The Bernoulli and beta-mixture examples test departures from the beta working model used by the parametric and MDP predictives. In the main text, we report the informative-prior comparison for the mixture example. The remaining synthetic distributions under informative priors are shown in Appendix G.1, while the non-informative and misspecified-prior regimes are considered in Appendix G.1.2. Detailed prior specifications are provided in Appendix F.1.1.

Figure 1 compares methods on the beta-mixture experiment in terms of oracle-normalised CS width and empirical cumulative miscoverage. With informative priors, Bayes-assisted methods are closer to the oracle than the empirical predictive and UP baselines over much of the sample-size range. The same overall qualitative pattern is observed for the remaining synthetic distributions in Figures 4 and 5: when the prior is informative, Bayes-assisted predictives usually reduce oracle-normalised width relative to prior-free baselines. Among the Bayes-assisted methods in Figure 1, MDP benefits strongly from prior information at small n, while BETEL and RETEL gradually catch up as the sample size grows. The prior-sensitivity experiments in Appendix G.1.2 show the complementary effect: diffuse or misspecified priors can increase oracle-normalised width, but empirical miscoverage remains controlled. Thus prior quality affects efficiency, whereas coverage is guaranteed by the underlying test-martingale construction. Sensitivity to the truncation level c and the MDP mixing parameter κ is studied in Appendices G.1.4 and G.1.5.

Best-arm identification for LLM evaluation. As a first real-data application, we consider adaptive LLM evaluation for identifying the best of four LLMs on a sorting task. We frame the problem as

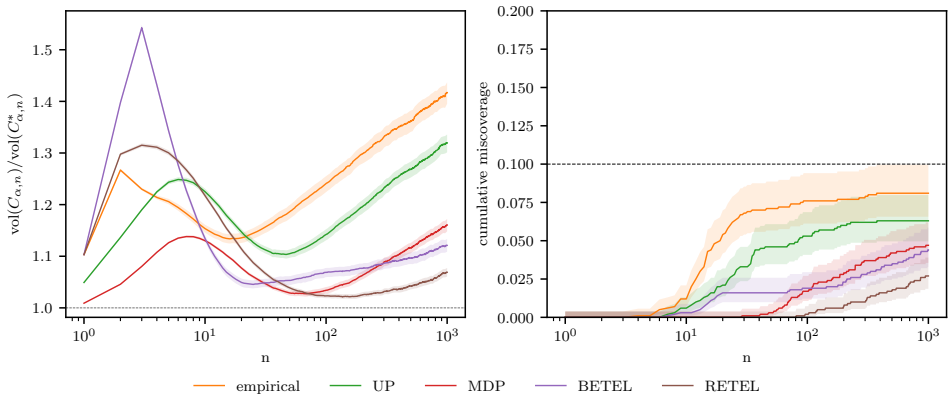

Figure 1: Synthetic beta-mixture experiment. The left and right panels show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 replicates; the dashed line marks $\alpha = 0.1$. Width ratios use replicates where both the method and oracle CSs are non-empty. At $n = 1000$, at most 2% are empty (0.8% on average). Shading shows pointwise 95% paired-bootstrap intervals for width ratios and Wilson intervals for cumulative miscoverage.

Table 1: Sequential stopping-time comparisons in the two real-data applications. Rank is computed within each replicate; W/T/L is against the empirical predictive baseline.

(a) LLM best-arm identification. Avg calls is the LUCB stopping time. Across all methods, no CS is empty at the stopping time.

Method	Avg calls (s.e.)	Avg rank (s.e.)	W/T/L
MDP	91.68 (3.90)	1.95 (0.13)	38/5/7
BETEL	99.36 (3.62)	3.28 (0.17)	23/7/20
Emp.	101.64 (4.03)	3.58 (0.21)	–
RETEL	109.76 (3.63)	4.95 (0.23)	18/2/30
UP	110.68 (3.53)	5.18 (0.16)	4/3/43

(b) ATTRIBUTEDQA PPI test. Avg labels is time to rejection. Across all methods, no CS is empty at the stopping time.

Method	Avg labels (s.e.)	Avg rank (s.e.)	W/T/L
BETEL	50.25 (3.37)	3.33 (0.18)	71/15/14
MDP	52.18 (3.50)	3.49 (0.18)	82/16/2
RETEL	55.03 (2.49)	4.24 (0.22)	62/8/30
Emp.	65.24 (4.12)	5.88 (0.25)	–
UP	69.73 (3.24)	6.48 (0.16)	38/4/58

best-arm identification: each LLM is a bandit arm, and pulling an arm queries the model on a random array of length 20 and returns a rescaled Spearman rank-correlation score in $[0, 1]$ measuring the quality of the sorted output. We use LUCB [58, 59] with arm-wise confidence sequences, stopping once the leading model is separated from its competitors. This avoids fixing the same benchmark size for all models and requires anytime-valid bounds because LUCB repeatedly inspects confidence sequences, adaptively samples arms, and stops at a data-dependent time, thereby potentially reducing the total number of LLM calls. For this example, we elicit priors from cheaper auxiliary evaluations: for each LLM, we evaluate the same task on smaller arrays, fit an exponential decay trend in array length, and centre the arm prior at the extrapolated score for length 20. Details on LUCB, auxiliary array lengths, and prior parameters are given in Appendix F.2.2.

Table 1a reports the average LUCB stopping time over 50 replicates, paired ranks, and win/tie/loss counts against the empirical predictive baseline, the strongest prior-free method in this experiment. MDP performs best in terms of both average calls and paired rank, stopping earlier than the empirical predictive in 38 of 50 paired replicates, while BETEL is close to empirical. Together, these observations suggest that the Beta prior is a good choice for this experiment. UP requires more calls on average than the empirical predictive. Full-baseline results and paired-comparison analyses are reported in Appendix G.2, while results under deliberately misspecified priors are reported in Appendix G.2.3.

Prediction-powered inference. As a second real-data application, we consider prediction-powered inference (PPI) [60] for bounded mean estimation. Given a fixed prediction rule f, PPI decomposes the target mean as $\theta^\star = \mathbb{E}[Y] = \mathbb{E}[f(X)] + \mathbb{E}[Y - f(X)] = m^\star + \Delta^\star$, where $m^\star$ is estimated from unlabelled data and the rectifier $\Delta^\star$ from labelled residuals. When labelled and unlabelled data both arrive sequentially, one can build confidence sequences for $m^\star$ and $\Delta^\star$ separately and

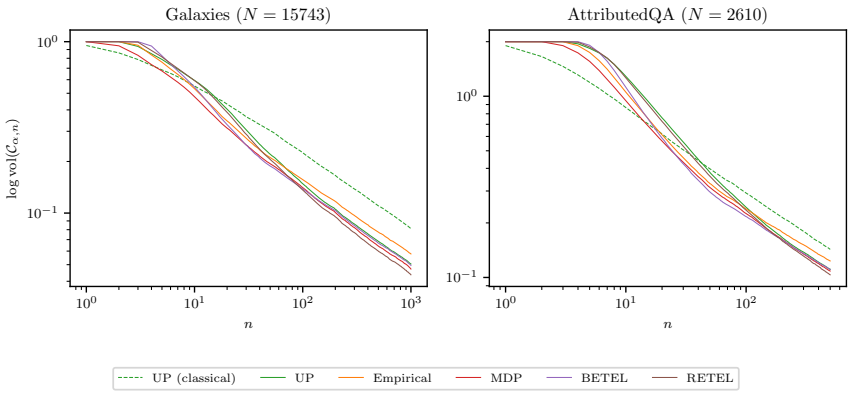

Figure 2: Prediction-powered mean estimation. The left and right panels show average CS length over 100 replicates for the GALAXIES and ATTRIBUTEDQA datasets. Curves labelled "classical" use only labelled observations, while the remaining curves use the PPI construction. Empty CSs are counted as width zero (at the final sample size, at most 1% are empty, 0.2% on average).

combine them by a union bound. As noted by Kilian et al. [50], however, this merging step can be loose and can obscure the contribution of the confidence-sequence method used for the rectifier. We therefore follow their simplified setting: a large pool of N unlabelled examples is available at the start, $m^\star$ is treated as fixed through the plug-in estimate $\frac{1}{N}\sum_{j=1}^{N} f(X_j)$, and sequential uncertainty is concentrated on the labelled residuals $Y_i - f(X_i)$. Since these residuals are bounded, we apply our bounded-mean CSs after rescaling to $[0,1]$; see Appendix F.3.1 for details. For this application, we use priors centred at zero for $\Delta^\star$, encoding the belief that the prediction rule is approximately unbiased. After rescaling the residual range to $[0,1]$, this corresponds to symmetric beta priors centred at $1/2$, with concentration controlling the strength of the prior belief; full specifications are given in Appendix F.3.3.

Figure 2 compares methods on two PPI tasks: estimating the proportion of galaxies with spiral arms in GALAXIES, and estimating the difference in attribution reliability between two retrieve-then-read QA systems, RTR-10 and RTR-4, on ATTRIBUTEDQA. For ATTRIBUTEDQA, both labels and predictions are differences between the two systems' binary attribution outcomes/probabilities; dataset details are in Appendix F.3.2. PPI-based CSs are substantially narrower than classical labelled-only CSs on both datasets. Among PPI methods, the Bayes-assisted procedures are competitive with the empirical and UP baselines: MDP performs best at small sample sizes, while RETEL is strongest at larger sample sizes in these experiments.

These width reductions translate into earlier anytime-valid decisions. On ATTRIBUTEDQA, testing $H_0 : \theta^\star \le 0$ asks whether RTR-10 has higher attribution reliability than RTR-4; rejection occurs once 0 is excluded from the CS. Table 1 shows that MDP, BETEL, and RETEL all reject using fewer labelled examples on average than the empirical predictive and UP baselines. The paired comparisons against the empirical predictive are strongest for MDP and BETEL, while RETEL also wins more often than it loses. Thus an informative prior on the rectifier can shorten CSs and reduce the labelled sample size needed for sequential conclusions. Full comparisons with additional baselines are reported in Appendix G.3, while the corresponding misspecified-prior experiment is reported in Appendix G.3.2.

7 Extensions beyond univariate i.i.d. bounded means

While our results focus on i.i.d. observations in $[0,1]$, the principle of using a predictive distribution to select a valid predictable bet also suggests extensions beyond this setting.

Quantiles. Let $(X_n)_{n\ge1}$ be i.i.d. from a distribution with cdf F, without requiring bounded observations. For a candidate quantile q, define $Z_n(q) = \mathbf{1}\{X_n \le q\}$. The variables $(Z_n(q))_{n\ge1}$ are i.i.d. and bounded with mean $F(q)$, so our construction can be applied to $F(q)$ and inverted over q. This connects directly with existing sequential inference for quantiles [61].

Relaxing the i.i.d. assumption. If $\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = \mu^*$ for all $n \geq 1$, independence is unnecessary for validity: every predictable admissible stake still yields a nonnegative martingale. Consequently, the Bayes-assisted procedures presented in this work remain anytime-valid under this conditional-mean assumption, irrespective of predictive-model misspecification. However, the log-optimality results in Theorems 1 and 2 rely on the i.i.d. assumption and do not directly extend to this larger class. For example, suppose $(X_n)_{n\geq 1}$ is Markov with transition kernel K and constant conditional mean. The relevant oracle stake is then state-dependent, maximising $\int \log\{1 + \lambda(y - \mu)\}\, K(x, dy)$ over $\lambda \in I_{\mu,c}$, whereas our methods learn a single stake from the marginal distribution. A natural Bayes-assisted extension would approximate the conditional law, e.g. the Markov transition kernel, rather than the marginal law, and select λ_i using the resulting conditional posterior predictive. Establishing log-optimality would then require a new analysis relative to this richer, state-dependent oracle.

Relaxing boundedness. Without additional moment or tail constraints, there is no nontrivial universal e-variable for the mean of arbitrary unbounded observations. Suppose instead that, under a candidate null, $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = \mu$ and $\mathrm{Var}(X_i \mid \mathcal{F}_{i-1}) \leq v$. Then $E_i(\mu) = (X_i - \mu)^2/v$ is a conditional e-variable. Hence, for any predictable $\eta_i \in [0, 1]$, $M_n(\mu) = \prod_{i=1}^n \{1 - \eta_i + \eta_i E_i(\mu)\}$ is a test supermartingale. This bounded-variance construction is developed by [62]. Our approach could be extended to select η_i by maximising posterior-predictive expected log-growth. The MDP predictive could combine the empirical distribution with a Gaussian or Student-t parametric component, while a possible empirical-likelihood analogue is the Gaussian-regularized RETEL construction of [18].

Multivariate extensions. Suppose $X_i \in K \subset \mathbb{R}^d$, where K is known, compact, and convex. For each candidate mean $\mu \in K$, replace the scalar stake by a predictable vector $\lambda_i(\mu)$ and define $W_n(\mu) = \prod_{i=1}^n \{1 + \lambda_i(\mu)^\top (X_i - \mu)\}$. A suitable admissible set is $\Lambda_{\mu,c} = \{\lambda \in \mathbb{R}^d : \inf_{x\in K}[1 + \lambda^\top (x - \mu)] \geq 1 - c\}$. This constraint keeps every wealth factor nonnegative and bounded away from zero. If $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = \mu^\star$, then $W_n(\mu^\star)$ is a nonnegative martingale for every predictable choice $\lambda_i(\mu^\star) \in \Lambda_{\mu^\star,c}$. Inverting these tests therefore gives an anytime-valid confidence sequence for the vector mean. Given a predictive distribution Q_i on K, the Bayes-assisted vector stake maximises $\int_K \log\{1 + \lambda^\top (x - \mu)\}\, Q_i(dx)$ over $\lambda \in \Lambda_{\mu,c}$. For fixed μ, this is a concave optimization problem over a convex admissible set. A parametric predictive can represent dependence between coordinates, while an MDP can mix that predictive with the empirical joint distribution. BETEL and RETEL also accommodate vector-valued moment restrictions. Theorem 3 extends Theorem 1 to vector means.

8 Discussion and limitations

This paper introduced a Bayes-assisted framework for constructing anytime-valid confidence sequences for bounded means: Bayesian or pseudo-Bayesian predictives are used to choose powerful predictable betting updates, while validity remains guaranteed by test-martingale inversion even under prior or working-model misspecification. The theory separates validity from efficiency, showing oracle-level asymptotic log-growth when the predictive distribution is consistent, and the MDP, BETEL, and RETEL constructions illustrate how informative prior information can yield shorter confidence sequences in practice. Several limitations remain: computing the sets can be substantially more expensive than closed-form baselines because it requires evaluating martingales over candidate means and repeatedly updating predictives (see Appendix D.3); while poor priors do not invalidate coverage, the efficiency gains depend on the prior or auxiliary predictive being sufficiently informative (see Appendix G.1.2); the inverted sets are not guaranteed to be single intervals, which can complicate reporting or downstream use; the pointwise numerical inversion used in the main experiments guarantees coverage only at the tested grid values, although we provide an alternative continuum-valid construction in Appendix E. These limitations are natural targets for future work on faster algorithms, interval-valued approximations, and adaptive or data-driven prior construction.

Acknowledgments and Disclosure of Funding

The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility [63] in carrying out this work. Valentin Kilian is supported by the Clarendon Funds Scholarship. Stefano Cortinovis is supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1).

References

[1] Jean Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, 1939.

[2] D. A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, July 1967. doi: 10.1073/pnas.58.1.66.

[3] Tze Leung Lai. On Confidence Sequences. *The Annals of Statistics*, 4(2):265–280, March 1976. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176343406.

[4] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), April 2021. ISSN 0090-5364. doi: 10.1214/20-AOS1991.

[5] Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H. Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences, March 2024.

[6] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2):1–390, 2025. doi: 10.1561/3600000002.

[7] I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

[8] M. Kumon, A. Takemura, and K. Takeuchi. Capital process and optimality properties of a Bayesian skeptic in coin-tossing games. *Stochastic Analysis and Applications*, 26(6):1161–1180, 2008.

[9] Masayuki Kumon, Akimichi Takemura, and Kei Takeuchi. Sequential optimizing strategy in multi-dimensional bounded forecasting games. *Stochastic Processes and their Applications*, 121(1):155–183, 2011. doi: 10.1016/j.spa.2010.09.004.

[10] Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL `https://jmlr.org/papers/v22/18-798.html`.

[11] Stefano Cortinovis, Valentin Kilian, and François Caron. Confidence sequences with informative, bounded-influence priors, July 2025.

[12] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[13] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.

[14] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[15] A. B. Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

[16] N. A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90(2):319–326, 2003.

[17] S. M. Schennach. Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92(1):31–46, 2005.

[18] Eunseop Kim, Steven N. MacEachern, and Mario Peruggia. Regularized exponentially tilted empirical likelihood for bayesian inference. *arXiv preprint arXiv:2312.17015*, 2026. doi: 10.1214/26-BA1620. URL `https://arxiv.org/abs/2312.17015`.

[19] Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 0162-1459. doi: 10.2307/2282952.

[20] George Bennett. Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. ISSN 0162-1459. doi: 10.2307/2282438.

[21] Sergei N. Bernshtein. *Theory of Probability*. Gostekhizdat, Moscow-Leningrad, 1927. In Russian, VIII+363 pages.

[22] Vidmantas Bentkus. On Hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, April 2004. ISSN 0091-1798, 2168-894X. doi: 10.1214/009117904000000360.

[23] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning Bandit Algorithms in Stochastic Environments. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory*, volume 4754, pages 150–165, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75224-0 978-3-540-75225-7. doi: 10.1007/978-3-540-75225-7_15.

[24] Andreas Maurer and M. Pontil. Empirical Bernstein Bounds and Sample-Variance Penalization. In *Annual Conference Computational Learning Theory*, July 2009.

[25] J. W. Pratt. Length of confidence intervals. *Journal of the American Statistical Association*, 56 (295):549–567, 1961.

[26] J. W. Pratt. Shorter confidence intervals for the mean of a normal distribution with known variance. *The Annals of Mathematical Statistics*, pages 574–586, 1963.

[27] L. D. Brown, G. Casella, and J.T. G. Hwang. Optimal confidence sets, bioequivalence, and the limacon of Pascal. *Journal of the American Statistical Association*, 90(431):880–889, 1995.

[28] C. Yu and P. D. Hoff. Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105(2):319–335, 2018.

[29] D. Farchione and P. Kabaila. Confidence intervals for the normal mean utilizing prior information. *Statistics & Probability Letters*, 78(9):1094–1100, 2008.

[30] P. Kabaila and K. Giri. Further properties of frequentist confidence intervals in regression that utilize uncertain prior information. *Australian & New Zealand Journal of Statistics*, 55(3): 259–270, 2013.

[31] P. Kabaila and D. Farchione. Confidence intervals that utilize sparsity. *Stat*, 11(1):e434, 2022.

[32] P. Hoff and C. Yu. Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, 13:94–119, 2019.

[33] P. Hoff. Bayes-optimal prediction with frequentist coverage control. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 29(2):901–928, 2023.

[34] Stefano Cortinovis and François Caron. Bayes-assisted Confidence Regions: Focal Point Estimator and Bounded-influence Priors. *arXiv preprint arXiv:2410.20169*, 2024.

[35] Stefano Cortinovis and François Caron. FAB-PPI: Frequentist, Assisted by Bayes, Prediction-Powered Inference, August 2025.

[36] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. doi: 10.1214/aoms/1177696786.

[37] Herbert Robbins and David Siegmund. Boundary crossing probabilities for the Wiener process and sample sums. *The Annals of Mathematical Statistics*, pages 1410–1429, 1970.

[38] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

[39] G. Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* John Wiley and Sons, 2001.

[40] G. Shafer and V. Vovk. *Game-theoretic foundations for probability and finance*. John Wiley & Sons, 2019.

[41] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

[42] P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024.

[43] F. Orabona and K.-S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 70(1):436–455, 2023.

[44] T. Cover. Universal portfolios. *Mathematical finance*, 1(1):1–29, 1991.

[45] Thomas M. Cover and Erik Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363, 1996. doi: 10.1109/18.485708.

[46] Shubhanshu Shekhar and Aaditya Ramdas. On the near-optimality of betting confidence sets for bounded means, 2023.

[47] E. Clerico. On the optimality of coin-betting for mean estimation. *International Journal of Approximate Reasoning*, 187, 2025.

[48] Ben Chugg and Aaditya Ramdas. Closed-form empirical bernstein confidence sequences for scalars and matrices, 2025.

[49] John L. Jr. Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926, 1956. doi: 10.1002/j.1538-7305.1956.tb03809.x.

[50] Valentin Kilian, Stefano Cortinovis, and François Caron. Anytime-valid, Bayes-assisted,Prediction-Powered Inference, May 2025.

[51] Jiahua Chen, Asokan Mulayath Variyath, and Bovas Abraham. Adjusted Empirical Likelihood and its Properties. *Journal of Computational and Graphical Statistics*, 17(2):426–443, June 2008. ISSN 1061-8600. doi: 10.1198/106186008X321068.

[52] Sarah C. Emerson and Art B. Owen. Calibration of the empirical likelihood method for a vector mean. *Electronic Journal of Statistics*, 3(none):1161–1192, January 2009. ISSN 1935-7524, 1935-7524. doi: 10.1214/09-EJS518.

[53] Hongtu Zhu, Haibo Zhou, Jiahua Chen, Yimei Li, Jeffrey Lieberman, and Martin Styner. Adjusted exponentially tilted likelihood with applications to brain morphology. *Biometrics*, 65 (3):919–927, September 2009. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2008.01124.x.

[54] Yukun Liu and Jiahua Chen. Adjusted empirical likelihood with high-order precision. *The Annals of Statistics*, 38(3), June 2010. ISSN 0090-5364. doi: 10.1214/09-AOS750.

[55] Stephen G. Walker. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10):1621–1633, October 2013. ISSN 0378-3758. doi: 10.1016/j.jspi.2013.05.013.

[56] Pier Giovanni Bissiri, Chris Holmes, and Stephen Walker. A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, November 2016. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12158.

[57] S. Lyddon, S. Walker, and C. Holmes. Nonparametric learning from Bayesian models with randomized objective functions. *Advances in neural information processing systems*, 31, 2018.

[58] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC Subset Selection in Stochastic Multi-armed Bandits. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1, 2012.

[59] Emilie Kaufmann and Shivaram Kalyanakrishnan. Information Complexity in Bandit Subset Selection. *Proceedings of the 26th Annual Conference on Learning Theory*, 30:228–251, 2013.

[60] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[61] Steven R. Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704 – 1728, 2022. doi: 10.3150/21-BEJ1388. URL `https://doi.org/10.3150/21-BEJ1388`.

[62] Yixuan Fan, Zhanyi Jiao, and Ruodu Wang. Testing the mean and variance by e-processes. *Biometrika*, 112(1):asae049, 02 2025. ISSN 1464-3510. doi: 10.1093/biomet/asae049. URL `https://doi.org/10.1093/biomet/asae049`.

[63] Andrew Richards. University of Oxford Advanced Research Computing, 2015.

[64] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 00087 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, [Cham], January 2015. ISBN 978-3-319-20827-5.

[65] P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic press, 1980.

[66] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Business and Economics. Springer Berlin Heidelberg, 2006. ISBN 9783540295877.

[67] Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

[68] Daniel Csillag, Claudio José Struchiner, and Guilherme Tegoni Goedert. Prediction-powered e-values. *arXiv preprint arXiv:2502.04294*, 2025.

[69] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[71] Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

[72] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83, 1945. ISSN 0099-4987. doi: 10.2307/3001968.

[73] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 0303-6898.

Technical Appendices and Supplementary Material

A Proofs

A.1 Proof of Proposition 1

Proposition 1 (Test-martingale property)**.** *Fix $\mu \in (0,1)$ and $c \in (0,1)$. If $\widehat{\lambda}_i(X_{1:i-1}, \mu)$ is predictable and belongs to $I_{\mu,c}$ for every $i \geq 1$, then $(W_n(X_{1:n}, \mu))_{n\geq 0}$ is nonnegative. Moreover, under the null hypothesis H_μ in* (2)*, $(W_n(X_{1:n}, \mu))_{n\geq 0}$ is a martingale with initial value one.*

Proof. For $\lambda \in I_{\mu,c}$ and $x \in [0,1]$, $1 + \lambda(x - \mu) \geq 1 - c \geq 0$, so the process is nonnegative. Since $\widehat{\lambda}_i(X_{1:i-1}, \mu)$ is $\mathcal{F}_{i-1}$-measurable, under H_μ,

$$\begin{aligned}\mathbb{E}\left[1 + \widehat{\lambda}_i(X_{1:i-1}, \mu)(X_i - \mu) \mid \mathcal{F}_{i-1}\right] &= 1 + \widehat{\lambda}_i(X_{1:i-1}, \mu)\mathbb{E}[X_i - \mu \mid \mathcal{F}_{i-1}] \\ &= 1.\end{aligned}$$

Hence $\mathbb{E}[W_i(X_{1:i}, \mu) \mid \mathcal{F}_{i-1}] = W_{i-1}(X_{1:i-1}, \mu)$. □

A.2 Proof of Proposition 2

Proposition 2 (Anytime-valid coverage)**.** *For any $\alpha \in (0,1)$, $(C_{\alpha,n}(X_{1:n}))_{n\geq 1}$ in Equation* (5) *is a $1 - \alpha$ confidence sequence for the mean parameter μ^*. That is, it satisfies Equation* (1)*.*

Proof. By Proposition 1, $(W_n(X_{1:n}, \mu^*))_{n\geq 0}$ is a nonnegative martingale with initial value one. Ville's inequality [1] gives

$$\mathbb{P}^*\left(\sup_{n\geq 1} W_n(X_{1:n}, \mu^*) > \frac{1}{\alpha}\right) \leq \alpha.$$

On the complement of this event, μ^* belongs to $C_{\alpha,n}(X_{1:n})$ for every $n \geq 1$, which proves the claim. □

A.3 Proof of Theorem 1

We first recall the theorem.

Theorem 1 (Asymptotic log-optimality)**.** *Let $(X_n)_{n\geq 1}$ be i.i.d. in $[0,1]$ from a non-degenerate distribution $\mathbb{P}^*$, and let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Let $(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}))_{n\geq 1}$ be an $\mathcal{F}_{n-1}$-measurable sequence of predictive distributions on $[0,1]$ such that*

$$\mathcal{W}_1(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}), \mathbb{P}^*) \to 0 \qquad \textit{in probability}. \tag{11}$$

Fix $\mu \in (0,1)$ and $c \in (0,1)$. Define

$$M(\lambda, \mu) := \mathbb{E}_{X_1 \sim \mathbb{P}^*}\left[\log\left(1 + \lambda(X_1 - \mu)\right)\right], \tag{12}$$

$$M_n(\lambda, X_{1:n-1}, \mu) := \mathbb{E}_{X'_n \sim \mathbb{Q}_{n|n-1}}\left[\log\left(1 + \lambda(X'_n - \mu)\right) \mid X_{1:n-1}\right]. \tag{13}$$

Let $\widehat{\lambda}_n(X_{1:n-1}, \mu)$ be an $\mathcal{F}_{n-1}$-measurable maximiser of $\lambda \mapsto M_n(\lambda, X_{1:n-1}, \mu)$ on $I_{\mu,c}$. Then $\lambda \mapsto M(\lambda, \mu)$ admits a unique maximiser $\lambda^*(\mu)$ over $I_{\mu,c}$, and

$$\widehat{\lambda}_n(X_{1:n-1}, \mu) \to \lambda^*(\mu) \qquad \textit{in probability}. \tag{14}$$

Moreover, for $W_n(X_{1:n}, \mu) = \prod_{i=1}^n (1 + \widehat{\lambda}_i(X_{1:i-1}, \mu)(X_i - \mu))$,

$$\frac{1}{n}\log W_n(X_{1:n}, \mu) \to M(\lambda^*(\mu), \mu) \qquad \textit{in probability}, \tag{15}$$

$$\frac{1}{n}\mathbb{E}_{\mathbb{P}^*}\left[\log W_n(X_{1:n}, \mu)\right] \to M(\lambda^*(\mu), \mu). \tag{16}$$

Proof. All probabilities and expectations in this proof are taken under the true data-generating law $\mathbb{P}^*$ of the i.i.d. sequence $(X_i)_{i\geq 1}$, unless explicitly stated otherwise. For readability, write

$$I_{\mu,c} := \left[-\frac{c}{1-\mu}, \frac{c}{\mu}\right], \qquad M_n(\lambda) := M_n(\lambda, X_{1:n-1}, \mu), \qquad M(\lambda) := M(\lambda, \mu),$$

and define

$$f_\lambda(x) := \log\{1 + \lambda(x-\mu)\}, \qquad \lambda \in I_{\mu,c},\ x \in [0,1].$$

For every $(\lambda, x) \in I_{\mu,c} \times [0,1]$, we have $1 + \lambda(x-\mu) \geq 1 - c > 0$. Hence f_λ is jointly continuous and uniformly bounded: with

$$B_{\mu,c} := \max\left\{|\log(1-c)|, \log\left(1 + c\frac{1-\mu}{\mu}\right), \log\left(1 + c\frac{\mu}{1-\mu}\right)\right\},$$

we have $|f_\lambda(x)| \leq B_{\mu,c}$ for all $(\lambda, x) \in I_{\mu,c} \times [0,1]$. Moreover, for fixed $\lambda \in I_{\mu,c}$, the map $x \mapsto f_\lambda(x)$ is Lipschitz with a constant that is uniform in λ:

$$|\partial_x f_\lambda(x)| = \left|\frac{\lambda}{1 + \lambda(x-\mu)}\right| \leq \frac{c}{(1-c)\min\{\mu, 1-\mu\}} =: L_{\mu,c}.$$

Therefore, by Kantorovich–Rubinstein duality for $\mathcal{W}_1$ [64, Sec. 1.2],

$$Z_n := \sup_{\lambda \in I_{\mu,c}} |M_n(\lambda) - M(\lambda)| \leq L_{\mu,c}\, \mathcal{W}_1\left(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}), \mathbb{P}^*\right). \tag{25}$$

Under the assumption of Theorem 1, $Z_n \to 0$ in probability. If the convergence in Wasserstein distance is strengthened to almost sure convergence, then the same display gives $Z_n \to 0$ almost surely.

We next show that M has a unique maximiser. For each $x \in [0,1]$, the function $\lambda \mapsto f_\lambda(x)$ is concave on $I_{\mu,c}$, and it is strictly concave whenever $x \neq \mu$. Since $\mathbb{P}^*$ is non-degenerate, $\mathbb{P}^*(X \neq \mu) > 0$. Hence, for any distinct $\lambda_1, \lambda_2 \in I_{\mu,c}$ and $t \in (0,1)$,

$$M(t\lambda_1 + (1-t)\lambda_2) > tM(\lambda_1) + (1-t)M(\lambda_2),$$

so M is strictly concave. Since M is continuous on the compact interval $I_{\mu,c}$, it admits a unique maximiser $\lambda^* = \lambda^*(\mu)$.

We now prove convergence of the maximisers. Fix $\epsilon > 0$ and set

$$A_\epsilon := \{\lambda \in I_{\mu,c} : |\lambda - \lambda^*| \geq \epsilon\}.$$

If $A_\epsilon = \varnothing$, there is nothing to prove. Otherwise, by compactness of A_ϵ and uniqueness of λ^*,

$$\Delta_\epsilon := M(\lambda^*) - \sup_{\lambda \in A_\epsilon} M(\lambda) > 0.$$

On the event $\{\widehat{\lambda}_n \in A_\epsilon\}$, the maximality of $\widehat{\lambda}_n$ for M_n implies

$$\begin{aligned}
\Delta_\epsilon &\leq M(\lambda^*) - M(\widehat{\lambda}_n) \\
&= \{M(\lambda^*) - M_n(\lambda^*)\} + \{M_n(\lambda^*) - M_n(\widehat{\lambda}_n)\} + \{M_n(\widehat{\lambda}_n) - M(\widehat{\lambda}_n)\} \\
&\leq 2Z_n.
\end{aligned}$$

Consequently,

$$\mathbb{P}\{|\widehat{\lambda}_n - \lambda^*| \geq \epsilon\} \leq \mathbb{P}\{Z_n \geq \Delta_\epsilon/2\} \to 0,$$

which proves Equation (14). The same argument also shows that if $Z_n \to 0$ almost surely, then $\widehat{\lambda}_n \to \lambda^*$ almost surely.

Define the log increment

$$Y_n := \log\{1 + \widehat{\lambda}_n(X_n - \mu)\}, \tag{26}$$

and let

$$D_n := Y_n - \mathbb{E}[Y_n \mid \mathcal{F}_{n-1}], \qquad S_n := \sum_{i=1}^n D_i.$$

Then $(S_n)_{n\geq 1}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 1}$. Since $|Y_n| \leq B_{\mu,c}$, we have $|D_n| \leq 2B_{\mu,c}$ and

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}[D_i^2 \mid \mathcal{F}_{i-1}]}{i^2} \leq 4B_{\mu,c}^2 \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty.$$

The martingale law of large numbers [65, Theorem 2.18] gives

$$\frac{S_n}{n} \to 0 \qquad \text{a.s.} \tag{27}$$

Moreover, since X_i is independent of $\mathcal{F}_{i-1}$ and $\widehat{\lambda}_i$ is $\mathcal{F}_{i-1}$-measurable,

$$\mathbb{E}[Y_i \mid \mathcal{F}_{i-1}] = \mathbb{E}\left[\log\{1 + \widehat{\lambda}_i(X_i - \mu)\} \mid \mathcal{F}_{i-1}\right] = M(\widehat{\lambda}_i).$$

Thus

$$\frac{1}{n}\log W_n = \frac{1}{n}\sum_{i=1}^{n} M(\widehat{\lambda}_i) + \frac{S_n}{n}. \tag{28}$$

Because $\widehat{\lambda}_n \to \lambda^*$ in probability and M is continuous and bounded on $I_{\mu,c}$, $M(\widehat{\lambda}_n) \to M(\lambda^*)$ in L^1. Hence, by Cesàro summability,

$$\frac{1}{n}\sum_{i=1}^{n} M(\widehat{\lambda}_i) \to M(\lambda^*) \qquad \text{in } L^1,$$

and therefore also in probability. Combining this convergence with Equations (27) and (28) gives

$$\frac{1}{n}\log W_n \to M(\lambda^*) \qquad \text{in probability,}$$

which proves Equation (15).

Finally,

$$\frac{1}{n}\mathbb{E}[\log W_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[M(\widehat{\lambda}_i)] \to M(\lambda^*)$$

by the same boundedness, L^1 convergence, and Cesàro argument. This proves Equation (16).

Almost-sure counterpart. If the assumption $\mathcal{W}_1(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}), \mathbb{P}^*) \to 0$ holds almost surely rather than merely in probability, then Equation (25) gives uniform convergence $Z_n \to 0$ almost surely. The separation argument for the argmax then yields $\widehat{\lambda}_n \to \lambda^*$ almost surely. Consequently, $M(\widehat{\lambda}_n) \to M(\lambda^*)$ almost surely, and Cesàro summability gives $n^{-1}\sum_{i=1}^{n} M(\widehat{\lambda}_i) \to M(\lambda^*)$ almost surely. Together with the martingale law of large numbers in Equation (27), this yields

$$\frac{1}{n}\log W_n \to M(\lambda^*) \qquad \text{almost surely.}$$

The expectation convergence in Equation (16) remains valid by the bounded L^1 argument above.

A.4 Proof of Theorem 2

We already proved that

$$\sup_{\lambda \in I_{\mu,c}} |M_n(\lambda) - M(\lambda)| \leq L_{\mu,c} \times \mathcal{W}_1\left(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}), \mathbb{P}^*\right)$$

We have

$$\begin{aligned} 0 \leq M(\lambda^*) - M(\widehat{\lambda}_n) &= (M(\lambda^*) - M_n(\lambda^*)) + (M_n(\lambda^*) - M_n(\widehat{\lambda}_n)) + (M_n(\widehat{\lambda}_n) - M(\widehat{\lambda}_n)) \\ &\leq 2 \sup_{\lambda \in I_{\mu,c}} |M_n(\lambda) - M(\lambda)| \\ &\leq 2L_{\mu,c} \times \mathcal{W}_1\left(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}), \mathbb{P}^*\right). \end{aligned}$$

Additionally,

$$\mathbb{E}[\log W_n] = \sum_{i=1}^{n}\mathbb{E}[Y_i] = \sum_{i=1}^{n}\mathbb{E}[\mathbb{E}[Y_i \mid \mathcal{F}_{i-1}]] = \sum_{i=1}^{n}\mathbb{E}[M(\widehat{\lambda}_i)],$$

where Y_n is the log increment defined in Equation (26). It follows that

$$0 \leq M(\lambda^*) - \frac{1}{n}\mathbb{E}[\log W_n] \leq 2L_{\mu,c} \times \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathcal{W}_1\left(\mathbb{Q}_{i|i-1}(\cdot \mid X_{1:i-1}), \mathbb{P}^*\right)\right].$$

A.5 Proof of Proposition 3

Let

$$\omega_n = \frac{\kappa}{\kappa + n}.$$

By convexity of $\mathcal{W}_1$ in its first argument,

$$\mathcal{W}_1(\mathbb{Q}_{n+1|n}, \mathbb{P}^*) \leq \omega_n \mathcal{W}_1(\mathbb{H}_{n+1|n}, \mathbb{P}^*) + (1 - \omega_n)\mathcal{W}_1(\mathbb{P}_n, \mathbb{P}^*).$$

Since both $\mathbb{H}_{n+1|n}$ and $\mathbb{P}^*$ are supported on $[0, 1]$, $\mathcal{W}_1(\mathbb{H}_{n+1|n}, \mathbb{P}^*) \leq 1$. Therefore

$$\mathbb{E}_{\mathbb{P}^*}\{\mathcal{W}_1(\mathbb{Q}_{n+1|n}, \mathbb{P}^*)\} \leq \frac{\kappa}{\kappa + n} + \mathbb{E}_{\mathbb{P}^*}\{\mathcal{W}_1(\mathbb{P}_n, \mathbb{P}^*)\}.$$

The second term is $O(n^{-1/2})$ for distributions supported on a bounded interval, so the displayed expectation is $O(n^{-1/2})$. Markov's inequality gives convergence in $\mathbb{P}^*$-probability, and Theorem 1 gives the final claim.

A.6 Proof of Proposition 4

We first recall the proposition.

Proposition 4 (Consistency of the RETEL predictive)**.** *Assume that $X_1, X_2, \ldots$ are i.i.d. from a non-degenerate distribution $\mathbb{P}^*$ supported on $[0, 1]$, with mean $\mu^* \in (0, 1)$ and variance $\sigma^2 > 0$. Assume that π is absolutely continuous on $[0, 1]$, with density bounded away from zero in a neighbourhood of μ^*, that $\tau_n = o(n)$, and that each $\mathbb{H}_\mu$ is supported on $[0, 1]$ with $\operatorname{conv}(\operatorname{supp} \mathbb{H}_\mu) = [0, 1]$. Then, for every $\epsilon > 0$,*

$$\Pi_n\big(|\mu - \mu^*| > \epsilon\big) \to 0, \quad \textit{and} \quad \Pi_n\big(|\gamma_{n,\mu}| > \epsilon\big) \to 0, \tag{24}$$

in probability. Moreover, in probability, $\mathcal{W}_1\big(\mathbb{Q}_{n+1|n}(\cdot \mid X_{1:n}), \mathbb{P}^*\big) \to 0$.

Proof. Write

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \qquad s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

and define the rescaled pseudo-likelihood

$$R_n(\mu) = \prod_{i=1}^n n w_{n,i}(\mu).$$

The factor n^n is independent of μ, so R_n and L_n induce the same pseudo-posterior. In the unregularized case $\tau_n = 0$, we use the convention from the main text that $R_n(\mu) = 0$ whenever $\mu \notin (\min_i X_i, \max_i X_i)$. At points where the likelihood is zero, the value assigned to $\gamma_{n,\mu}$ is irrelevant for the posterior statements below.

Lemma 1 (Separation from the empirical mean)**.** *For every $t \in (0, 1]$,*

$$\sup_{\mu \in [0,1]:\, |\mu - \bar{X}_n| \geq t} R_n(\mu) \leq \exp\{-nc(t)\}, \qquad c(t) = \min\left\{-\log(1 - t/2),\, \frac{t^2}{2}\right\} > 0.$$

Proof. Fix μ. If $R_n(\mu) = 0$, there is nothing to prove. Otherwise, write

$$a = \sum_{i=1}^n w_{n,i}(\mu) = 1 - w_{n,H}(\mu) > 0, \qquad p_i = \frac{w_{n,i}(\mu)}{a}, \quad i = 1, \ldots, n.$$

Then $p = (p_1, \ldots, p_n)$ is a probability vector. Since $\widetilde{\mathbb{P}}_{n,\mu}$ has mean μ, and its regularizing component is supported on $[0, 1]$, there is an $m_H \in [0, 1]$ such that

$$\mu = a \sum_{i=1}^n p_i X_i + (1 - a) m_H.$$

Let $m_p = \sum_i p_i X_i$ and let $u_i = 1/n$. The last display implies $|\mu - m_p| \leq 1 - a$. Moreover,

$$\log R_n(\mu) = \sum_{i=1}^n \log(n w_{n,i}(\mu)) = n \log a + \sum_{i=1}^n \log(n p_i) = n \log a - n\, \mathrm{KL}(u \| p),$$

with the usual convention that $\mathrm{KL}(u\|p) = \infty$ if $p_i = 0$ for some i.

Suppose $|\mu - \bar{X}_n| \geq t$. If $1 - a \geq t/2$, then

$$\log R_n(\mu) \leq n \log a \leq n \log(1 - t/2).$$

If instead $1 - a < t/2$, then $|m_p - \bar{X}_n| \geq t/2$. Since $X_i \in [0, 1]$,

$$|m_p - \bar{X}_n| = \left| \sum_{i=1}^{n} (p_i - u_i) X_i \right| \leq \mathrm{TV}(p, u),$$

where $\mathrm{TV}(p, u) = \frac{1}{2} \sum_i |p_i - u_i|$. Pinsker's inequality therefore gives

$$\mathrm{KL}(u\|p) \geq 2\,\mathrm{TV}(u, p)^2 \geq \frac{t^2}{2}.$$

Thus $\log R_n(\mu) \leq -nt^2/2$ in the second case. Combining the two cases proves the lemma. □

Lemma 2 (Local control of the tilting parameter)**.** *For every $v_0 > 0$ there exist constants $\eta > 0$ and $C < \infty$, depending only on v_0, such that, for all sufficiently large n, whenever $s_n^2 \geq v_0$,*

$$|\gamma_{n,\mu}| \leq C \left(|\mu - \bar{X}_n| + \frac{\tau_n}{n} \right)$$

for every μ for which $|\mu - \bar{X}_n| \leq \eta$ and the tilt equation is defined.

Proof. For fixed μ, write

$$\psi_{n,\mu}(\gamma) = \log \int e^{\gamma(z-\mu)}\, d\mathbb{P}_{n,\mu}(z).$$

The tilting parameter solves $\psi'_{n,\mu}(\gamma_{n,\mu}) = 0$. Let $h_{n,\mu} = \int z\, d\mathbb{H}_\mu(z) \in [0, 1]$. The mean of the base measure is

$$m_{n,\mu} = \int z\, d\mathbb{P}_{n,\mu}(z) = \frac{n}{n + \tau_n} \bar{X}_n + \frac{\tau_n}{n + \tau_n} h_{n,\mu},$$

so

$$|\psi'_{n,\mu}(0)| = |m_{n,\mu} - \mu| \leq |\bar{X}_n - \mu| + \frac{\tau_n}{n + \tau_n}. \tag{29}$$

Moreover, by the variance decomposition for mixtures,

$$\psi''_{n,\mu}(0) = \mathrm{var}_{\mathbb{P}_{n,\mu}}(Z) \geq \frac{n}{n + \tau_n} s_n^2.$$

If $s_n^2 \geq v_0$, then, since $\tau_n = o(n)$, the last quantity is bounded below by $v_0/2$ for all sufficiently large n.

Fix a small constant $\delta > 0$. For $|\gamma| \leq \delta$, let $\mathbb{P}^\gamma_{n,\mu}$ be the exponentially tilted version of $\mathbb{P}_{n,\mu}$. Since $z - \mu \in [-1, 1]$,

$$\frac{d\mathbb{P}^\gamma_{n,\mu}}{d\mathbb{P}_{n,\mu}}(z) = \exp\{\gamma(z - \mu) - \psi_{n,\mu}(\gamma)\} \geq e^{-2\delta}.$$

Using

$$\mathrm{var}_Q(Z) = \frac{1}{2} \int\int (z - z')^2\, dQ(z) dQ(z'),$$

we obtain

$$\psi''_{n,\mu}(\gamma) = \mathrm{var}_{\mathbb{P}^\gamma_{n,\mu}}(Z) \geq e^{-4\delta} \mathrm{var}_{\mathbb{P}_{n,\mu}}(Z) \geq e^{-4\delta} v_0/2 =: c_0$$

for all $|\gamma| \leq \delta$ and all large n.

Choose $\eta > 0$ such that $\eta < c_0\delta/2$. Since $\tau_n/n \to 0$, (29) implies that, uniformly over $|\mu - \bar{X}_n| \leq \eta$,

$$|\psi'_{n,\mu}(0)| < c_0\delta$$

for all large n. Hence $\psi'_{n,\mu}(-\delta) < 0 < \psi'_{n,\mu}(\delta)$, and the root lies in $(-\delta, \delta)$. By the mean-value theorem,

$$|\gamma_{n,\mu}| \leq c_0^{-1} |\psi'_{n,\mu}(0)| \leq c_0^{-1} \left(|\mu - \bar{X}_n| + \frac{\tau_n}{n + \tau_n} \right),$$

which proves the claim, after increasing the constant if necessary. □

Lemma 3 (Subexponential denominator lower bound)**.** *Let*

$$D_n = \int_0^1 R_n(\mu)\pi(d\mu).$$

Under the assumptions of Proposition 4, with $\mathbb{P}^*$-probability tending to one,

$$D_n \geq \frac{c}{n}\exp\{-C(1+\tau_n)\}$$

for constants $c, C > 0$ independent of n. In particular, $\log D_n = o_{\mathbb{P}^*}(n)$.

Proof. Since $\bar{X}_n \to \mu^*$ and $s_n^2 \to \sigma^2 > 0$ almost surely, and since $\mu^* \in (0,1)$, with probability tending to one the interval

$$B_n = \left[\bar{X}_n - \frac{1}{n}, \bar{X}_n + \frac{1}{n}\right]$$

is contained in $(0,1)$, lies in a neighbourhood on which the prior density is bounded below, and has $s_n^2 \geq \sigma^2/2$. Applying Lemma 2 with $v_0 = \sigma^2/2$, we also have the local bound in that lemma on this event. In the unregularized case $\tau_n = 0$, the same event may also be taken to include $B_n \subset (\min_i X_i, \max_i X_i)$: this holds with probability tending to one because $\sigma^2 > 0$ implies positive probability both below and above μ^*.

On this event, for every $\mu \in B_n$,

$$|\gamma_{n,\mu}| \leq C_1 \frac{1+\tau_n}{n}.$$

Using $|X_i - \mu| \leq 1$ and $M_{\mathbb{H}_\mu}(\gamma_{n,\mu}) \leq e^{|\gamma_{n,\mu}|}$, we have

$$nw_{n,i}(\mu) = \frac{ne^{\gamma_{n,\mu}(X_i-\mu)}}{\sum_{j=1}^n e^{\gamma_{n,\mu}(X_j-\mu)} + \tau_n M_{\mathbb{H}_\mu}(\gamma_{n,\mu})} \geq \frac{n}{n+\tau_n} e^{-2|\gamma_{n,\mu}|}.$$

Therefore, uniformly over $\mu \in B_n$,

$$R_n(\mu) \geq \left(\frac{n}{n+\tau_n}\right)^n \exp\{-2n|\gamma_{n,\mu}|\} \geq \exp\{-C_2(1+\tau_n)\},$$

where we used $\log(1+x) \leq x$. The prior mass of B_n is at least c/n on the same event, so

$$D_n \geq \int_{B_n} R_n(\mu)\pi(d\mu) \geq \frac{c}{n}\exp\{-C_2(1+\tau_n)\}.$$

Since $\tau_n = o(n)$, this gives $\log D_n = o_{\mathbb{P}^*}(n)$. □

We now prove the posterior concentration assertions. Fix $\epsilon > 0$. If $\epsilon > 1$, then $\{\mu \in [0,1] : |\mu - \mu^*| > \epsilon\}$ is empty, so assume $\epsilon \leq 1$. Let

$$A_\epsilon = \{\mu \in [0,1] : |\mu - \mu^*| > \epsilon\}.$$

On the event $|\bar{X}_n - \mu^*| \leq \epsilon/2$, $A_\epsilon \subseteq \{\mu : |\mu - \bar{X}_n| \geq \epsilon/2\}$. Hence, by Lemma 1,

$$\int_{A_\epsilon} R_n(\mu)\pi(d\mu) \leq \exp\{-nc(\epsilon/2)\}.$$

Combining this exponential numerator bound with the subexponential lower bound from Lemma 3 gives

$$\Pi_n(A_\epsilon) = \frac{\int_{A_\epsilon} R_n(\mu)\pi(d\mu)}{D_n} \to 0$$

in $\mathbb{P}^*$-probability. This proves the first part of (24).

Next fix $\epsilon > 0$. Let η and C be as in Lemma 2, and choose $\rho \in (0, \eta)$ such that $C\rho < \epsilon/2$. Since $\tau_n/n \to 0$, with probability tending to one,

$$|\mu - \bar{X}_n| \leq \rho \quad \Longrightarrow \quad |\gamma_{n,\mu}| \leq C\left(\rho + \frac{\tau_n}{n}\right) < \epsilon.$$

Thus

$$\Pi_n(|\gamma_{n,\mu}| > \epsilon) \le \Pi_n(|\mu - \bar{X}_n| > \rho) + o_{\mathbb{P}^*}(1).$$

Furthermore,

$$\Pi_n(|\mu - \bar{X}_n| > \rho) \le \Pi_n(|\mu - \mu^*| > \rho/2) + \mathbf{1}\{|\bar{X}_n - \mu^*| > \rho/2\}.$$

The first term tends to zero by the first part of (24), and the second term by the law of large numbers. Hence the second part of (24) follows.

It remains to prove predictive consistency. We first record a deterministic bound.

Lemma 4. *For every μ for which the fitted distribution is defined,*

$$\mathcal{W}_1(\widetilde{\mathbb{P}}_{n,\mu}, \mathbb{P}_n) \le 2|\gamma_{n,\mu}| + \frac{\tau_n}{n+\tau_n} + \frac{\tau_n}{n} e^{2|\gamma_{n,\mu}|}. \tag{30}$$

Proof. Indeed, by Kantorovich–Rubinstein duality, or by the one-dimensional CDF formula for $\mathcal{W}_1$, and because all distributions are supported on $[0,1]$,

$$\mathcal{W}_1(\widetilde{\mathbb{P}}_{n,\mu}, \mathbb{P}_n) \le \sum_{i=1}^{n} \left| w_{n,i}(\mu) - \frac{1}{n} \right| + w_{n,H}(\mu).$$

The regularizer weight is bounded by

$$w_{n,H}(\mu) \le \frac{\tau_n e^{|\gamma_{n,\mu}|}}{n e^{-|\gamma_{n,\mu}|}} = \frac{\tau_n}{n} e^{2|\gamma_{n,\mu}|}.$$

For the empirical weights, view

$$w_{n,i}(\mu;\gamma) = \frac{e^{\gamma(X_i-\mu)}}{\sum_{j=1}^{n} e^{\gamma(X_j-\mu)} + \tau_n M_{\mathbb{H}_\mu}(\gamma)}.$$

Then

$$\frac{d}{d\gamma} w_{n,i}(\mu;\gamma) = w_{n,i}(\mu;\gamma)\{(X_i - \mu) - m(\gamma)\},$$

where $m(\gamma) \in [-1,1]$. Therefore

$$\sum_{i=1}^{n} \left| \frac{d}{d\gamma} w_{n,i}(\mu;\gamma) \right| \le 2,$$

and hence

$$\sum_{i=1}^{n} |w_{n,i}(\mu;\gamma) - w_{n,i}(\mu;0)| \le 2|\gamma|.$$

Since $w_{n,i}(\mu;0) = 1/(n+\tau_n)$,

$$\sum_{i=1}^{n} \left| w_{n,i}(\mu) - \frac{1}{n} \right| \le 2|\gamma_{n,\mu}| + \frac{\tau_n}{n+\tau_n},$$

which proves (30). □

Finally,

$$\begin{aligned}\mathcal{W}_1(\mathbb{Q}_{n+1|n}, \mathbb{P}^*) &\le \mathcal{W}_1(\mathbb{P}_n, \mathbb{P}^*) + \mathcal{W}_1(\mathbb{Q}_{n+1|n}, \mathbb{P}_n) \\ &\le \mathcal{W}_1(\mathbb{P}_n, \mathbb{P}^*) + \int_0^1 \mathcal{W}_1(\widetilde{\mathbb{P}}_{n,\mu}, \mathbb{P}_n)\Pi_n(d\mu).\end{aligned}$$

The empirical term satisfies $\mathcal{W}_1(\mathbb{P}_n, \mathbb{P}^*) \to 0$ almost surely, because $[0,1]$ is compact and $\mathbb{P}_n$ converges weakly to $\mathbb{P}^*$ almost surely. For the second term, fix $\delta > 0$. Since $\mathcal{W}_1 \le 1$ on $[0,1]$, (30) gives

$$\int_0^1 \mathcal{W}_1(\widetilde{\mathbb{P}}_{n,\mu}, \mathbb{P}_n)\Pi_n(d\mu) \le 2\delta + \frac{\tau_n}{n+\tau_n} + \frac{\tau_n}{n} e^{2\delta} + \Pi_n(|\gamma_{n,\mu}| > \delta).$$

Taking $n \to \infty$ and using $\tau_n = o(n)$ and the second part of (24), the limsup of the second term is at most 2δ in $\mathbb{P}^*$-probability. Since $\delta > 0$ is arbitrary, this term converges to zero in $\mathbb{P}^*$-probability, proving the consistency of the RETEL predictive in Wasserstein distance.

B Asymptotic log-optimality for multivariate bounded means

We extend Theorem 1 to the vector betting rule of Section 7. For a fixed candidate mean μ and $c \in (0,1)$, write

$$\Lambda_{\mu,c} = \big\{\lambda \in \mathbb{R}^d : \inf_{x\in K}\{1+\lambda^\top(x-\mu)\} \geq 1-c\big\}, \qquad f_\lambda(x) = \log\{1+\lambda^\top(x-\mu)\}.$$

Theorem 3 (Multivariate asymptotic log-optimality)**.** *Let $(X_n)_{n\geq 1}$ be i.i.d. from a distribution $\mathbb{P}^*$ supported on a known compact convex set $K \subset \mathbb{R}^d$, and suppose that X_1 has positive definite covariance. Fix μ in the interior of K and $c \in (0,1)$. Let $\mathcal{F}_n = \sigma(X_1,\dots,X_n)$ and let $\mathbb{Q}_{n|n-1}$ be an $\mathcal{F}_{n-1}$-measurable predictive distribution on K such that $\mathcal{W}_1(\mathbb{Q}_{n|n-1},\mathbb{P}^*) \to 0$ in probability, where $\mathcal{W}_1$ uses Euclidean distance. Define*

$$M(\lambda,\mu) = \mathbb{E}_{\mathbb{P}^*}[f_\lambda(X_1)], \qquad M_n(\lambda, X_{1:n-1},\mu) = \int_K f_\lambda(x)\,\mathbb{Q}_{n|n-1}(dx \mid X_{1:n-1}).$$

Let $\widehat{\lambda}_n(\mu)$ be an $\mathcal{F}_{n-1}$-measurable maximiser of M_n over $\Lambda_{\mu,c}$, and set $W_n(X_{1:n},\mu) = \prod_{i=1}^n\{1+\widehat{\lambda}_i(\mu)^\top(X_i-\mu)\}$. Then $M(\cdot,\mu)$ has a unique maximiser $\lambda^*(\mu)$ over $\Lambda_{\mu,c}$, and $\widehat{\lambda}_n(\mu) \to \lambda^*(\mu)$ in probability. Moreover,

$$\frac{1}{n}\log W_n(X_{1:n},\mu) \longrightarrow M(\lambda^*(\mu),\mu) \qquad \textit{in probability},$$
$$\frac{1}{n}\mathbb{E}_{\mathbb{P}^*}[\log W_n(X_{1:n},\mu)] \longrightarrow M(\lambda^*(\mu),\mu).$$

Proof. Fix μ and c throughout. Write $Q_n = \mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1})$, $M_n(\lambda) = M_n(\lambda, X_{1:n-1},\mu)$ and $M(\lambda) = M(\lambda,\mu)$. Expectations and probabilities below are under the true law of the data. All norms are Euclidean.

1. The admissible set and uniform bounds. Since μ is in the interior of K, there is an $r>0$ such that the closed ball of radius r about μ lies in K. If $\lambda \in \Lambda_{\mu,c}$ is nonzero, then $x = \mu - r\lambda/\|\lambda\|$ belongs to that ball. Admissibility at this x gives $1 - r\|\lambda\| \geq 1-c$, so $\|\lambda\| \leq R := c/r$. Moreover,

$$\Lambda_{\mu,c} = \bigcap_{x\in K}\{\lambda \in \mathbb{R}^d : \lambda^\top(x-\mu) \geq -c\}.$$

This intersection of closed half-spaces is closed and convex, contains zero, and is bounded by the preceding calculation. It is therefore compact.

Put $D = \operatorname{diam}(K)$. For all $\lambda \in \Lambda_{\mu,c}$ and $x \in K$, the lower bound follows from admissibility and the upper bound from Cauchy–Schwarz. Together they give

$$1-c \leq 1+\lambda^\top(x-\mu) \leq 1+RD.$$

Thus $|f_\lambda(x)| \leq B := \max\{-\log(1-c), \log(1+RD)\}$. On $[1-c,\infty)$ the derivative of the logarithm is at most $1/(1-c)$. Applying the mean-value theorem, first with x varying and then with λ varying, gives

$$|f_\lambda(x) - f_\lambda(y)| \leq \frac{|\lambda^\top(x-y)|}{1-c} \leq L\|x-y\|, \qquad L := \frac{R}{1-c},$$
$$|f_\lambda(x) - f_{\lambda'}(x)| \leq \frac{|(\lambda-\lambda')^\top(x-\mu)|}{1-c} \leq \frac{D}{1-c}\|\lambda-\lambda'\|.$$

The second inequality shows that M and, at each history, M_n are continuous in λ. Then, both attain their maxima on the compact $\Lambda_{\mu,c}$. For each fixed λ, $M_n(\lambda)$ is $\mathcal{F}_{n-1}$-measurable because Q_n is a measurable probability kernel. The measurable maximum theorem [66, Theorem 18.19] for a continuous objective on a fixed compact metric space therefore gives an $\mathcal{F}_{n-1}$-measurable choice from its maximisers.

2. A uniform bound on the predictive objective. Fix a realised history, so that Q_n is a fixed probability measure. Let $\Gamma(Q_n,\mathbb{P}^*)$ be the set of probability measures γ on $K\times K$ whose first marginal is Q_n

and whose second marginal is $\mathbb{P}^*$. For any $\gamma \in \Gamma(Q_n, \mathbb{P}^*)$ and any admissible λ, the two marginal identities and the first inequality above give

$$\begin{aligned} |M_n(\lambda) - M(\lambda)| &= \left| \int_K f_\lambda \, dQ_n - \int_K f_\lambda \, d\mathbb{P}^* \right| \\ &= \left| \int_{K \times K} \{f_\lambda(x) - f_\lambda(y)\} \, \gamma(dx, dy) \right| \\ &\leq \int_{K \times K} |f_\lambda(x) - f_\lambda(y)| \, \gamma(dx, dy) \\ &\leq L \int_{K \times K} \|x - y\| \, \gamma(dx, dy). \end{aligned}$$

The left-hand side is the same for every coupling. By definition, the infimum of the last integral over $\Gamma(Q_n, \mathbb{P}^*)$ is $\mathcal{W}_1(Q_n, \mathbb{P}^*)$. The constant L is the same for every admissible stake. Taking first the infimum over couplings and then the supremum over stakes yields

$$Z_n := \sup_{\lambda \in \Lambda_{\mu,c}} |M_n(\lambda) - M(\lambda)| \leq L \mathcal{W}_1(Q_n, \mathbb{P}^*). \tag{31}$$

By continuity, the supremum equals the supremum over any fixed countable dense subset of $\Lambda_{\mu,c}$, so Z_n is measurable. Predictive consistency implies $Z_n \to 0$ in probability.

3. Uniqueness of the oracle and convergence of the stakes. Write Σ for the covariance matrix of X_1. For distinct admissible stakes λ, λ', put $v = \lambda - \lambda'$. Then $\operatorname{var}(v^\top (X_1 - \mu)) = v^\top \Sigma v > 0$, so the two positive wealth factors $1 + \lambda^\top (X_1 - \mu)$ and $1 + {\lambda'}^\top (X_1 - \mu)$ differ with positive probability. For $0 < t < 1$ the wealth factor at $t\lambda + (1-t)\lambda'$ is their convex combination. Concavity of the logarithm gives a pointwise concavity inequality which is strict whenever the two factors differ. Taking expectations therefore gives

$$M(t\lambda + (1-t)\lambda') > tM(\lambda) + (1-t)M(\lambda').$$

Thus M is strictly concave on the convex admissible set and its maximiser $\lambda^* := \lambda^*(\mu)$ is unique.

Write $\widehat{\lambda}_n = \widehat{\lambda}_n(\mu)$. Since $M_n(\lambda^*) - M_n(\widehat{\lambda}_n) \leq 0$, adding and subtracting both predictive objective values yields the pathwise bound

$$\begin{aligned} 0 \leq M(\lambda^*) - M(\widehat{\lambda}_n) = [M(\lambda^*) - M_n(\lambda^*)] + [M_n(\lambda^*) - M_n(\widehat{\lambda}_n)] \\ + [M_n(\widehat{\lambda}_n) - M(\widehat{\lambda}_n)] \leq 2Z_n. \end{aligned} \tag{32}$$

The same uniform bound applies at the random stake because it holds for every stake at each realised history. Fix $\varepsilon > 0$ and let $A_\varepsilon = \{\lambda \in \Lambda_{\mu,c} : \|\lambda - \lambda^*\| \geq \varepsilon\}$. If A_ε is empty, then $\|\widehat{\lambda}_n - \lambda^*\| < \varepsilon$ for all n. Otherwise A_ε is compact and excludes the unique maximiser, so $\delta_\varepsilon = M(\lambda^*) - \max_{\lambda \in A_\varepsilon} M(\lambda) > 0$. On the event $\{\widehat{\lambda}_n \in A_\varepsilon\}$, the left-hand gap in (32) is at least δ_ε. Hence

$$\mathbb{P}(\|\widehat{\lambda}_n - \lambda^*\| \geq \varepsilon) \leq \mathbb{P}(Z_n \geq \delta_\varepsilon / 2) \longrightarrow 0.$$

4. Realised and expected log-growth. Let $Y_i = f_{\widehat{\lambda}_i}(X_i)$. Conditional on $\mathcal{F}_{i-1}$, the stake $\widehat{\lambda}_i$ is known and X_i has law $\mathbb{P}^*$ by independence. Thus

$$\mathbb{E}[Y_i \mid \mathcal{F}_{i-1}] = \int_K f_{\widehat{\lambda}_i}(x) \, \mathbb{P}^*(dx) = M(\widehat{\lambda}_i).$$

The centred increments $D_i = Y_i - M(\widehat{\lambda}_i)$ have conditional mean zero. For $i < j$, D_i is $\mathcal{F}_{j-1}$-measurable, so $\mathbb{E}[D_i D_j] = \mathbb{E}[D_i \, \mathbb{E}(D_j \mid \mathcal{F}_{j-1})] = 0$. Also $\mathbb{E}[D_i^2] = \mathbb{E}[\operatorname{var}(Y_i \mid \mathcal{F}_{i-1})] \leq \mathbb{E}[Y_i^2] \leq B^2$. Expanding the square and using these two facts gives

$$\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n D_i \right)^2 \right] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[D_i^2] \leq \frac{B^2}{n}.$$

Since $\log W_n(X_{1:n},\mu) = \sum_{i=1}^n Y_i$, subtracting $M(\lambda^*)$ gives

$$\frac{1}{n}\log W_n(X_{1:n},\mu) - M(\lambda^*) = \frac{1}{n}\sum_{i=1}^n D_i - \frac{1}{n}\sum_{i=1}^n \{M(\lambda^*) - M(\widehat{\lambda}_i)\}.$$

The terms in braces are nonnegative. Applying Cauchy–Schwarz to the first average and (31)–(32) to the second yields

$$\mathbb{E}\left|\frac{1}{n}\log W_n(X_{1:n},\mu) - M(\lambda^*)\right| \le \frac{B}{\sqrt{n}} + \frac{2L}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{W}_1(Q_i,\mathbb{P}^*)].$$

Both laws are supported on K, so $\mathcal{W}_1(Q_i,\mathbb{P}^*) \le D$. For every $\eta > 0$, $\mathbb{E}[\mathcal{W}_1(Q_i,\mathbb{P}^*)] \le \eta + D\,\mathbb{P}(\mathcal{W}_1(Q_i,\mathbb{P}^*) > \eta)$. First letting $i \to \infty$ and then $\eta \downarrow 0$ shows that these expectations tend to zero. Their Cesàro averages therefore tend to zero as well. This proves L^1 convergence of $n^{-1}\log W_n(X_{1:n},\mu)$ to $M(\lambda^*)$. Convergence in probability follows by Markov's inequality. The expected log-growth conclusion follows from $|\mathbb{E}[n^{-1}\log W_n(X_{1:n},\mu)] - M(\lambda^*)| \le \mathbb{E}|n^{-1}\log W_n(X_{1:n},\mu) - M(\lambda^*)|$. □

Scalar specialisation. For $d = 1$, $K = [0,1]$ and $0 < \mu < 1$, the affine wealth factor attains its minimum at an endpoint. Its two constraints are $1 - \lambda\mu \ge 1 - c$ and $1 + \lambda(1-\mu) \ge 1 - c$, giving exactly $\Lambda_{\mu,c} = [-c/(1-\mu), c/\mu] = I_{\mu,c}$. Positive definite covariance reduces to positive variance, and taking $r = \min\{\mu, 1-\mu\}$ gives $L = c/[(1-c)\min\{\mu, 1-\mu\}]$. Thus the theorem yields all three conclusions of Theorem 1 with the same admissible interval and Lipschitz constant.

C Computational details for BETEL/RETEL predictives

For fixed μ, let

$$M_{\mathbb{H}_\mu}(\gamma) = \int e^{\gamma(z-\mu)}\, d\mathbb{H}_\mu(z), \qquad Z_{n,\mu} = \sum_{j=1}^n e^{\gamma_{n,\mu}(X_j-\mu)} + \tau_n M_{\mathbb{H}_\mu}(\gamma_{n,\mu}).$$

Then the tilted distribution in (22) can be written as

$$\widetilde{\mathbb{P}}_{n,\mu} = \sum_{i=1}^n w_{n,i}(\mu)\delta_{X_i} + w_{n,H}(\mu)\widetilde{\mathbb{H}}_{n,\mu}, \tag{33}$$

where

$$w_{n,i}(\mu) = \frac{e^{\gamma_{n,\mu}(X_i-\mu)}}{Z_{n,\mu}}, \qquad i = 1,\dots,n, \tag{34}$$

$$w_{n,H}(\mu) = \frac{\tau_n M_{\mathbb{H}_\mu}(\gamma_{n,\mu})}{Z_{n,\mu}}, \tag{35}$$

and

$$\widetilde{\mathbb{H}}_{n,\mu}(A) = \frac{\int_A e^{\gamma_{n,\mu}(z-\mu)}\, d\mathbb{H}_\mu(z)}{M_{\mathbb{H}_\mu}(\gamma_{n,\mu})}.$$

Some regularized ETEL likelihoods include the atom correction

$$\prod_{i=1}^n \{1 + \tau_n \mathbb{H}_\mu(\{X_i\})\}.$$

This factor is equal to one for continuous regularizers and is independent of μ for the two-point regularizer used in our experiments, so it does not change the pseudo-posterior.

For the two-point choice $\mathbb{H} = \frac{1}{2}(\delta_0 + \delta_1)$, define

$$S_0(\gamma) = \sum_{i=1}^n e^{\gamma X_i}, \qquad S_1(\gamma) = \sum_{i=1}^n X_i e^{\gamma X_i}.$$

The tilting equation (22) becomes

$$\mu = \frac{S_1(\gamma_{n,\mu}) + \frac{\tau_n}{2} e^{\gamma_{n,\mu}}}{S_0(\gamma_{n,\mu}) + \frac{\tau_n}{2}(1 + e^{\gamma_{n,\mu}})}. \tag{36}$$

The corresponding weights are

$$w_{n,i}(\mu) = \frac{e^{\gamma_{n,\mu} X_i}}{S_0(\gamma_{n,\mu}) + \frac{\tau_n}{2}(1 + e^{\gamma_{n,\mu}})}, \qquad w_{n,H}(\mu) = \frac{\frac{\tau_n}{2}(1 + e^{\gamma_{n,\mu}})}{S_0(\gamma_{n,\mu}) + \frac{\tau_n}{2}(1 + e^{\gamma_{n,\mu}})}. \tag{37}$$

Finally, the tilted regularizer remains supported on $\{0, 1\}$:

$$\widetilde{\mathbb{H}}_{n,\mu} = (1 - p_{n,\mu})\delta_0 + p_{n,\mu}\delta_1, \qquad p_{n,\mu} = \frac{e^{\gamma_{n,\mu}}}{1 + e^{\gamma_{n,\mu}}}.$$

D Implementation details

The methods proposed in this paper were implemented in Python 3.12. The code to reproduce all experiments is available at `https://github.com/ValentinKil/bacsbm.git`. Except for the LLM experiment, all experiments were run on a CPU-only cluster with a mix of Intel and AMD nodes (Intel Xeon Platinum 8268, AMD EPYC 9634, and AMD EPYC 9645). The LLM experiment was run on an Apple M4 Max system with 36 GB of RAM. A runtime comparison for a subset of the synthetic experiments is reported separately by processor in Table 2.

D.1 Computation of predictive-assisted confidence sequences

All confidence sequences based on the predictive-assisted construction in Section 4 are computed by discretising the test inversion (5). For G equal subdivisions of $[0, 1]$, we use the equally spaced grid

$$\mathcal{G} = \left\{ \frac{1}{G}, \frac{2}{G}, \ldots, \frac{G-1}{G} \right\} \subset (0, 1).$$

For each candidate mean $\mu \in \mathcal{G}$ and each time n, the method computes a predictable coefficient

$$\widehat{\lambda}_n(\mu) \in I_{\mu,c} = \left[-\frac{c}{1-\mu}, \frac{c}{\mu} \right],$$

for a given truncation parameter $c \in (0, 1)$, using only the history $X_{1:n-1}$. The corresponding test-martingale value is then updated through the one-step factor

$$1 + \widehat{\lambda}_t(\mu)(X_t - \mu).$$

For numerical stability, the implementation works with cumulative log-values,

$$\log W_n(X_{1:n}, \mu) = \sum_{i=1}^{n} \log\{1 + \widehat{\lambda}_i(\mu)(X_i - \mu)\}, \qquad \mu \in \mathcal{G}.$$

At time n, the retained grid points are

$$\mathcal{A}_n = \{\mu \in \mathcal{G} : \log W_n(X_{1:n}, \mu) \le \log(1/\alpha)\}.$$

The reported discretised confidence set is the smallest interval covering the retained grid points, enlarged by one grid step and clipped to $[0, 1]$:

$$\widetilde{C}_{\alpha,n} = \left[\max\{0, \min \mathcal{A}_n - G^{-1}\}, \min\{1, \max \mathcal{A}_n + G^{-1}\}\right].$$

The enlargement by one grid step accounts for the finite grid resolution when reporting an interval from the retained grid points. Unless otherwise stated, the displayed confidence sequence is the running intersection

$$C^{\mathrm{run}}_{\alpha,n} = \bigcap_{s=1}^{n} \widetilde{C}_{\alpha,s},$$

which enforces nested intervals while preserving anytime-validity [see, e.g., 7].

For a generic predictable distribution $\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1})$, the coefficient is computed according to (10): for each $\mu \in \mathcal{G}$,

$$\widehat{\lambda}_n(X_{1:n-1}, \mu) \in \arg\max_{\lambda \in I_{\mu,c}} \left\{ \mathbb{E}_{\widetilde{X}_n \sim \mathbb{Q}_{n|n-1}(\cdot|X_{1:n-1})} \left[\log\left(1 + \lambda(\widetilde{X}_n - \mu)\right) \right] \right\},$$

which is a concave objective in λ. In the implementation, the maximiser is obtained by solving the first-order condition

$$\mathbb{E}_{\widetilde{X}_n \sim \mathbb{Q}_{n|n-1}(\cdot|X_{1:n-1})} \left[\frac{\widetilde{X}_n - \mu}{1 + \lambda(\widetilde{X}_n - \mu)} = 0 \right]$$

whenever the solution lies in the interior of $I_{\mu,c}$. Otherwise, the appropriate endpoint of $I_{\mu,c}$ is selected. For discrete predictive distributions, the integral above is evaluated as a finite weighted sum.

D.2 Predictive distributions

The implemented methods differ only in the construction of the predictable distribution $\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1})$ used in (10).

Empirical predictive. The empirical method uses the empirical distribution of the previous observations,

$$\mathbb{P}_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n-1} \delta_{X_i}, \qquad n \geq 2,$$

as the predictive distribution. Thus, for each candidate value μ, the coefficient is computed by maximising

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \log\{1 + \lambda(X_i - \mu)\}$$

over $\lambda \in I_{\mu,c}$. Equivalently, when the optimum is interior, it solves

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0. \tag{38}$$

Parametric Bayesian predictive. The parametric Bayesian method uses a beta working model in mean–concentration parametrisation,

$$X_i \mid \rho, \nu \sim \mathrm{Beta}(\rho\nu, (1-\rho)\nu), \qquad \rho \in (0,1), \quad \nu > 0. \tag{39}$$

In experiments, observations supplied to the parametric and MDP predictive calculations are clipped to $[\epsilon_{\mathrm{mach}}, 1 - \epsilon_{\mathrm{mach}}]$, where ϵ_{mach} is double-precision machine epsilon, preventing singular beta-likelihood evaluations; the resulting stakes are evaluated on the original observations in the test martingale. Posterior inference is approximated using the IBIS/waste-free SMC sampler from the `particles` package [67]. With N_{smc} particles and Markov chains of length L_{smc}, each time point stores

$$M = N_{\mathrm{smc}} \times L_{\mathrm{smc}}$$

posterior samples. Conditional on these samples, the predictive score

$$\mathbb{E}\left[\frac{X - \mu}{1 + \lambda(X - \mu)} \,\middle|\, X \sim \mathrm{Beta}(a, b) \right]$$

is evaluated for each sampled beta distribution using the closed-form hypergeometric expression, and the resulting scores are averaged over the SMC particle cloud. The root of the averaged score is then solved independently for each $\mu \in \mathcal{G}$.

MDP predictive. The MDP method implements the predictive distribution in (19). Let $\mathbb{H}_{n|n-1}(\cdot \mid X_{1:n-1})$ and $\mathbb{P}_{n-1}$ denote the beta posterior predictive and the empirical distribution after observing $X_{1:n-1}$, respectively. For $\kappa > 0$, the implemented predictive is

$$\mathbb{Q}_{n|n-1} = \frac{\kappa}{\kappa + n - 1} \mathbb{H}_{n|n-1} + \frac{n-1}{\kappa + n - 1} \mathbb{P}_{n-1}.$$

Accordingly, the first-order condition for $\widehat{\lambda}_t(\mu)$ is evaluated as the same convex combination of the parametric Bayesian score and the empirical score. As n increases, the empirical component dominates, so the predictive distribution becomes increasingly data-driven.

BETEL predictive. The BETEL method implements the unregularised empirical-likelihood predictive from Section 5.3, corresponding to $\tau_n = 0$. After observing $X_{1:n-1}$, the pseudo-posterior integral (23) is approximated on a deterministic grid of mean values $\mu_0 \in (0, 1)$. In the unregularised case, the implementation constructs this grid inside the interior of the empirical convex hull of $X_{1:n-1}$, corresponding to the convention that the empirical likelihood is zero outside this hull. For each retained μ_0, the tilting parameter γ_{n,μ_0} and the weights $w_{n,i}(\mu_0)$ are computed numerically as described in Section 5.3. The pseudo-posterior weights are then normalised over the retained grid points, using a log-sum-exp calculation for numerical stability.

The resulting predictive distribution is the grid approximation to

$$\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}) = \int_0^1 \widetilde{P}_{n,\mu_0}(\cdot)\, \Pi_{n-1}(d\mu_0),$$

and is represented in the code as a discrete distribution supported on the observed sample points $X_1, \ldots, X_{n-1}$. This discrete predictive is then used in (10) to compute $\widehat{\lambda}_n(\mu)$ for each candidate value μ in the confidence-sequence inversion grid. If the history is empty or degenerate, the implementation uses the safe default $\widehat{\lambda}_n(\mu) = 0$ for all candidate values μ.

RETEL predictive. The RETEL method implements the regularised empirical-likelihood predictive from Section 5.3. In all reported experiments, we use the two-point regulariser

$$\mathbb{H}_{\mu_0} = \mathbb{H} = \frac{1}{2}(\delta_0 + \delta_1), \qquad \tau_n = 1.$$

After observing $X_{1:n-1}$, the pseudo-posterior integral (23) is approximated on a deterministic grid of values $\mu_0 \in (0, 1)$. For each grid value of μ_0, the tilting parameter γ_{n,μ_0}, the sample-point weights, and the endpoint weights are computed using the formulas in Appendix C. The pseudo-posterior weights are then normalised on the grid, again using log-sum-exp normalisation.

The resulting predictive distribution is the grid approximation to

$$\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}) = \int_0^1 \widetilde{P}_{n,\mu_0}(\cdot)\, \Pi_{n-1}(d\mu_0).$$

In contrast to BETEL, this predictive has support on both the observed sample points and the two endpoints $\{0, 1\}$. It is then substituted into (10) to compute $\widehat{\lambda}_n(\mu)$ over the confidence-sequence inversion grid. The endpoint masses stabilise the empirical-likelihood approximation and make the regularised construction well defined over the full mean range.

D.3 Computational complexity

Let n denote the sequence length and let G be the number of equal subdivisions used for test inversion, giving $G - 1$ interior candidate means. All CS methods based on (5), like the ones above, evaluate wealth processes over the $n \times (G - 1)$ time–mean grid. Thus, after the predictable coefficients have been computed, test inversion itself contributes a common $\mathcal{O}(nG)$ time cost.

However, the dominant cost is usually the computation of the predictable coefficients itself. For the empirical predictive, exact evaluation of the first-order condition (38) requires summing over the current history $X_{1:t-1}$ at each time t and each grid value. This gives $\mathcal{O}(Gn^2)$ operations up to root-solver iterations. The MDP method inherits the same empirical-support term, and also averages the parametric predictive contribution over posterior particles. BETEL and RETEL also have growing empirical-support dependence: at each time they solve empirical-likelihood tilting equations and form pseudo-predictive weights using sums over $X_{1:t-1}$. With a fixed pseudo-posterior grid size G_{etel}, their exact implementation therefore contains terms of order $\mathcal{O}(G_{\text{etel}} n^2)$ for the tilting and pseudo-posterior construction, as well as $\mathcal{O}(Gn^2)$ for the subsequent weighted empirical betting optimisation, again up to solver iterations. Consequently, the empirical, MDP, BETEL, and RETEL methods have quadratic dependence on n and linear dependence on G in the current exact implementation.

By contrast, the parametric Bayesian SMC method avoids the growing empirical support in the coefficient computation. With a fixed particle budget, its coefficient-computation cost is linear in n and G, up to particle, root-solver, and SMC-update constants.

Table 2: Wall-clock time in seconds of one method call computing the CS at $n = 1000$ on the Beta mixture scenario, by CPU: mean (standard error) over replicates.

Method	Prior	Intel Xeon Platinum 8268	AMD EPYC 9634	AMD EPYC 9645
Empirical	-	18.63 (0.31)	7.78 (0.12)	5.90 (0.65)
MDP	Informative	2223.19 (25.77)	1445.81 (12.11)	1276.17 (74.21)
BETEL	Informative	189.96 (2.77)	108.80 (1.38)	106.04 (11.48)
RETEL	Informative	551.69 (7.30)	432.89 (5.96)	448.47 (40.94)
Parametric	Informative	1864.92 (20.55)	1336.11 (14.09)	1208.72 (82.34)
Replicates	-	265	694	41

The quadratic dependence on sequence length is determined by the predictive representation and its numerical implementation, rather than by the use of prior information itself. Nonetheless, wall-clock costs can still differ substantially among the Bayes-assisted methods proposed in this work because of the additional grids, posterior computations, and numerical optimizations. Table 2 reports method-only wall-clock times for computing the CS until $n = 1000$ in the beta-mixture experiment for the empirical predictive, parametric Bayes, and the proposed Bayes-assisted methods. Since these methods share the predictive-assisted CS construction, the comparison isolates the computational cost of their different predictive representations. Results are stratified by processor across all 1000 paired replicates.

Thus, in their current implementations, the Bayes-assisted methods are substantially slower than the empirical predictive. This additional computation may nevertheless be acceptable when the cost of acquiring observations dominates the cost of computing the CS.

D.4 Default numerical parameters

Unless otherwise specified, all experiments use confidence level $\alpha = 0.1$ and $G = 500$ equal grid subdivisions, giving spacing $1/500$ and 499 interior candidate means for test inversion. The empirical and Bayes-assisted methods use $c = 0.95$, HC uses its default $c = 0.5$, and PRECiSE-CO96 and UP use their default full-range constructions, corresponding to $c = 1$. For the SMC-based beta predictives we use $N_{\text{smc}} = 20$ and $L_{\text{smc}} = 50$, giving $M = 1000$ posterior samples at each time point. For BETEL and RETEL we use $G_{\text{etel}} = 1000$ grid points for the pseudo-posterior integral. For the MDP predictive, we use $\kappa = 50$; RETEL uses the two-point regulariser with $\tau_n = 1$, while BETEL corresponds to $\tau_n = 0$.

Moreover, the Bayes-assisted methods, including the parametric Bayesian, MDP, BETEL, and RETEL predictives, require prior specifications. In particular, the parametric Bayesian and MDP methods require a prior distribution over the beta working model mean–concentration parameters ρ and ν in Equation (39). This usually takes the form

$$\rho \sim \mathrm{Beta}(a_\rho, b_\rho), \qquad \nu \sim \mathrm{Gamma}(a_\nu, b_\nu)$$

for some hyperparameters $a_\rho, b_\rho, a_\nu, b_\nu$. The BETEL and RETEL methods require a prior distribution directly over the mean μ, which we take to be a beta distribution of the form

$$\mu \sim \mathrm{Beta}(a_m, b_m).$$

In practice, we usually set $a_m = a_\rho$ and $b_m = b_\rho$.

E Continuum-valid cellwise inversion

The main experiments invert test martingales only at finitely many candidate means, so their validity guarantee applies only to those grid points. Taking the hull of the accepted points and enlarging it by one grid step is a useful reporting convention, but does not by itself certify coverage between grid points. Here we instead partition the parameter space into cells, use stakes that are valid uniformly within each cell, and reject a cell only when endpoint evaluations certify rejection throughout it. The union of the retained cells is therefore a confidence sequence over all $\mu \in [0, 1]$.

E.1 Cellwise Betting Construction

Fix a partition of $[0,1]$: $0 = a_0 < a_1 < \cdots < a_G = 1$, and define cells

$$B_j = [a_{j-1}, a_j], \qquad j = 1, \ldots, G.$$

Write $B_j = [\ell_j, u_j]$. For each cell, define the cellwise-safe stake interval

$$J_j = \bigcap_{\mu \in B_j \cap (0,1)} I_{\mu,c} = \left[-\frac{c}{1-\ell_j}, \frac{c}{u_j} \right].$$

Thus every $\lambda \in J_j$ is valid simultaneously for every $\mu \in B_j \cap (0,1)$. At time t, before observing X_t, choose a predictable stake $\lambda_{t,j} \in J_j$. This choice can be arbitrary from the perspective of validity. For efficiency, one natural choice is to optimize predictive log-growth at the cell midpoint $m_j = \frac{\ell_j + u_j}{2}$. Given a predictable distribution $Q_{t|t-1}$, set

$$\lambda_{t,j} \in \arg\max_{\lambda \in J_j} \mathbb{E}_{X \sim Q_{t|t-1}} \log\{1 + \lambda(X - m_j)\}.$$

This is a valid predictable stake for all $\mu \in B_j$, hence validity is preserved.

E.2 Endpoint Certificate

For each cell B_j, define the cellwise log-wealth

$$\Lambda_{n,j}(\mu) = \sum_{t=1}^{n} \log\{1 + \lambda_{t,j}(X_t - \mu)\}, \qquad \mu \in B_j.$$

For fixed $\lambda_{t,j}$, each summand is concave in μ. Therefore $\Lambda_{n,j}$ is concave on B_j. A concave function on an interval attains its minimum at an endpoint, so

$$\inf_{\mu \in B_j} \Lambda_{n,j}(\mu) = \min\{\Lambda_{n,j}(\ell_j), \Lambda_{n,j}(u_j)\}.$$

We reject the entire cell B_j at time n only if

$$\min\{\Lambda_{n,j}(\ell_j), \Lambda_{n,j}(u_j)\} > \log(1/\alpha).$$

Equivalently, the reported confidence set is

$$\widehat{C}_{\alpha,n} = \bigcup_{j:\, \min\{\Lambda_{n,j}(\ell_j), \Lambda_{n,j}(u_j)\} \le \log(1/\alpha)} B_j.$$

If a single interval is desired, one may report the convex hull of $\widehat{C}_{\alpha,n}$, which remains conservative.

An empirical comparison with pointwise grid inversion is reported in Appendix G.1.6.

F Experimental details

This section provides additional details on the experiments presented in Section 6 and Appendix G.

F.1 Synthetic experiments

F.1.1 Prior specification

For each of the six distributions considered for the synthetic experiments, we consider three prior regimes: informative, non-informative, and misspecified. Informative priors place significant mass around the true mean and represent the case in which the prior information is directionally correct. Non-informative priors are diffuse over $[0,1]$. Misspecified priors are concentrated away from the true mean and are used to test sensitivity to misleading prior information. The specific prior specifications for each distribution and regime are as follows.

Bernoulli distributions. Let $X_n \sim \text{Bernoulli}(p)$, $p \in \{0.1, 0.5\}$. For Parametric/MDP, we set:

- Informative prior: $\rho \sim \text{Beta}(500p, 500(1-p))$, $\quad \nu \sim \text{Gamma}(1, 100)$.
- Non-informative prior: $\rho \sim \text{Beta}(1, 1)$, $\quad \nu \sim \text{Gamma}(1.5, 1)$.
- Misspecified prior: $\rho \sim \text{Beta}(500p_{\text{bad}}, 500(1-p_{\text{bad}}))$, $\quad \nu \sim \text{Gamma}(7.5, 1.0)$, where $p_{\text{bad}} = 0.5$ if $p = 0.1$ and $p_{\text{bad}} = 0.1$ if $p = 0.5$.

For BETEL/RETEL, we set the μ priors to be equal to the corresponding ρ priors above.

Beta distributions. Let $X_n \sim \text{Beta}(a, b)$. Consider the three cases $(a, b) = (0.5, 0.5), (1, 1), (10, 30)$, which give $\mu^\star = a/(a+b) = 0.5, 0.5, 0.25$. For Parametric/MDP, we set:

- Informative prior: $\rho \sim \text{Beta}(500\mu^\star, 500(1-\mu^\star))$, $\quad \nu \sim \text{Gamma}(a_\nu, b_\nu)$, where $(a_\nu, b_\nu) = (2, 1), (7.5, 1), (2, 0.1)$ in the three cases.
- Non-informative prior: $\rho \sim \text{Beta}(1, 1)$, $\quad \nu \sim \text{Gamma}(1.5, 1)$.
- Misspecified prior: $\rho \sim \text{Beta}(500\mu_{\text{bad}}, 500(1-\mu_{\text{bad}}))$, $\quad \nu \sim \text{Gamma}(a_\nu, b_\nu)$, where $\mu_{\text{bad}} = 0.9$ and $(a_\nu, b_\nu) = (2, 1), (7.5, 1), (2, 0.1)$ in the three cases.

For BETEL/RETEL, we set the μ priors to be equal to the corresponding ρ priors above.

Beta mixture distribution. Let X_n be distributed as $0.25\,\text{Beta}(5, 15) + 0.75\,\text{Beta}(15, 5)$, which gives $\mu^\star = 0.625$. For Parametric/MDP, we set:

- Informative prior: $\rho \sim \text{Beta}(500\mu^\star, 500(1-\mu^\star))$, $\quad \nu \sim \text{Gamma}(2, 2)$.
- Non-informative prior: $\rho \sim \text{Beta}(1, 1)$, $\quad \nu \sim \text{Gamma}(1.5, 1)$.
- Misspecified prior: $\rho \sim \text{Beta}(500\mu_{\text{bad}}, 500(1-\mu_{\text{bad}}))$, $\quad \nu \sim \text{Gamma}(2, 2)$, where $\mu_{\text{bad}} = 0.1$.

For BETEL/RETEL, we set the μ priors to be equal to the corresponding ρ priors above.

F.2 Best-arm identification for LLM evaluation

Appendix F.2.1 provides a detailed description of the LUCB algorithm [58, 59], whereas Appendix F.2.2 details the implementation of the LUCB algorithm for best-arm identification in LLM evaluation, and explains our experimental setup.

F.2.1 Lower Upper Confidence Bounds algorithm

Imagine you have access to n independent arms (or bandits); when played, these arms provide a reward sampled from an unknown distribution that characterizes them. The Best Arm Identification problem aims to identify the arm with the highest expected reward with high probability, using the minimum number of plays possible. This problem, known as the multi-armed bandit problem, is a classical challenge in decision theory.

The Lower Upper Confidence Bounds (LUCB) algorithm is a method designed to perform best arm identification in multi-armed bandit problems. It was introduced by [58] and refined by [59]. Unlike standard bandit algorithms which aim to maximize the cumulative reward while playing, LUCB is a purely exploratory algorithm. Its goal is strictly to identify the top-m arms with high confidence using as few samples as possible. In our experiment, we focus only on best arm identification, but we present in this appendix the general algorithm that can perform best-m-arms identification.

The central idea is that one does not need to know the precise mean of every arm, but only to identify the boundary between the top-m arms and the remaining arms. To explain the algorithm, let us assume we have n arms and want to identify the subset of size m with the highest means. Consider a confidence level $\alpha \in (0, 1)$. At any time step t and for each arm a, we maintain:

- $\hat{\mu}_a(t)$, the empirical mean.
- $[L_a(t), U_a(t)]$ a $(1-\alpha)$-two-sided confidence sequence for the true mean of a.

Note that these values are not updated for every arm at every time step, but only for the arms that have been pulled. LUCB is agnostic to the methods used to derive the upper and lower confidence bounds, provided they offer proper statistical guarantees. While the choice of bound does not alter the correctness of the algorithm, it impacts its performance in terms of the number of sample pulls required for identification.

At each time step t, the algorithm sorts all arms based on their empirical means to form a tentative set of top-m arms, denoted J_t. From this ranking, we can identify two critical arms that define the decision boundary:

1. The Contender $h_t = \arg\min_{a \in J_t} L_a(t)$: this is the arm in J_t that we are least likely to keep among the top-m.
2. The Challenger $l_t = \arg\max_{a \notin J_t} U_a(t)$: this is the arm in J_t^c that we are most likely to include among the top-m.

We then draw samples only from the Contender and the Challenger to update their values. By doing so, we ignore the arms that are clearly safe winners or safe losers. The algorithm stops when:

$$L_{h_t}(t) > U_{l_t}(t) - \epsilon$$

This occurs when the confidence sets of these two critical arms no longer overlap up to a tolerance ϵ (which can be chosen equal to 0). The full algorithm is described in Algorithm 1. When the algorithm stops, we are $(1 - n\alpha)$-confident that every single arm in the returned set is better than, or at most ϵ-worse than, any arm rejected.

Algorithm 1 LUCB: Lower Upper Confidence Bound

Input: Set of arms $\mathcal{K} = \{1, \ldots, n\}$, target number of arms m, confidence parameter $\alpha \in (0, 1)$, tolerance $\epsilon \geq 0$.
Initialization: Sample each arm $a \in \mathcal{K}$ once. Set time $t \leftarrow n$.
Calculate empirical means $\hat{\mu}_a(t)$ and confidence bounds $U_a(t), L_a(t)$.
while $U_{l_t}(t) - L_{h_t}(t) \geq \epsilon$ **do**
  Draw one sample from arm h_t and one sample from arm l_t.
  Update empirical means $\hat{\mu}$ and confidence bounds for h_t and l_t.
  Sort arms based on empirical means: $\hat{\mu}_{(1)}(t) \geq \hat{\mu}_{(2)}(t) \geq \cdots \geq \hat{\mu}_{(n)}(t)$.
  Define the set of current Top-m candidates: J_t.
  Identify the *contender* $h_t = \arg\min_{a \in J_t} L_a(t)$.
  Identify the *challenger* $l_t = \arg\max_{a \notin J_t} U_a(t)$.
  $t \leftarrow t + 2$.
end while
return J_t

F.2.2 Application of LUCB to LLM evaluation

For our experiments we apply the LUCB algorithm using several underlying CS methods. While LUCB is designed to pull arms sequentially, we ensure each LUCB variant receives the same LLM outputs by first performing a large number of LLM calls and then simulating the online behavior for each LUCB variant. Note that in practice, when the goal is strictly best-arm identification rather than method comparison, the LLM calls would be performed iteratively. The four LLMs we compare are:

- **Gemma 3 1B:** `google/gemma-3-1b-it-qat` (`https://huggingface.co/google/gemma-3-1b-it-qat-q4_0-gguf` *license gemma*)
- **Qwen 3 1.7B:** `qwen3-1.7b`: (`https://huggingface.co/Qwen/Qwen3-1.7B-GGUF` *license apache 2.0*)
- **Ministral 3 3B:** `ministral-3-3b-instruct-2512`: (`https://huggingface.co/mistralai/Ministral-3-3B-Instruct-2512-GGUF` *license apache 2.0*)
- **Llama 3.2 3B:** `llama-3.2-3b-instruct`: (`https://huggingface.co/lmstudio-community/Llama-3.2-3B-Instruct-GGUF` *license llama 3.2*)

When available, reasoning abilities were disabled. Each time a LLM is queried, we generate a random array of integers between 0 and 100 without replacement, and the LLM is tasked with sorting the array using the following prompt:

```
You will receive 20 indexed items in the form:
i0: value i1: value ...
Task: Assign one numeric sorting score to each item.
Rules: - Return exactly one numeric field for each item id -
Smaller score means smaller value - Larger score means larger
value - Use all item ids exactly once as field names - Do not
return any extra fields
Example: Input: i0: 42 i1: 7 i2: 15
A correct response would be something like: {{ "i0": 3,
"i1": 1, "i2": 2 }}
because 7 < 15 < 42.
```

The LLMs were called using LM Studio version 0.4.25+1 locally on an Apple Silicon M4 Max Chip with 36 GB of RAM. The temperature of the LLMs was set to 0.0. We utilise a structured output format to ensure the LLMs return only the array ranking. The performance of the LLM ordering is evaluated against the ground truth using the Spearman rank correlation coefficient, defined as

$$\mathcal{S} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{40}$$

where d_i is the difference between the predicted and the target rank of the i-nth number in the array, which is returned as reward after an arm pull and rescaled to be in $[0, 1]$ before being fed to the LUCB algorithm.

In all our experiments, LUCB is used with $\alpha = 0.025$ and $\epsilon = 0.1$.

Lastly, we emphasise that the goal of this experiment is to compare confidence-sequence methods within a best-arm identification procedure, rather than to benchmark the LLMs themselves. Accordingly, our primary outcome is the number of model queries required for LUCB to identify the best arm. In all replicates, and for all confidence-sequence methods and score metrics considered, Ministral 3 3B is selected as the best arm; the differences between methods therefore reflect sampling efficiency rather than disagreement about the final selected model.

F.2.3 Prior specification

For the Bayes-assisted CS methods, we elicit a prior for every LLM tested. In particular, given an LLM, we evaluate its performance on the same sorting task, using the same prompt, for smaller arrays ($n \in \{16, 18, 19\}$), for each smaller array we perform 20 sorting tasks and fit an exponential decay function to extrapolate the predicted performance at $n = 20$, denoted by $\hat{\mu}$. We also build an estimator $\hat{\nu} = \hat{\mu}(1-\hat{\mu})/\hat{\sigma}$ where $\hat{\sigma}$ is the empirical variance of the 20 data points at size 19. With this, we then set the following priors for the Bayes-assisted CS methods:

- Parametric/MDP: $\rho \sim \mathrm{Beta}(500\hat{\mu}, 500(1-\hat{\mu})), \quad \nu \sim \mathrm{Gamma}(2, 2/\hat{\nu})$.
- BETEL/RETEL: $\mu \sim \mathrm{Beta}(500\hat{\mu}, 500(1-\hat{\mu}))$.

The prior on ν is chosen so its mean is $\hat{\nu}$ while maintaining the same relative uncertainty as the $\mathrm{Gamma}(2, 2)$ used in the synthetic experiment.

F.3 Prediction-powered inference

F.3.1 Bounded-mean PPI construction

Prediction-powered inference (PPI) uses predictions from a fixed model to improve inference from a limited number of labelled observations and a larger pool of unlabelled observations [60]. We use PPI only in the bounded-mean setting considered throughout the paper. Let (X_i, Y_i) denote a labelled example, let f be a fixed prediction rule, and let the target be

$$\theta^\star = \mathbb{E}[Y].$$

PPI decomposes this target as

$$\theta^\star = \mathbb{E}[f(X)] + \mathbb{E}[Y - f(X)] = m^\star + \Delta^\star,$$

where $m^\star$ is the prediction-only term and $\Delta^\star$ is the rectifier correcting the bias of the prediction rule. Recent work has studied confidence sequences for this decomposition, either by constructing prediction-powered confidence sequences with asymptotic guarantees [50] or by building prediction-powered e-values in more general sequential settings [68]. Our use of PPI is narrower but fits directly into the nonasymptotic bounded-mean framework of the present paper: after fixing the unlabelled plug-in estimate of $m^\star$, the remaining inferential problem is an ordinary bounded-mean problem for the rectifier.

Specifically, given a pool of N unlabelled examples, we estimate the prediction-only term by

$$\widehat{m}_N = \frac{1}{N}\sum_{j=1}^{N} f(X_j).$$

Following the simplified setting of Kilian et al. [50], we treat $\widehat{m}_N$ as fixed and focus the sequential uncertainty on the labelled residuals

$$R_i = Y_i - f(X_i), \qquad \Delta^\star = \mathbb{E}[R_i].$$

If $R_i \in [\ell, u]$, we rescale the residuals to the unit interval by

$$Z_i = \frac{R_i - \ell}{u - \ell} \in [0,1], \qquad \mu_Z^\star = \mathbb{E}[Z_i] = \frac{\Delta^\star - \ell}{u - \ell}.$$

We then apply the bounded-mean confidence-sequence methods of the paper to the sequence $(Z_i)_{i\geq 1}$. If $C_{\alpha,n}^Z$ denotes the resulting confidence sequence for $\mu_Z^\star$, then the corresponding confidence sequence for the rectifier is

$$C_{\alpha,n}^\Delta = \ell + (u - \ell)C_{\alpha,n}^Z,$$

and the reported PPI confidence sequence for the original target is

$$C_{\alpha,n}^\theta = \widehat{m}_N + C_{\alpha,n}^\Delta.$$

Thus all Bayes-assisted and prior-free bounded-mean methods are applied only to the rescaled residual sequence Z_i; the unlabelled predictions enter through the fixed shift $\widehat{m}_N$.

F.3.2 Datasets

Galaxies. The GALAXIES dataset contains a total of 16743 images from the Galaxy Zoo 2 initiative [69]. For each image, the dataset contains a binary label indicating whether the galaxy has spiral arms ($Y_i \in \{0,1\}$) and the predicted probability of a ResNet50 model [70] for the same quantity ($f(X_i) \in [0,1]$). The estimand is the population fraction of galaxies with spiral arms, i.e., $\theta^\star = \mathbb{E}[Y_i] \in [0,1]$. The PPI rectifier is based on residuals $R_i = Y_i - f(X_i)$, which lie in $[-1,1]$. We therefore apply our bounded-mean methods to the rescaled residuals $Z_i = (R_i + 1)/2 \in [0,1]$.

AttributedQA. The ATTRIBUTEDQA dataset evaluates attribution in question answering systems [71]. For each system output, the dataset reports an automatic attribution score from AutoAIS, an NLI-based evaluator, and for a subset of examples also reports human attribution labels. We compare the retrieve-then-read systems RTR-10 and RTR-4, for which there are 1000 examples with both human labels and AutoAIS scores, and 2610 additional examples with AutoAIS scores only. For each human-labelled example and system $k \in \{10, 4\}$, let $Y_i^{(k)} \in \{0,1\}$ indicate whether the answer is judged by humans to be correctly attributed, and let $f^{(k)}(X_i) \in [0,1]$ denote the corresponding AutoAIS score. We form paired differences

$$Y_i = Y_i^{(10)} - Y_i^{(4)} \in \{-1, 0, 1\}, \qquad f(X_i) = f^{(10)}(X_i) - f^{(4)}(X_i) \in [-1,1].$$

The estimand is the difference in human attribution reliability between RTR-10 and RTR-4, i.e., $\theta^\star = \mathbb{E}[Y_i] \in [-1,1]$. The PPI rectifier is based on residuals $R_i = Y_i - f(X_i)$, which lie in $[-2,2]$. We therefore apply our bounded-mean methods to the rescaled residuals $Z_i = (R_i + 2)/4 \in [0,1]$.

F.3.3 Prior specification

We choose the PPI priors to match the heuristic of Kilian et al. [50], who place a zero-centred prior on the scaled mean rectifier with variance $1/n_0$, where n_0 is a reference labelled sample size. Let $R_i = Y_i - f(X_i) \in [\ell, u]$ denote the rectifier residual, and let

$$Z_i = \frac{R_i - \ell}{u - \ell} \in [0, 1], \qquad \mu_Z^\star = \mathbb{E}[Z_i].$$

In both PPI datasets the residual bounds are symmetric around zero, so the prior belief $\mathbb{E}[R_i] \approx 0$ maps to $\mu_Z^\star \approx 1/2$. We therefore use a symmetric beta prior,

$$\mu_Z^\star \sim \mathrm{Beta}(\xi, \xi),$$

with ξ chosen by matching the variance on the original rectifier scale. Since

$$\mathrm{Var}\left(\ell + (u - \ell)\mu_Z^\star\right) = (u - \ell)^2 \frac{1}{4(2\xi + 1)},$$

matching this variance to $1/n_0$ gives

$$\xi(n_0) = \frac{1}{2}\left(\frac{n_0(u - \ell)^2}{4} - 1\right),$$

where n_0 is taken to be the largest labelled sample size considered in the PPI experiments, i.e., $n_0 = 1000$ for GALAXIES and $n_0 = 500$ for ATTRIBUTEDQA.

This prior enters the different Bayes-assisted predictives following Appendix D.4. For the parametric Bayesian and MDP methods, we use

$$\rho \sim \mathrm{Beta}(\xi(n_0), \xi(n_0)), \qquad \nu \sim \mathrm{Gamma}(7.5, 1),$$

where ρ and ν are the mean and concentration parameters of the beta working likelihood, respectively. For BETEL and RETEL, we place the corresponding prior directly on the rescaled mean,

$$\mu \sim \mathrm{Beta}(\xi(n_0), \xi(n_0)).$$

Thus all Bayes-assisted methods encode the same prior belief that the rectifier is close to zero, while differing in whether this belief enters through a beta working likelihood or directly through the empirical-likelihood pseudo-posterior.

G Additional experimental results

This section presents additional experimental results to complement the ones in Section 6.

G.1 Synthetic data

G.1.1 Full baseline comparison

Figure 3 shows the full-baseline version of the beta-mixture experiments in Figure 1, while Figures 4 and 5 show the full results on the Bernoulli and beta experiments, respectively. Overall, the results are consistent with the main text: accurate predictive distributions can substantially reduce confidence-sequence width without sacrificing coverage. It is interesting to see that the parametric Bayesian predictive severely underperforms the other Bayes-assisted methods in the Bernoulli experiment with $p = 0.1$, where the true distribution is far from the beta working model.

G.1.2 Prior sensitivity

Figures 7 to 8 compare the three prior regimes for the Bayes-assisted methods on the beta-mixture, Bernoulli, and beta simulations, respectively; detailed prior specifications are provided in Appendix F.1.1. The overall message is clear: the quality of the prior information incorporated into the predictive distribution can have a significant impact on efficiency, with misspecified priors leading to wider confidence sequences, but coverage is preserved across all prior regimes (up to random fluctuations in the $\mathrm{Bernoulli}(0.5)$ example, where one of the misspecified Bayes-assisted methods slightly undercovers).

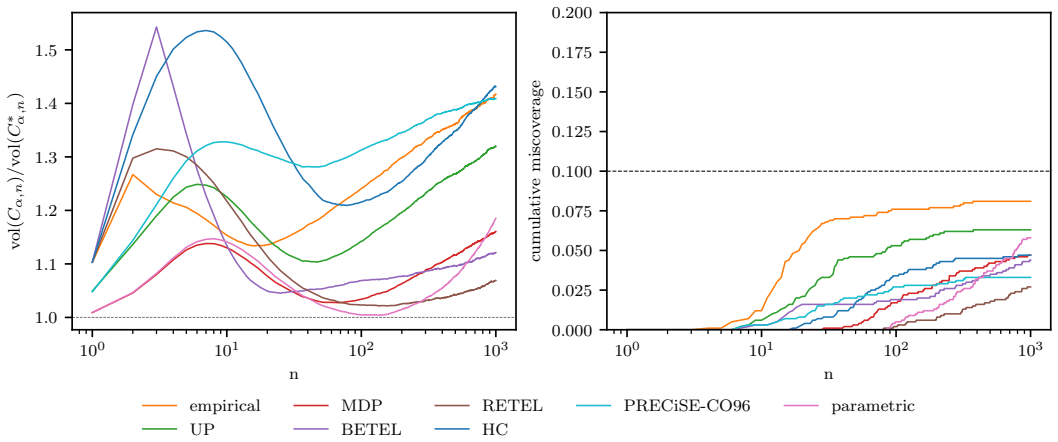

Figure 3: Full-baseline version of the synthetic beta-mixture experiment in Figure 1. The left and right panels show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 replicates; the dashed line marks $\alpha = 0.1$. Width ratios use replicates where both the method and oracle CSs are non-empty. At $n = 1000$, at most 2% are empty (0.5% on average).

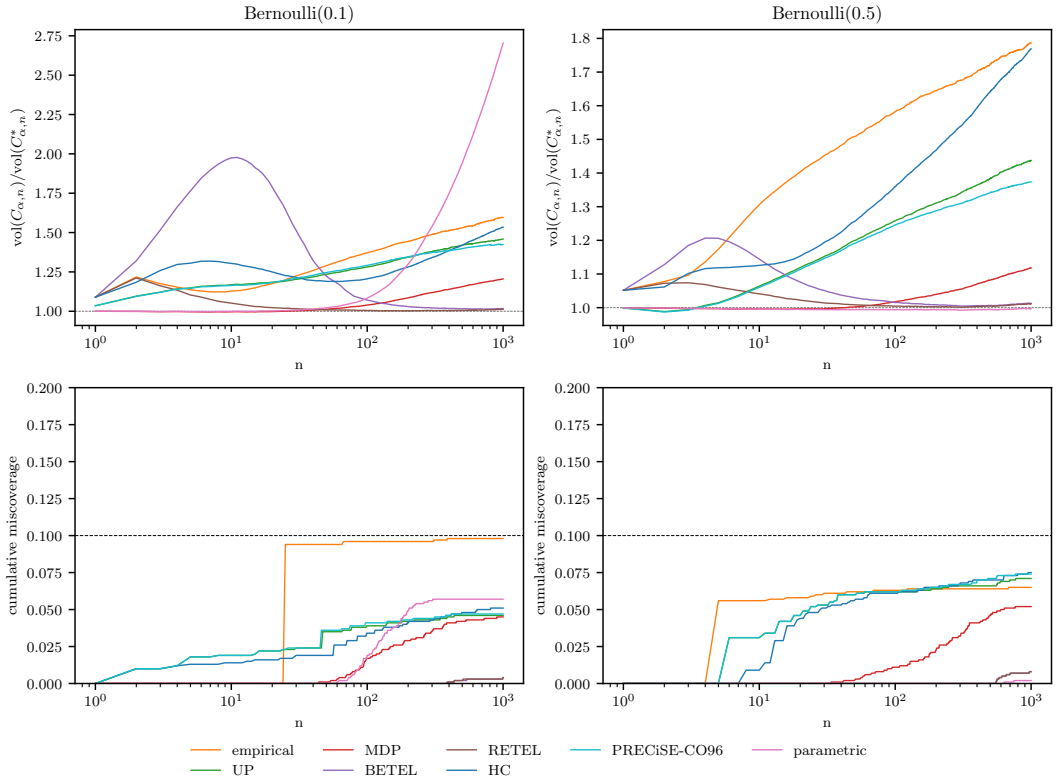

Figure 4: Synthetic Bernoulli experiments with all baselines. The left and right columns correspond to $X_n \sim \text{Bernoulli}(0.1)$ and $X_n \sim \text{Bernoulli}(0.5)$, respectively. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 replicates; the dashed line marks $\alpha = 0.1$. Width ratios use replicates where both the method and oracle CSs are non-empty. At $n = 1000$, at most 4.4% are empty (0.7% on average).

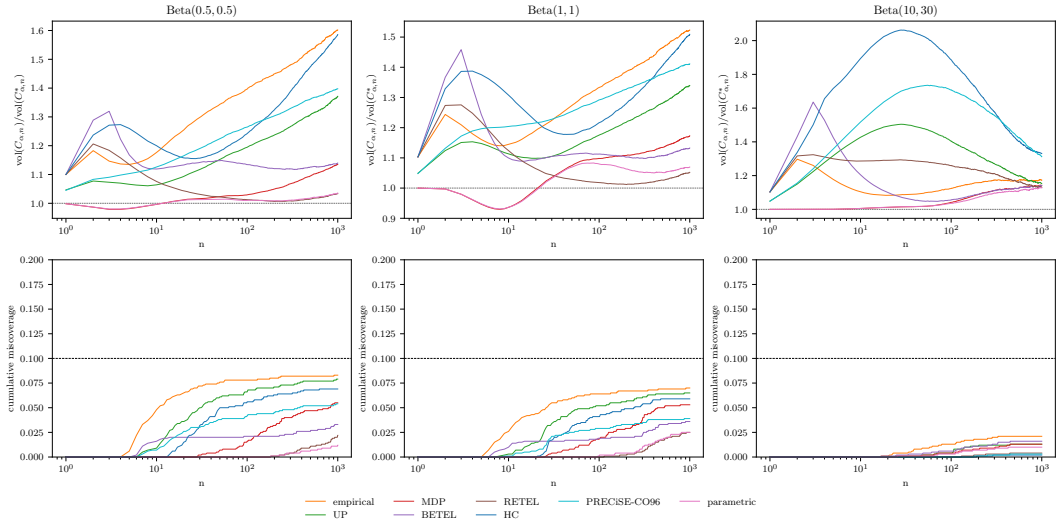

Figure 5: Synthetic beta experiments with all baselines. The left, centre, and right columns correspond to $X_n \sim \text{Beta}(0.5, 0.5)$, $X_n \sim \text{Beta}(1, 1)$, and $X_n \sim \text{Beta}(10, 30)$, respectively. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 replicates; the dashed line marks $\alpha = 0.1$. Width ratios use replicates where both the method and oracle CSs are non-empty. At $n = 1000$, at most 3.5% are empty (0.4% on average).

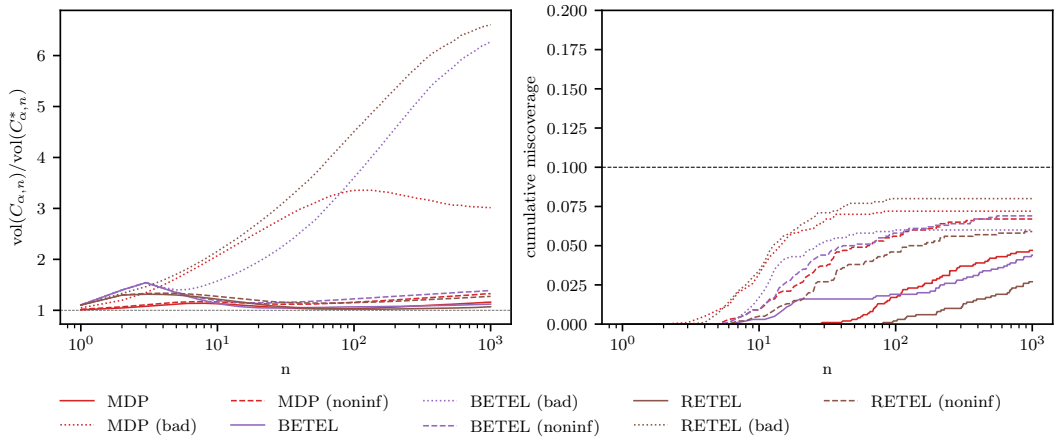

Figure 6: Synthetic beta-mixture experiment with different prior specifications for the same setting as Figures 1 and 3. The left and right panels show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 replicates. Solid, dotted, and faint solid curves correspond to informative, misspecified, and non-informative priors, respectively; the dashed horizontal line marks $\alpha = 0.1$. Width ratios use replicates where both the method and oracle CSs are non-empty. At $n = 1000$, at most 1.4% are empty (0.6% on average).

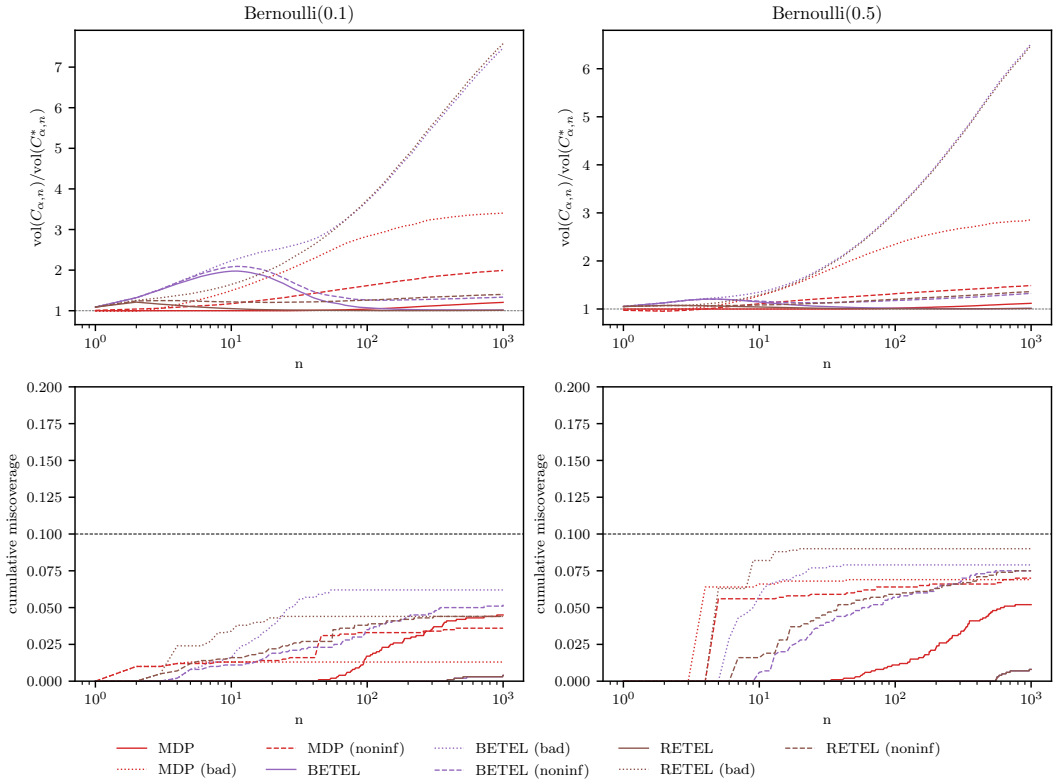

Figure 7: Synthetic Bernoulli experiment with different prior specifications for the same setting as in Figure 4. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 replicates. Solid, dotted, and faint solid curves correspond to informative, misspecified, and non-informative priors, respectively; the dashed horizontal line marks $\alpha = 0.1$. Width ratios use replicates where both the method and oracle CSs are non-empty. At $n = 1000$, at most 3.2% are empty (0.8% on average).

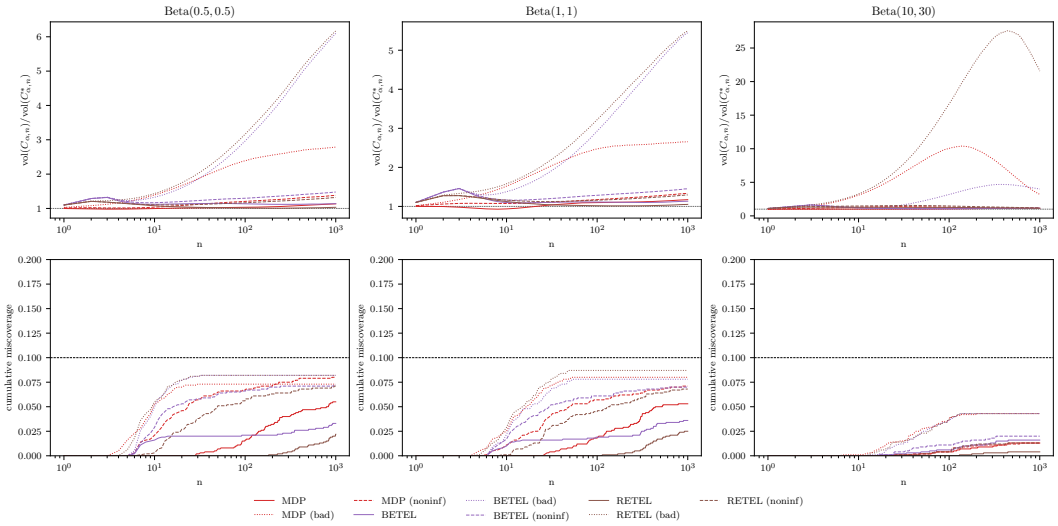

Figure 8: Synthetic beta experiment with different prior specifications for the same setting as in Figure 5. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 replicates. Solid, dotted, and faint solid curves correspond to informative, misspecified, and non-informative priors, respectively; the dashed horizontal line marks $\alpha = 0.1$. Width ratios use replicates where both the method and oracle CSs are non-empty. At $n = 1000$, at most 2.4% are empty (0.5% on average).

G.1.3 Prior-informed universal portfolio

We consider a prior-informed variant of the original Cover-Ordentlich Universal Portfolio (UP) [45] construction, which we call Informed Universal Portfolio (I-UP). For each candidate null mean μ, I-UP mixes the constant betting strategies

$$\lambda_{\mu,u} = \frac{u-\mu}{\mu(1-\mu)}, \quad u \in [0,1],$$

using the prior $\pi(du)$, giving the wealth process

$$W_n^{\pi}(\mu) = \int \prod_{i=1}^{n} \{1 + \lambda_{\mu,u}\,(X_i - \mu)\}\,\pi(du),$$

which is a valid test martingale for any fixed choice of π. The form of $\lambda_{\mu,u}$ arises from

$$\lambda_{\mu,u} \in \operatorname{argmax}_{\lambda \in I_{\mu,1}} \inf_{\mathbb{P}:\mathbb{E}_{\mathbb{P}}X=u} \mathbb{E}_{\mathbb{P}} \log\{1+\lambda(X-\mu)\} = \operatorname{argmax}_{\lambda \in I_{\mu,1}} \mathbb{E}_{\mathbb{P}=\mathrm{Bernoulli}(u)} \log\{1+\lambda(X-\mu)\},$$

and therefore represents a robust Kelly bet against the composite null $H_\mu : \mu^\star = \mu$ against the composite alternative $H_u : \mu^\star = u$, where robustness refers to maximising the worst-case expected log growth over all distributions on $[0,1]$ with mean u. It can be thought of as a worst-case analogue of the predictive objective in Equation (10). Standard UP is recovered by choosing $\pi(du) = \mathrm{Beta}(1/2, 1/2)$. By setting $\pi(du)$ equal to the same Beta prior on the mean used for BETEL/RETEL, one can incorporate the same prior information through a structurally different mechanism: rather than using the prior-induced predictive to select the coefficient that maximises predictive expected log-growth, I-UP uses the prior as a fixed mixing distribution over constant, alternative-mean-indexed betting strategies.

To evaluate the performance of I-UP, we run additional experiments for all our synthetic scenarios. Table 3 compares average widths over 100 replicates between the different methods; the Bayes-assisted methods are compared under the same priors.

The Bernoulli setting provides a positive control: because the least-favourable bounded distribution with mean u is $\mathrm{Bernoulli}(u)$, the robust bet indexed by u coincides with the oracle's. Accordingly, informative I-UP essentially matches the oracle and is roughly 15–20% narrower than standard UP at $n = 50$ and 25% narrower at $n = 200$. Informative MDP is similarly close to the oracle. Thus, I-UP can exploit prior information nearly optimally when its robust mean-indexed bets align with the data-generating distribution.

For the non-Bernoulli distributions, informative MDP remains close to the oracle and is narrower than standard UP. In contrast, informative I-UP is generally wider than standard UP; the exception is $\mathrm{Beta}(0.5, 0.5)$, where the two are similar and I-UP is slightly narrower at $n = 200$. While the prior is correctly centred on the true mean, I-UP concentrates on bets that are worst-case optimal over all bounded distributions with a given mean, whereas the coefficient optimal for the actual distribution can correspond to an index far from that mean. MDP can exploit this additional distributional information through its adaptive predictive formulation.

Finally, non-informative variants remain broadly comparable to their prior-free counterparts, while strongly misspecified priors substantially reduce the finite-sample efficiency of both approaches without affecting validity. Under misspecification, the efficiency loss is consistently smaller for MDP, particularly at $n = 200$, because its predictive increasingly relies on the empirical distribution.

Overall, the comparison shows that the gains of our framework do not arise merely from access to prior information, but from translating that information into candidate-specific bets that adapt well to the data-generating process at hand.

G.1.4 Truncation parameter c

The truncation parameter c in Equation (3) is common in betting-based confidence-sequence methods. This section explains its role, the choice $c = 0.95$ for the empirical and Bayes-assisted experiments, and the sensitivity of the resulting confidence sequences to this choice.

The truncation parameter c does not affect validity, although it can substantially affect efficiency. For a given $c \in (0,1)$ and candidate mean μ, the set of admissible betting coefficients

$$I_{\mu,c} = \left[-\frac{c}{1-\mu}, \frac{c}{\mu}\right]$$

Table 3: Average CS width (standard error) over 100 replicates, at $n = 50$ and $n = 200$. Empty CSs are counted as width zero; across all CSs shown, at most 3% are empty at $n = 50$ or $n = 200$ (0.1% on average).

		Bernoulli(0.1)		Bernoulli(0.5)		Beta(0.5, 0.5)		Beta(1, 1)		Beta(10, 30)		Beta mixture	
		$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$
Method	Prior												
Oracle	-	0.170 (0.0030)	0.086 (0.0011)	0.284 (0.0035)	0.145 (0.0019)	0.205 (0.0021)	0.104 (0.0017)	0.168 (0.0017)	0.083 (0.0011)	0.059 (0.0003)	0.022 (0.0002)	0.137 (0.0016)	0.069 (0.0008)
UP	$\mathrm{Beta}(1/2, 1/2)$	0.212 (0.0044)	0.117 (0.0022)	0.340 (0.0051)	0.189 (0.0039)	0.237 (0.0034)	0.130 (0.0025)	0.194 (0.0027)	0.102 (0.0018)	0.087 (0.0003)	0.030 (0.0002)	0.151 (0.0022)	0.082 (0.0015)
I-UP	Informative	0.170 (0.0030)	0.086 (0.0011)	0.284 (0.0033)	0.145 (0.0016)	0.241 (0.0011)	0.120 (0.0007)	0.228 (0.0007)	0.112 (0.0004)	0.184 (0.0001)	0.088 (0.0001)	0.216 (0.0007)	0.106 (0.0003)
	Non-informative	0.219 (0.0036)	0.119 (0.0021)	0.325 (0.0042)	0.179 (0.0034)	0.228 (0.0029)	0.124 (0.0022)	0.189 (0.0023)	0.098 (0.0016)	0.100 (0.0002)	0.033 (0.0002)	0.154 (0.0019)	0.080 (0.0012)
	Misspecified	0.490 (0.0035)	0.424 (0.0035)	0.646 (0.0147)	0.575 (0.0128)	0.516 (0.0074)	0.440 (0.0055)	0.478 (0.0048)	0.404 (0.0037)	0.648 (0.0009)	0.562 (0.0005)	0.564 (0.0039)	0.484 (0.0029)
MDP	Informative	0.173 (0.0032)	0.094 (0.0014)	0.286 (0.0028)	0.151 (0.0016)	0.210 (0.0020)	0.110 (0.0014)	0.184 (0.0026)	0.094 (0.0013)	0.060 (0.0003)	0.024 (0.0002)	0.141 (0.0015)	0.074 (0.0009)
	Non-informative	0.252 (0.0065)	0.152 (0.0033)	0.352 (0.0071)	0.193 (0.0045)	0.239 (0.0038)	0.131 (0.0027)	0.195 (0.0027)	0.103 (0.0019)	0.080 (0.0004)	0.028 (0.0003)	0.153 (0.0026)	0.082 (0.0017)
	Misspecified	0.418 (0.0043)	0.270 (0.0039)	0.575 (0.0147)	0.367 (0.0099)	0.436 (0.0074)	0.269 (0.0050)	0.381 (0.0048)	0.222 (0.0035)	0.478 (0.0008)	0.223 (0.0005)	0.422 (0.0037)	0.225 (0.0030)
BETEL	Informative	0.208 (0.0096)	0.089 (0.0012)	0.291 (0.0032)	0.146 (0.0016)	0.236 (0.0051)	0.115 (0.0029)	0.192 (0.0028)	0.093 (0.0017)	0.062 (0.0004)	0.024 (0.0002)	0.145 (0.0027)	0.074 (0.0015)
	Non-informative	0.236 (0.0092)	0.109 (0.0020)	0.324 (0.0042)	0.176 (0.0029)	0.260 (0.0054)	0.140 (0.0034)	0.215 (0.0035)	0.114 (0.0022)	0.064 (0.0004)	0.026 (0.0003)	0.162 (0.0030)	0.086 (0.0021)
	Misspecified	0.503 (0.0060)	0.420 (0.0031)	0.656 (0.0135)	0.583 (0.0116)	0.456 (0.0121)	0.392 (0.0099)	0.393 (0.0087)	0.321 (0.0065)	0.127 (0.0031)	0.095 (0.0018)	0.379 (0.0097)	0.320 (0.0074)
RETEL	Informative	0.171 (0.0029)	0.086 (0.0011)	0.286 (0.0032)	0.145 (0.0017)	0.210 (0.0020)	0.105 (0.0014)	0.172 (0.0015)	0.084 (0.0009)	0.075 (0.0003)	0.028 (0.0002)	0.142 (0.0014)	0.071 (0.0007)
	Non-informative	0.210 (0.0041)	0.113 (0.0022)	0.329 (0.0045)	0.180 (0.0036)	0.233 (0.0032)	0.126 (0.0023)	0.193 (0.0025)	0.100 (0.0017)	0.090 (0.0003)	0.031 (0.0003)	0.155 (0.0020)	0.081 (0.0013)
	Misspecified	0.490 (0.0035)	0.424 (0.0035)	0.654 (0.0136)	0.583 (0.0118)	0.502 (0.0081)	0.419 (0.0068)	0.437 (0.0052)	0.340 (0.0044)	0.633 (0.0010)	0.525 (0.0006)	0.497 (0.0044)	0.368 (0.0030)
Parametric	Informative	0.174 (0.0026)	0.112 (0.0015)	0.283 (0.0036)	0.144 (0.0018)	0.209 (0.0023)	0.105 (0.0014)	0.182 (0.0025)	0.089 (0.0012)	0.060 (0.0003)	0.024 (0.0002)	0.140 (0.0014)	0.070 (0.0007)
Empirical	-	0.218 (0.0084)	0.124 (0.0039)	0.418 (0.0091)	0.229 (0.0058)	0.279 (0.0060)	0.153 (0.0039)	0.218 (0.0042)	0.118 (0.0026)	0.065 (0.0005)	0.026 (0.0003)	0.161 (0.0034)	0.087 (0.0023)
PRECiSE-CO96	-	0.211 (0.0047)	0.115 (0.0021)	0.337 (0.0051)	0.187 (0.0039)	0.251 (0.0032)	0.136 (0.0023)	0.213 (0.0024)	0.111 (0.0015)	0.101 (0.0002)	0.036 (0.0002)	0.174 (0.0022)	0.093 (0.0012)
HC	-	0.205 (0.0033)	0.110 (0.0019)	0.357 (0.0023)	0.213 (0.0018)	0.242 (0.0022)	0.137 (0.0015)	0.198 (0.0019)	0.108 (0.0010)	0.119 (0.0002)	0.038 (0.0002)	0.166 (0.0015)	0.087 (0.0007)

ensures that, for $x \in [0,1]$, each one-step factor in the wealth process satisfies

$$1 + \lambda(x - \mu) \geq 1 - c > 0$$

thereby inducing a valid nonnegative test martingale.

The truncation level c controls the amount of leverage available to the betting strategy, i.e., the largest fraction of the current wealth that can be lost in a single time step. This exposes a genuine trade-off. If $c_1 < c_2$, then $I_{\mu,c_1} \subset I_{\mu,c_2}$; that is, increasing c enlarges the oracle's feasible set and thus cannot decrease the optimal expected log growth. As a result, a larger c can improve efficiency when the predictive distribution used to select the betting coefficient is accurate, but increases exposure to bad bets if that is not the case.

This also explains the dependence on c through $L_{\mu,c}$ in the upper bound of Theorem 2. Theorem 2 bounds the per-observation expected log-growth gap relative to the oracle in terms of the average predictive error measured in Wasserstein distance. As shown in the proof of Theorem 1, $L_{\mu,c}$ is a Lipschitz constant in x, uniform over $\lambda \in I_{\mu,c}$, for the one-step log factor $\log\{1 + \lambda(x - \mu)\}$. The latter's derivative with respect to x becomes large when the linear factor $1 + \lambda(x - \mu)$ approaches zero. Since the admissible coefficient range guarantees the lower bound $1 + \lambda(x - \mu) \geq 1 - c$, this worst-case sensitivity necessarily deteriorates as $c \to 1$, for fixed μ. However, $L_{\mu,c}$ is obtained by taking the worst case uniformly over all $\lambda \in I_{\mu,c}$ and $x \in [0,1]$, and can therefore be conservative in a particular experiment. Nevertheless, its behaviour as $c \to 1$ captures the potential vulnerability of highly leveraged bets to predictive misspecification: an inaccurate predictive may select an aggressive yet suboptimal coefficient whose one-step log loss is extremely costly.

To assess this trade-off empirically, we run additional synthetic experiments with $c \in \{0.5, 0.75, 0.9, 0.95, 0.99\}$ under three contrasting data-generating processes: $\mathrm{Bernoulli}(0.5)$, which has maximal variance among distributions on $[0,1]$ and whose observations lie at the support boundaries; $\mathrm{Beta}(10, 30)$, which is concentrated in the interior; and the asymmetric beta mixture, which departs from the beta working model. Table 4 reports the average widths at $n = 100$ and $n = 1000$ over 100 replicates.

The results reveal a clear leverage-robustness trade-off. For $\mathrm{Bernoulli}(0.5)$, neither the oracle nor informative Bayes-assisted methods benefit from increasing c, indicating that $c = 0.5$ already contains the relevant optimal coefficients. Compared with $c = 0.5$, at $c = 0.95$ empirical is about 23% wider at both horizons, HC is 36–56% wider, and misspecified MDP and BETEL are 23–28% wider. Because every Bernoulli observation occurs at a support endpoint, an overly aggressive bet is repeatedly exposed to the most adverse possible outcome.

For $\mathrm{Beta}(10, 30)$, the pattern reverses. At $n = 100$, increasing c from 0.5 to 0.95 narrows substantially the oracle, empirical, and informative Bayes-assisted CSs. Here, the smaller admissible range excludes useful betting coefficients, while the concentration of the observations away from the endpoints makes greater leverage beneficial. The gains begin to saturate by $c = 0.95$: at $n = 100$, increasing c further to 0.99 changes the oracle and informative Bayes-assisted widths by at most 2.2%, while at $n = 1000$ the changes are negligible. The misspecified-prior results are more method- and sample-dependent, and therefore do not support a universal monotonic robustness claim; nevertheless, the accurate-prior results clearly show that a larger c can exploit better oracle bets when they are available.

The beta-mixture results are intermediate: the oracle and most informative-prior methods are nearly insensitive to c, while empirical, HC, and BETEL widen modestly as c increases.

Overall, we use $c = 0.95$ for the empirical and Bayes-assisted experiments as an aggressive, nearly unconstrained experimental choice, not as a universal default. When the predictive is accurate, this choice captures most of the available oracle benefit when larger coefficients are useful, while not forcing their use when a smaller coefficient is optimal. Moderate choices such as $c \in [0.5, 0.75]$ provide stronger regularisation when the predictive is uncertain or potentially misspecified. In these experiments, values closer to one provide little additional oracle improvement while reducing the worst-case one-step wealth floor from 0.05 at $c = 0.95$ to 0.01 at $c = 0.99$.

G.1.5 MDP mixing parameter κ

The MDP predictive takes the form (Equation (19))

Table 4: Average CS width at $n = 100$ / $n = 1000$ for each truncation level c. Widths are averaged over the non-empty CSs; across all CSs shown, at most 6% are empty at $n = 100$ or $n = 1000$ (0.4% on average).

(a) Bernoulli(0.5)

Method	Oracle	MDP			BETEL			RETEL			Parametric			Empirical	HC
Prior	-	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	-	-
c															
0.5	0.201 / 0.064	0.206 / 0.074	0.232 / 0.084	0.370 / 0.148	0.204 / 0.066	0.238 / 0.087	0.507 / 0.349	0.202 / 0.066	0.240 / 0.089	0.507 / 0.349	0.200 / 0.065	0.224 / 0.077	0.487 / 0.304	0.252 / 0.094	0.276 / 0.115
0.75	0.201 / 0.064	0.206 / 0.074	0.250 / 0.092	0.426 / 0.172	0.204 / 0.066	0.239 / 0.086	0.580 / 0.415	0.202 / 0.066	0.244 / 0.089	0.578 / 0.415	0.200 / 0.065	0.242 / 0.085	0.555 / 0.362	0.276 / 0.102	0.318 / 0.140
0.9	0.201 / 0.064	0.206 / 0.074	0.260 / 0.098	0.459 / 0.180	0.204 / 0.066	0.239 / 0.086	0.624 / 0.433	0.202 / 0.066	0.244 / 0.089	0.623 / 0.433	0.200 / 0.065	0.253 / 0.090	0.613 / 0.390	0.297 / 0.111	0.354 / 0.164
0.95	0.201 / 0.064	0.206 / 0.074	0.261 / 0.099	0.472 / 0.184	0.204 / 0.066	0.239 / 0.086	0.624 / 0.433	0.202 / 0.066	0.244 / 0.089	0.623 / 0.433	0.200 / 0.065	0.254 / 0.091	0.633 / 0.398	0.313 / 0.117	0.375 / 0.180
0.99	0.201 / 0.064	0.206 / 0.074	0.262 / 0.099	0.491 / 0.190	0.204 / 0.066	0.239 / 0.086	0.624 / 0.433	0.202 / 0.066	0.244 / 0.089	0.623 / 0.433	0.200 / 0.065	0.254 / 0.091	0.658 / 0.407	0.349 / 0.129	0.418 / 0.213

(b) Beta(10, 30)

Method	Oracle	MDP			BETEL			RETEL			Parametric			Empirical	HC
Prior	-	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	-	-
c															
0.5	0.051 / 0.011	0.052 / 0.012	0.062 / 0.013	0.332 / 0.034	0.053 / 0.012	0.054 / 0.012	0.121 / 0.039	0.054 / 0.012	0.062 / 0.013	0.591 / 0.213	0.052 / 0.012	0.063 / 0.013	0.579 / 0.053	0.055 / 0.012	0.066 / 0.014
0.75	0.040 / 0.010	0.041 / 0.012	0.050 / 0.012	0.347 / 0.034	0.041 / 0.012	0.043 / 0.012	0.115 / 0.040	0.045 / 0.012	0.053 / 0.012	0.589 / 0.221	0.041 / 0.011	0.053 / 0.012	0.581 / 0.057	0.044 / 0.012	0.053 / 0.014
0.9	0.036 / 0.010	0.038 / 0.012	0.047 / 0.012	0.354 / 0.034	0.038 / 0.012	0.040 / 0.012	0.113 / 0.041	0.044 / 0.012	0.052 / 0.012	0.589 / 0.226	0.037 / 0.012	0.049 / 0.012	0.582 / 0.062	0.041 / 0.012	0.048 / 0.014
0.95	0.036 / 0.010	0.037 / 0.012	0.046 / 0.012	0.356 / 0.034	0.037 / 0.012	0.039 / 0.012	0.112 / 0.042	0.044 / 0.012	0.052 / 0.012	0.589 / 0.227	0.037 / 0.012	0.049 / 0.012	0.582 / 0.063	0.040 / 0.012	0.047 / 0.013
0.99	0.035 / 0.010	0.036 / 0.012	0.045 / 0.012	0.357 / 0.034	0.037 / 0.012	0.039 / 0.012	0.112 / 0.042	0.044 / 0.012	0.052 / 0.012	0.589 / 0.229	0.036 / 0.012	0.048 / 0.012	0.582 / 0.064	0.039 / 0.012	0.046 / 0.014

(c) Beta mixture

Method	Oracle	MDP			BETEL			RETEL			Parametric			Empirical	HC
Prior	-	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	Informative	Non-informative	Misspecified	-	-
c															
0.5	0.095 / 0.031	0.098 / 0.036	0.107 / 0.040	0.293 / 0.072	0.097 / 0.033	0.109 / 0.040	0.355 / 0.167	0.098 / 0.033	0.109 / 0.039	0.408 / 0.171	0.096 / 0.037	0.108 / 0.045	0.480 / 0.169	0.109 / 0.040	0.118 / 0.046
0.75	0.096 / 0.031	0.099 / 0.036	0.109 / 0.041	0.311 / 0.088	0.100 / 0.034	0.113 / 0.042	0.352 / 0.187	0.098 / 0.033	0.110 / 0.040	0.423 / 0.193	0.096 / 0.037	0.109 / 0.046	0.487 / 0.192	0.113 / 0.042	0.119 / 0.049
0.9	0.096 / 0.031	0.099 / 0.036	0.110 / 0.041	0.320 / 0.095	0.103 / 0.034	0.116 / 0.043	0.352 / 0.198	0.098 / 0.033	0.110 / 0.040	0.432 / 0.206	0.096 / 0.037	0.110 / 0.046	0.492 / 0.205	0.116 / 0.044	0.122 / 0.052
0.95	0.096 / 0.031	0.099 / 0.036	0.110 / 0.041	0.322 / 0.097	0.103 / 0.035	0.117 / 0.043	0.352 / 0.202	0.098 / 0.033	0.110 / 0.040	0.434 / 0.210	0.096 / 0.037	0.110 / 0.046	0.493 / 0.208	0.118 / 0.044	0.123 / 0.053
0.99	0.096 / 0.031	0.099 / 0.036	0.110 / 0.042	0.324 / 0.098	0.104 / 0.035	0.118 / 0.043	0.352 / 0.204	0.098 / 0.033	0.110 / 0.040	0.436 / 0.213	0.096 / 0.037	0.110 / 0.046	0.494 / 0.210	0.119 / 0.044	0.125 / 0.053

$$\mathbb{Q}_{n+1|n} = \frac{\kappa}{\kappa + n}\mathbb{H}_{n+1|n} + \frac{n}{\kappa + n}\mathbb{P}_n,$$

where κ controls the prior-driven component's weight relative to the empirical distribution and hence how long the predictive remains prior-driven. For every fixed κ, this component vanishes asymptotically.

The distinction between empirical and MDP is therefore primarily finite-sample, which is relevant in sequential settings. At small sample sizes, the empirical predictive is supported on only a few observations and can yield noisy betting coefficients. When accurate external information is available, the Bayesian component can regularise this predictive, improve early evidence growth, and thus reduce stopping times. For example, in the AttributedQA sequential test in Section 6, MDP requires 52.18 labels on average, compared with 65.24 for the empirical predictive, a reduction of about 20%, and wins or ties against it in 98 of 100 paired replicates. Conversely, an inaccurate prior can reduce finite-sample efficiency relative to empirical.

To quantify this trade-off, we provide additional results for $\mathrm{Bernoulli}(0.5)$, $\mathrm{Beta}(10, 30)$, and the beta mixture, varying $\kappa \in \{5, 20, 50, 200\}$. Table 5 compares average CS widths at $n = 100$ and $n = 1000$ for all three data-generating processes.

In the Bernoulli experiment, over jointly non-empty paired replicates, an accurate informative prior makes MDP about 27% narrower than empirical for $\kappa = 5$ and 34–37% narrower for $\kappa = 50$ across the two horizons. These results show that the finite-sample benefit can remain substantial well beyond the very earliest observations.

The same experiment also makes the cost of misspecification explicit. Larger κ generally magnifies both the potential gain and the potential loss. Although the prior weight vanishes for fixed κ, early differences in accumulated wealth can remain visible at substantially larger sample sizes. In the low-variance Beta experiment, there is comparatively little room for an accurate prior to improve upon empirical, while a severely misspecified prior can cause a substantial efficiency loss. The contrast between Bernoulli and Beta is consistent with the intuition that prior regularisation is most useful when the empirical predictive is intrinsically noisy and takes time to adapt to the data.

A practical calibration is to specify the desired weight w_{ref} assigned to the prior-driven predictive at a reference sample size n_{ref}, giving $\kappa = w_{\mathrm{ref}} / (1 - w_{\mathrm{ref}})\, n_{\mathrm{ref}}$. Outside this ablation, we use $\kappa = 50$, corresponding to prior predictive weights of 0.5 at $n = 50$ and approximately 0.05 at $n = 1000$, so that the prior can materially affect the early sequential regime while the empirical predictive rapidly becomes dominant.

G.1.6 Cellwise-certified inversion

We compare the pointwise and cellwise implementations on the same length-1000 data sequence and method seed in each of 100 paired replicates of all six synthetic experiments. We use $\alpha = 0.1$ and a uniform partition of $[0, 1]$ into $G = 500$ cells of width $1/500$, matching the pointwise grid resolution; all other method, prior, and running-intersection settings match the main synthetic suite. Tables 6 and 7 report mean-width ratios over replicates in which both CSs are non-empty: cellwise/pointwise for the predictive methods and, because UP uses the same grid in the main experiments to match their numerical resolution, direct-interval/grid for UP. For the predictive methods, these ratios range from 0.983 to 1.045 at $n = 100$ and from 0.942 to 1.085 at $n = 1000$, so the widths differ by at most 9%. Overall, this close agreement indicates that the pointwise grid implementation used for the predictive methods in the other experiments has little effect on the reported widths in these settings.

G.2 Best-arm identification for LLM evaluation

G.2.1 Full baseline comparison

Table 9 reports the full results for the best-arm identification experiment in Section 6.

Table 5: Average CS width at $n = 100$ / $n = 1000$ for each MDP mixing parameter κ. Widths are averaged over the non-empty CSs; across all CSs shown, at most 6% are empty at $n = 100$ or $n = 1000$ (0.9% on average).

(a) Bernoulli(0.5)

Method	Oracle	MDP			Empirical
Prior	-	Informative	Non-informative	Misspecified	-
κ					
5	0.201 / 0.064	0.229 / 0.084	0.278 / 0.107	0.328 / 0.119	0.313 / 0.117
20	0.201 / 0.064	0.212 / 0.077	0.267 / 0.102	0.401 / 0.147	0.313 / 0.117
50	0.201 / 0.064	0.206 / 0.074	0.261 / 0.099	0.472 / 0.184	0.313 / 0.117
200	0.201 / 0.064	0.201 / 0.069	0.256 / 0.095	0.569 / 0.258	0.313 / 0.117

(b) Beta(10, 30)

Method	Oracle	MDP			Empirical
Prior	-	Informative	Non-informative	Misspecified	-
κ					
5	0.036 / 0.010	0.038 / 0.012	0.042 / 0.012	0.118 / 0.017	0.040 / 0.012
20	0.036 / 0.010	0.037 / 0.012	0.044 / 0.012	0.244 / 0.025	0.040 / 0.012
50	0.036 / 0.010	0.037 / 0.012	0.046 / 0.012	0.356 / 0.034	0.040 / 0.012
200	0.036 / 0.010	0.037 / 0.012	0.048 / 0.012	0.494 / 0.049	0.040 / 0.012

(c) Beta mixture

Method	Oracle	MDP			Empirical
Prior	-	Informative	Non-informative	Misspecified	-
κ					
5	0.096 / 0.031	0.107 / 0.039	0.113 / 0.042	0.159 / 0.055	0.118 / 0.044
20	0.096 / 0.031	0.102 / 0.038	0.111 / 0.042	0.243 / 0.075	0.118 / 0.044
50	0.096 / 0.031	0.099 / 0.036	0.110 / 0.041	0.322 / 0.097	0.118 / 0.044
200	0.096 / 0.031	0.097 / 0.035	0.110 / 0.042	0.425 / 0.137	0.118 / 0.044

Table 6: Cellwise-to-pointwise mean-width ratio at $n = 100$. For UP, the analogous direct-interval/grid ratio is included for reference. Averages are taken over the replicates in which both CSs are non-empty (see Table 8); across all CSs, at most 2% are empty at $n = 100$ (0.1% on average).

	Bernoulli(0.1)	Bernoulli(0.5)	Beta(0.5, 0.5)	Beta(1, 1)	Beta(10, 30)	Beta mixture
Emp.	0.998	0.996	0.996	0.996	0.984	0.994
Oracle	0.997	0.998	0.997	0.997	0.995	0.995
Param. (good)	1.003	0.998	0.998	0.999	0.990	0.994
Param. (bad)	0.999	1.003	0.998	0.998	0.998	0.998
Param. (noninf)	1.045	0.999	0.997	0.995	0.984	0.994
MDP (good)	0.996	0.998	0.996	0.998	0.983	0.991
MDP (bad)	1.005	1.001	0.999	0.996	0.997	0.997
MDP (noninf)	1.026	0.999	0.998	0.995	0.986	0.993
BETEL (good)	0.997	0.998	0.998	0.997	0.990	0.994
BETEL (bad)	1.021	1.004	0.998	0.997	0.991	0.998
BETEL (noninf)	0.996	0.998	0.997	0.997	0.988	0.996
RETEL (good)	0.995	0.997	0.998	0.996	0.985	0.994
RETEL (bad)	1.019	1.004	0.999	0.998	0.998	0.998
RETEL (noninf)	0.999	0.999	0.997	0.995	0.987	0.994
UP	0.987	0.992	0.989	0.985	0.959	0.982

Table 7: Cellwise-to-pointwise mean-width ratio at $n = 1000$. For UP, the analogous direct-interval/grid ratio is included for reference. Averages are taken over the replicates in which both CSs are non-empty (see Table 8); across all CSs, at most 6% are empty at $n = 1000$ (0.8% on average).

	Bernoulli(0.1)	Bernoulli(0.5)	Beta(0.5, 0.5)	Beta(1, 1)	Beta(10, 30)	Beta mixture
Emp.	0.992	0.996	0.995	0.994	0.966	0.990
Oracle	0.988	0.992	0.994	0.988	0.942	0.981
Param. (good)	1.012	0.993	0.984	0.988	0.960	0.987
Param. (bad)	1.002	1.001	0.998	0.996	0.984	0.995
Param. (noninf)	1.085	0.997	0.992	0.991	0.949	0.991
MDP (good)	0.993	0.992	0.990	0.987	0.968	0.979
MDP (bad)	1.000	0.999	0.998	0.997	0.982	0.997
MDP (noninf)	1.003	0.997	0.992	0.991	0.952	0.990
BETEL (good)	0.987	0.992	0.992	0.984	0.966	0.979
BETEL (bad)	1.042	1.002	0.998	0.995	0.981	0.995
BETEL (noninf)	0.994	0.996	0.994	0.994	0.958	0.994
RETEL (good)	0.988	0.992	0.988	0.989	0.944	0.981
RETEL (bad)	1.040	1.002	0.998	0.996	0.997	0.995
RETEL (noninf)	0.992	0.996	0.993	0.991	0.949	0.984
UP	0.966	0.977	0.969	0.961	0.833	0.951

Table 8: Proportion of replicates in which the cellwise CS is empty. Methods and scenarios not listed have no empty cellwise CS; the corresponding pointwise rates differ by at most one percentage point. Across both implementations, at most 6% are empty (0.5% on average).

		Bernoulli(0.1)	Bernoulli(0.5)	Beta(0.5, 0.5)	Beta(1, 1)	Beta mixture
$n = 100$	Emp.	0.00	0.02	0.01	0.00	0.00
	Param. (noninf)	0.00	0.01	0.00	0.00	0.00
	MDP (noninf)	0.00	0.01	0.00	0.00	0.00
	BETEL (good)	0.00	0.00	0.02	0.00	0.01
	BETEL (bad)	0.00	0.00	0.02	0.00	0.00
$n = 1000$	Emp.	0.03	0.05	0.03	0.01	0.06
	Param. (bad)	0.00	0.02	0.00	0.00	0.00
	Param. (noninf)	0.00	0.05	0.02	0.00	0.02
	MDP (bad)	0.01	0.02	0.01	0.00	0.03
	MDP (noninf)	0.02	0.05	0.02	0.00	0.03
	BETEL (good)	0.00	0.00	0.03	0.00	0.02
	BETEL (bad)	0.00	0.00	0.02	0.00	0.00
	BETEL (noninf)	0.00	0.00	0.02	0.00	0.04
	RETEL (noninf)	0.02	0.02	0.01	0.00	0.02
	UP	0.02	0.02	0.01	0.00	0.03

Table 9: Full-baseline version of the LLM best-arm identification experiment. Across all methods, no CS is empty at the stopping time.

Method	Avg Pulls (s.e.)	Avg rank (s.e.)	W/T/L vs Empirical
MDP	91.68 (3.90)	1.95 (0.13)	38/5/7
Parametric	95.92 (4.75)	2.66 (0.26)	37/1/12
BETEL	99.36 (3.62)	3.28 (0.17)	23/7/20
Empirical	101.64 (4.03)	3.58 (0.21)	–
RETEL	109.76 (3.63)	4.95 (0.23)	18/2/30
UP	110.68 (3.53)	5.18 (0.16)	4/3/43
PRECiSE-CO96	123.84 (4.16)	6.54 (0.13)	0/0/50
HC	147.96 (3.51)	7.86 (0.09)	0/0/50

Overall, the results presented in the main text stand, with the parametric Bayes-assisted methods typically outperforming the prior-free baselines on this task. The additional prior-free baselines, PRECiSE-CO96 and HC, decisively underperform all methods reported in the main text. The nonparametric Bayesian predictive performs similarly to the empirical predictive. All this suggests that the Beta working model is a good fit for this example.

G.2.2 Paired significance comparisons

To establish the statistical significance of the performance disparities among the evaluated methods in Tables 1a and 9, we conduct pairwise comparisons across all unique combinations of algorithms. For every unordered pair, denoted as $\{A, B\}$, we employ a two-sided Wilcoxon signed-rank test [72] to detect structural differences in the number of pulls required to converge. To rigorously control the family-wise error rate stemming from this testing matrix, the resulting p-values are adjusted utilizing the Holm-Bonferroni step-down procedure [73], bounding the global significance level at $\alpha = 0.05$. When a statistically significant difference is established, the superior algorithm is subsequently identified as the one exhibiting the lower mean number of pulls. Ties between paired samples are handled via the zero-split method. The results are presented in Table 10.

	MDP	Parametric	BETEL	Empirical	RETEL	UP	PRECiSE-CO96	HC
MDP	-	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Parametric	No	-	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
BETEL	No	No	-	?	**Yes**	**Yes**	**Yes**	**Yes**
Empirical	No	No	?	-	?	**Yes**	**Yes**	**Yes**
RETEL	No	No	No	?	-	?	**Yes**	**Yes**
UP	No	No	No	No	?	-	**Yes**	**Yes**
PRECiSE-CO96	No	No	No	No	No	No	-	**Yes**
HC	No	No	No	No	No	No	No	-

Table 10: Pairwise statistical comparison: Is Method A (row) significantly faster than Method B (column)? Across all methods, no CS is empty at the stopping time.

G.2.3 Prior misspecification

Misspecification of the working prior does not affect statistical validity, but it can reduce efficiency. For the LUCB experiment of Section 6 and Appendix G.2, we centre the priors at the well-specified mean minus one standard deviation of the calibration data used to compute that mean, while leaving everything else unchanged. Table 11 reports average LUCB stopping times and paired ranks for accurate and inaccurate priors. All methods identify the same LLM, but the inaccurate priors require considerably more LLM calls.

Table 11: LLM best-arm identification experiment with informative and misspecified priors. Across all methods, no CS is empty at the stopping time.

Method	Avg Pulls (s.e.)	Avg rank (s.e.)
MDP	91.68 (3.90)	1.67 (0.09)
Parametric	95.92 (4.75)	2.16 (0.13)
BETEL	99.36 (3.62)	2.73 (0.15)
RETEL	109.76 (3.63)	3.90 (0.16)
MDP (bad)	149.64 (7.26)	5.00 (0.11)
BETEL (bad)	171.40 (10.76)	5.85 (0.08)
Parametric (bad)	246.04 (19.21)	7.19 (0.16)
RETEL (bad)	229.84 (14.54)	7.50 (0.08)

G.3 Prediction-powered inference

G.3.1 Full baseline comparison

Figure 9 shows the full-baseline version of the prediction-powered inference experiment in Figure 2. While the results remain consistent with the main text, it is interesting to see that the parametric Bayesian predictive performs well on this task, especially on the ATTRIBUTEDQA dataset, where the working model seems to be a good fit for the data distribution.

Table 12 reports the full comparison for the sequential test of $H_0 : \theta^\star \leq 0$ on the ATTRIBUTEDQA dataset, including additional classical and PPI baselines not reported in the main text. While remaining consistent with the comments in the main text, the full comparison highlights again the strong performance of the parametric Bayesian predictive on this task, where it competes closely with MDP.

Table 12: Full comparison for the AttributedQA sequential test of $H_0 : \theta^\star \leq 0$. Avg labels is the average number of labelled examples required to reject the null over 100 replicates. Avg rank ranks methods by stopping time within each replicate. W/T/L gives paired wins, ties, and losses against the PPI empirical predictive baseline. Across all methods, at most 1% of CSs are empty at the stopping time (0.1% on average).

Method	Avg labels (s.e.)	Avg rank (s.e.)	W/T/L vs Empirical
Parametric	49.43 (3.23)	2.91 (0.16)	80 / 18 / 2
BETEL	50.25 (3.37)	3.33 (0.18)	71 / 15 / 14
MDP	52.18 (3.50)	3.49 (0.18)	82 / 16 / 2
RETEL	55.03 (2.49)	4.24 (0.22)	62 / 8 / 30
Empirical	65.24 (4.12)	5.88 (0.25)	–
UP	69.73 (3.24)	6.48 (0.16)	38 / 4 / 58
HC	87.82 (2.94)	8.42 (0.18)	18 / 5 / 77
PRECiSE-CO96	89.87 (3.71)	8.90 (0.16)	9 / 1 / 90
UP (classical)	106.27 (7.48)	7.24 (0.32)	30 / 0 / 70
PRECiSE-CO96 (classical)	112.23 (7.46)	8.43 (0.32)	27 / 0 / 73
HC (classical)	114.01 (9.02)	7.78 (0.33)	25 / 2 / 73
Empirical (classical)	188.91 (10.06)	10.93 (0.30)	11 / 0 / 89

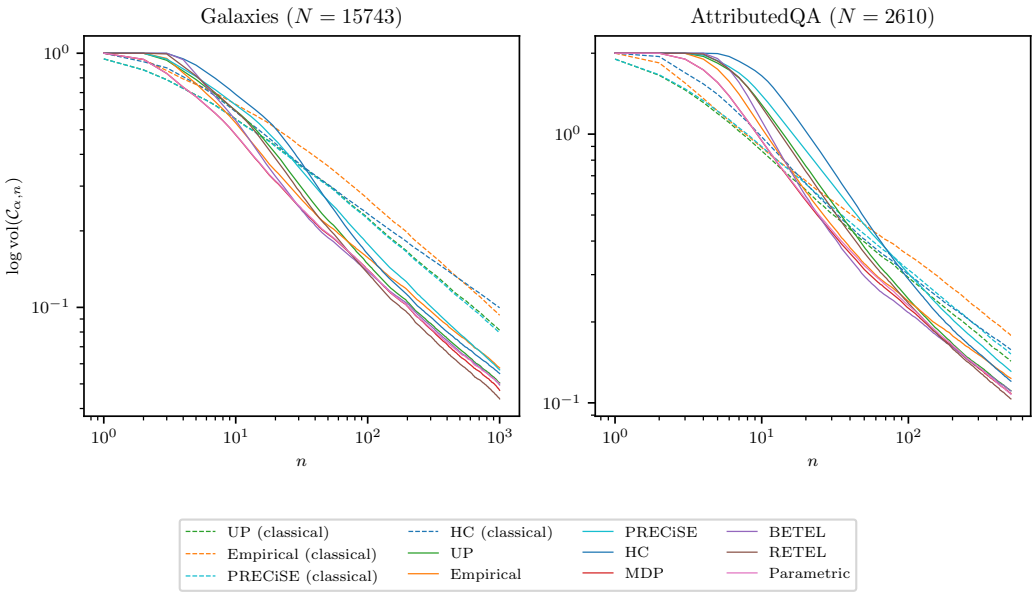

Figure 9: Full-baseline version of prediction-powered mean estimation experiment in Figure 2. The left and right panels show average CS length over 100 replicates for the GALAXIES and ATTRIBUTEDQA datasets. Curves labelled "classical" use only labelled observations, while the remaining curves use the PPI construction. Empty CSs are counted as width zero; across all CSs shown, at most 6% are empty at the final sample size (0.5% on average).

G.3.2 Prior misspecification

For the AttributedQA experiment described in Section 6 and Appendix F.3, we shift the prior means from 0.5 to 0.45 while leaving the variances unchanged. Although the change in the mean is modest, these priors are strongly concentrated and therefore encode a substantially shifted belief about the quality of the black-box predictor. The table below reports the average capped number of labels to reject $H_0 : \theta^* \leq 0$ and paired ranks across 100 replicates. All four methods reject within the 500-label budget in every replicate. The shifted priors nevertheless delay rejection substantially for some methods. MDP exhibits the smallest degradation among the prior-assisted methods, consistent with its increasing reliance on the empirical component as additional observations are collected.

Table 13: AttributedQA PPI experiment with informative and misspecified priors. Across all methods, no CS is empty at the stopping time.

Method	Avg labels (s.e.)	Avg rank (s.e.)
Parametric	49.43 (3.23)	2.08 (0.07)
BETEL	50.25 (3.37)	2.46 (0.11)
MDP	52.18 (3.50)	2.63 (0.11)
RETEL	55.03 (2.49)	3.18 (0.13)
MDP (bad)	77.16 (3.01)	5.11 (0.06)
BETEL (bad)	87.07 (4.18)	5.79 (0.12)
Parametric (bad)	107.92 (3.36)	7.18 (0.05)
RETEL (bad)	112.75 (2.79)	7.59 (0.06)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's actual contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 discusses limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Proofs of all results are included in full detail in the supplementary material, except for some very classical results for which we provide only references.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix D discusses implementations details, and Appendix F provides the information needed to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A link to a public GitHub repository containing code to reproduce all experiments is provided.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details necessary to understand the results are provided both in the main text and in the supplementary material; see, e.g., Appendices D and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: standard errors are reported for averages reported in all tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: information on compute resources is provided in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper is mainly theoretical and uses only publicly available datasets, which do not contain any sensitive information.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work performed is mainly theoretical, and we do not foresee any societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work performed is mainly theoretical and doesn't pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external datasets and LLM checkpoints are credited and linked in the supplementary material, where the applicable model licences are identified. LM Studio is identified by version and was used locally.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release such assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core method development in this research does not involve LLMs as important, original, or non-standard components. Some experiments explicitly involve LLMs as subjects of study; their use is described in full detail in the appendix as part of the detailed description of the experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.