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

Valentin Kilian [*1] **Stefano Cortinovis** [*1] **François Caron** [1]

Abstract

Confidence sequences based on test martingales provide time-uniform uncertainty quantification for the mean of bounded IID 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 (Ville, 1939; Darling & Robbins, 1967; Lai, 1976; Howard et al., 2021; Waudby-Smith et al., 2024; Ramdas & Wang, 2024). 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.

The key practical question is how to choose these predictable coefficients. GRAPA, the Growth Rate Adaptive to the Particular Alternative strategy of Waudby-Smith & Ramdas (2024), 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 (Kumon et al., 2008; 2011; Kaufmann & Koolen, 2021; Cortinovis et al., 2025). 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

[*]Equal contribution [1]Department of Statistics, University of Oxford, Oxford, United Kingdom. Correspondence to: Valentin Kilian <kilian@stats.ox.ac.uk>.

Accepted to the ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea. Copyright 2026 by the author(s).

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 (Ferguson, 1973; Blackwell & MacQueen, 1973; Antoniak, 1974); 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 (Owen, 2001; Lazar, 2003; Schennach, 2005; Kim et al., 2024).

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. 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 (Hoeffding, 1963; Bennett, 1962; Bernshtein, 1927; Bentkus, 2004; Audibert et al., 2007; Maurer & Pontil, 2009). 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 (Pratt, 1961; 1963; Brown et al., 1995; Yu & Hoff, 2018; Farchione & Kabaila, 2008; Kabaila & Giri, 2013; Kabaila & Farchione, 2022; Hoff & Yu, 2019; Hoff, 2023; Cortinovis & Caron, 2024; 2025). 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 & Robbins (1967); Robbins (1970); Robbins & Siegmund (1970); Lai (1976), building on martingale and sequential-testing foundations (Ville, 1939; Wald, 1945). Modern treatments use nonnegative supermartingales, e-values, and e-processes to obtain validity under optional stopping and continuation (Howard et al., 2021; Waudby-Smith et al., 2024; Shafer & Vovk, 2001; 2019; Ramdas et al., 2023; Ramdas & Wang, 2025; 2024; Grünwald et al., 2024). For bounded means, betting martingales are especially relevant: Waudby-Smith & Ramdas (2024) introduced betting CIs and CSs, with GRAPA choosing stakes by empirical log-growth optimization, while Orabona & Jun (2023) obtained PRECiSE sequences via universal-portfolio regret bounds (Cover, 1991; Cover & Ordentlich, 2002). Near-optimality and optimality of betting confidence sets have since been studied by Shekhar & Ramdas (2023); Clerico (2025), and closed-form empirical-Bernstein sequences give strong variance-adaptive baselines (Chugg & Ramdas, 2025). Our method follows the Kelly log-growth principle (Kelly, 1956), 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 (Kumon

et al., 2008; 2011; Kaufmann & Koolen, 2021; Cortinovis et al., 2025). The nonparametric Bayes-assisted confidence sequences of Kilian et al. (2025) are related, but their guarantees are asymptotic, whereas ours are nonasymptotic for bounded observations. We instantiate the betting rule using Dirichlet-process posterior predictives (Ferguson, 1973; Blackwell & MacQueen, 1973; Antoniak, 1974) and Bayesian empirical or exponentially tilted empirical likelihood (Owen, 2001; Lazar, 2003; Schennach, 2005; Chen et al., 2008; Emerson & Owen, 2009; Zhu et al., 2009; Liu & Chen, 2010; Kim et al., 2024). 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. (Ramdas & Wang, 2024). 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

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

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 (Waudby-Smith & Ramdas, 2024; Clerico, 2025).

Proposition 3.1 (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 3.2 (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

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

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):

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

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 4.1 (Asymptotic log-optimality)**.** *Let $(X_n)_{n\geq1}$ be i.i.d. in $[0,1]$ from a non-degenerate distribution $\mathbb{P}^*$, and let $\mathcal{F}_n=\sigma(X_1,\dots,X_n)$. Let $(\mathbb{Q}_{n|n-1}(\cdot\mid X_{1:n-1}))_{n\geq1}$ be an $\mathcal{F}_{n-1}$-measurable sequence of predictive distributions on $[0,1]$ such that*

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

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

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

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

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

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

Moreover, for

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

$$\frac{1}{n}\log W_n(X_{1:n},\mu)\to M(\lambda^*(\mu),\mu) \quad (15)$$

in probability, and

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

Proposition 3.2 gives anytime-valid coverage for any sequence of predictive distributions. In contrast, Theorem 4.1 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 4.2 (Predictive error and expected log-evidence)**.** *With the notation of Theorem 4.1,*

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

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 3.2 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 4.1 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

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

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

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

The form in (19) is inspired by the Pólya-urn posterior predictive of a Dirichlet process (Ferguson, 1973; Blackwell & MacQueen, 1973). When G_0 is non-atomic and all observations are distinct, it corresponds to the posterior predictive in Antoniak's mixture of Dirichlet processes (Antoniak, 1974). 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 & Ramdas (2024).

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 (Walker, 2013; Bissiri et al., 2016; Lyddon et al., 2018).

Proposition 5.1 (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*

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

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 4.1, 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) (Lazar, 2003; Schennach, 2005) and a regularized version (RETEL), related to adjusted and regularized empirical likelihood methods (Chen et al., 2008; Emerson & Owen, 2009; Liu & Chen, 2010; Zhu et al., 2009; Kim et al., 2024).

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

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

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

$$\frac{d\widetilde{\mathbb{P}}_{n,\mu}}{d\mathbb{P}_{n,\mu}}(z) = \exp\{\gamma_{n,\mu}(z-\mu) - \psi_{n,\mu}(\gamma_{n,\mu})\}, \quad (22)$$

$$\psi'_{n,\mu}(\gamma_{n,\mu}) = 0. \quad (23)$$

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 $\mathrm{conv}(\mathrm{supp}\,\mathbb{H}_\mu) = [0,1]$, the construction is defined for every $\mu \in (0,1)$. Let

$$M_{\mathbb{H}_\mu}(\gamma) = \int e^{\gamma(z-\mu)}\,d\mathbb{H}_\mu(z),$$

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

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

$$\Pi_n(d\mu) = \frac{L_n(\mu)\pi(d\mu)}{\int_0^1 L_n(u)\pi(du)}, \tag{24}$$

$$\mathbb{Q}_{n+1|n}(A \mid X_{1:n}) = \int_0^1 \widetilde{\mathbb{P}}_{n,\mu}(A)\,\Pi_n(d\mu). \tag{25}$$

The next proposition verifies the Wasserstein consistency condition needed for Theorem 4.1; its proof is given in Appendix A.6.

Proposition 5.2 (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 $\epsilon > 0$,*

$$\begin{aligned} \Pi_n\big(|\mu - \mu^*| > \epsilon\big) &\to 0, \\ \Pi_n\big(|\gamma_{n,\mu}| > \epsilon\big) &\to 0, \end{aligned} \tag{26}$$

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 4.1, 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 B.

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 (24) with $\tau_n = 0$, and RETEL (24) 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 (Waudby-Smith & Ramdas, 2024), PRECiSE-CO96 (Orabona & Jun, 2023), and the full Cover–Ordentlich universal portfolio (UP) method (Cover & Ordentlich, 1996). 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 Hedged-Capital, PRECiSE-CO96, and parametric Bayes are deferred to Appendix E. 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, all experiments use confidence level $\alpha = 0.1$, truncation parameter $c = 0.95$, and an equally spaced grid of 500 candidate means for test inversion. Method-specific details are provided in Appendix C.

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 E.1, while the non-informative and misspecified-prior regimes are considered in Appendix E.1.2. Detailed prior specifications are provided in Appendix D.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 E.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.

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 *best-arm identification*: each LLM is a bandit arm, and pulling an arm queries the model on a

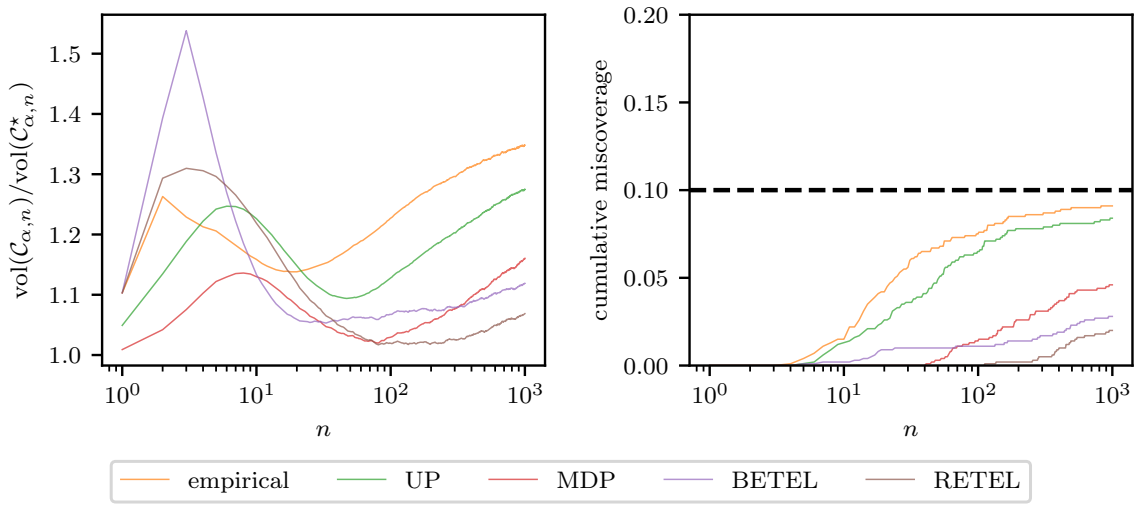

Figure 1. Synthetic beta-mixture experiment. The left and right panels show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 repetitions; the dashed line marks $\alpha = 0.1$.

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 (Kalyanakrishnan et al., 2012; Kaufmann & Kalyanakrishnan, 2013) 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 a sigmoid 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 D.2.2.

Table 1a reports average LUCB stopping time over 50 repetitions, paired ranks, and win/tie/loss counts against the empirical predictive baseline, the strongest prior-free method in this experiment. BETEL performs best in terms of both average calls and paired rank, stopping earlier than the empirical predictive in 35 of 50 paired repetitions. RETEL also improves on the empirical predictive more often than not, while MDP is close to empirical, suggesting that its beta-based prior component may be less suitable in this experiment. UP requires more calls on average than the empirical predictive. Full-baseline results and paired-comparison analyses are reported in Appendix E.2.

Prediction-powered inference. As a second real-data application, we consider prediction-powered inference (PPI) (Angelopoulos et al., 2023) 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 combine them by a union bound. As noted by Kilian et al. (2025), 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 D.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 D.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 D.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

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

(a) LLM best-arm identification. Avg calls is the LUCB stopping time.

Method	Avg calls (s.e.)	Avg rank (s.e.)	W/T/L
BETEL	85.28 (3.94)	1.78 (0.15)	35/2/13
RETEL	93.12 (3.65)	2.64 (0.21)	27/2/21
MDP	97.64 (4.30)	3.04 (0.13)	24/4/22
Emp.	98.48 (5.67)	3.56 (0.32)	–
UP	106.68 (4.71)	4.96 (0.15)	15/5/30

(b) ATTRIBUTEDQA PPI test. Avg labels is time to rejection.

Method	Avg labels (s.e.)	Avg rank (s.e.)	W/T/L
MDP	52.79 (3.65)	3.58 (0.16)	65/31/4
BETEL	54.60 (3.98)	3.95 (0.23)	63/13/24
RETEL	58.24 (3.12)	4.87 (0.25)	53/3/44
Emp.	65.17 (4.77)	4.98 (0.29)	–
UP	66.95 (3.61)	6.27 (0.18)	33/11/56

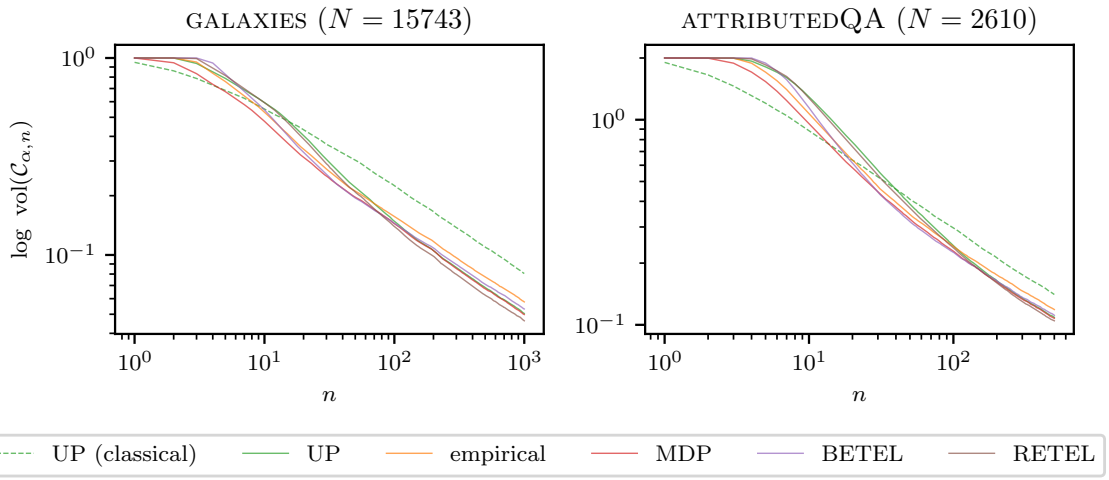

Figure 2. Prediction-powered mean estimation. The left and right panels show average CS length over 1000 repetitions for the GALAXIES and ATTRIBUTEDQA datasets. Curves labelled "classical" use only labelled observations, while the remaining curves use the PPI construction.

sample sizes in these experiments.

These width reductions translate into earlier anytime-valid decisions. On ATTRIBUTEDQA, testing $H_0 : \theta^\star \leq 0$ 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 E.2.3.

7. 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 C.3); the inverted sets are not guaranteed to be single intervals, which can complicate reporting or downstream use; and while poor priors do not invalidate coverage, the efficiency gains depend on the prior or auxiliary predictive being sufficiently informative (see Appendix E.1.2). These limitations are natural targets for future work on faster algorithms, interval-valued approximations, and adaptive or data-driven prior construction.

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382 (6671):669–674, 2023.

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

Audibert, J.-Y., Munos, R., and Szepesvári, C. Tuning Bandit Algorithms in Stochastic Environments. In Hutter, M., Servedio, R. A., and Takimoto, E. (eds.), *Algorithmic Learning Theory*, volume 4754, pp. 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.

Bennett, G. 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.

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

Bernshtein, S. N. *Theory of Probability*. Gostekhizdat, Moscow-Leningrad, 1927. In Russian, VIII+363 pages.

Bissiri, P. G., Holmes, C., and Walker, S. 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.

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

Bohnet, B., Tran, V. Q., Verga, P., Aharoni, R., Andor, D., Soares, L. B., Ciaramita, M., Eisenstein, J., Ganchev, K., Herzig, J., et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

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

Chen, J., Variyath, A. M., and Abraham, B. 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.

Chopin, N., Papaspiliopoulos, O., et al. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

Chugg, B. and Ramdas, A. Closed-form empirical bernstein confidence sequences for scalars and matrices, 2025.

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

Cortinovis, S. and Caron, F. Bayes-assisted Confidence Regions: Focal Point Estimator and Bounded-influence Priors. *arXiv preprint arXiv:2410.20169*, 2024.

Cortinovis, S. and Caron, F. FAB-PPI: Frequentist, Assisted by Bayes, Prediction-Powered Inference, August 2025.

Cortinovis, S., Kilian, V., and Caron, F. Confidence sequences with informative, bounded-influence priors, July 2025.

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

Cover, T. and Ordentlich, E. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363, 2002.

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

Csillag, D., Struchiner, C. J., and Goedert, G. T. Prediction-powered e-values. *arXiv preprint arXiv:2502.04294*, 2025.

Darling, D. A. and Robbins, H. 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.

Emerson, S. C. and Owen, A. B. 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.

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

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

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

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hoeffding, W. 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.

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

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

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

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), April 2021. ISSN 0090-5364. doi: 10.1214/20-AOS1991.

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

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

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. PAC Subset Selection in Stochastic Multi-armed Bandits. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 1, 2012.

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

Kaufmann, E. and Koolen, W. M. 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.

Kelly, J. L. J. 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.

Kilian, V., Cortinovis, S., and Caron, F. Anytime-valid, Bayes-assisted,Prediction-Powered Inference, May 2025.

Kim, E., MacEachern, S. N., and Peruggia, M. Regularized Exponentially Tilted Empirical Likelihood for Bayesian Inference, February 2024.

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

Kumon, M., Takemura, A., and Takeuchi, K. 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.

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

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

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

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

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

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

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

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

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

Ramdas, A. and Wang, R. Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*, 2024.

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

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Robbins, H. 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.

Robbins, H. and Siegmund, D. Boundary crossing probabilities for the Wiener process and sample sums. *The Annals of Mathematical Statistics*, pp. 1410–1429, 1970.

Santambrogio, F. *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.

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

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

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

Shekhar, S. and Ramdas, A. On the near-optimality of betting confidence sets for bounded means, 2023.

Ville, J. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, 1939.

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

Walker, S. G. 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.

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

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H., and Ramdas, A. Time-uniform central limit theory and asymptotic confidence sequences, March 2024.

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

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., 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.

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

Zhu, H., Zhou, H., Chen, J., Li, Y., Lieberman, J., and Styner, M. 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.

Technical Appendices and Supplementary Material

A. Proofs

A.1. Proof of Proposition 3.1

Proposition A.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 3.2

Proposition A.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 3.1, $(W_n(X_{1:n}, \mu^*))_{n\geq 0}$ is a nonnegative martingale with initial value one. Ville's inequality (Ville, 1939) 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 4.1

We first recall the theorem.

Theorem 4.1 (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

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

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

in probability, and

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

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 $\mathcal{W}_1$ (Santambrogio, 2015 Section1.2),

$$Z_n := \sup_{\lambda\in I_{\mu,c}} |M_n(\lambda) - M(\lambda)| \leq L_{\mu,c}\,\mathcal{W}_1\left(\mathbb{Q}_{n|n-1}(\cdot \mid X_{1:n-1}), \mathbb{P}^*\right). \tag{27}$$

Under the assumption of Theorem 4.1, $Z_n \to 0$ 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

$$Y_n := \log\{1 + \widehat{\lambda}_n(X_n - \mu)\}, \tag{28}$$

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 (Hall & Heyde, 1980 Theorem 2.18) gives

$$\frac{S_n}{n} \to 0 \qquad \text{a.s.} \tag{29}$$

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

$$\frac{1}{n}\log W_n = \frac{1}{n}\sum_{i=1}^{n} M(\widehat{\lambda}_i) + \frac{S_n}{n}. \tag{30}$$

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 (29) and (30) 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 (27) 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 (29), 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 4.2

We already proved that

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

We have

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

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 (28). It follows that

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

A.5. Proof of Proposition 5.1

Let

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

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

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

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

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

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 4.1 gives the final claim.

A.6. Proof of Proposition 5.2

We first recall the proposition.

Proposition A.3 (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 $\epsilon > 0$,*

$$\begin{aligned} \Pi_n\big(|\mu - \mu^*| > \epsilon\big) &\to 0, \\ \Pi_n\big(|\gamma_{n,\mu}| > \epsilon\big) &\to 0, \end{aligned} \tag{26}$$

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 A.4 (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 A.5 (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

$$|\psi'_{n,\mu}(0)| = |m_{n,\mu} - \mu| \le |\bar{X}_n - \mu| + \frac{\tau_n}{n+\tau_n}. \tag{31}$$

Moreover, by the variance decomposition for mixtures,

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

If $s_n^2 \ge v_0$, 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 $|\gamma| \le \delta$, let $\mathbb{P}^\gamma_{n,\mu}$ be the exponentially tilted version of $\mathbb{P}_{n,\mu}$. Since $z - \mu \in [-1,1]$,

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

Using

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

we obtain

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

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

Choose $\eta > 0$ such that $\eta < c_0\delta/2$. Since $\tau_n/n \to 0$, (31) implies that, uniformly over $|\mu - \bar{X}_n| \le \eta$,

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

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,

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

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

Lemma A.6 (Subexponential denominator lower bound)**.** *Let*

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

Under the assumptions of Proposition 5.2, with $\mathbb{P}^*$-probability tending to one,

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

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 $s_n^2 \ge \sigma^2/2$. Applying Lemma A.5 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 $\mu \in B_n$,

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

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

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

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

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

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

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

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 $\epsilon \le 1$. Let

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

On the event $|\bar{X}_n - \mu^*| \le \epsilon/2$, $A_\epsilon \subseteq \{\mu : |\mu - \bar{X}_n| \ge \epsilon/2\}$. Hence, by Lemma A.4,

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

Combining this exponential numerator bound with the subexponential lower bound from Lemma A.6 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 (26).

Next fix $\epsilon > 0$. Let η and C be as in Lemma A.5, and choose $\rho \in (0,\eta)$ such that $C\rho < \epsilon/2$. Since $\tau_n/n \to 0$, with probability tending to one,

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

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 (26), and the second term by the law of large numbers. Hence the second part of (26) follows.

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

Lemma A.7. *For every μ for which the fitted distribution is defined,*

$$\mathcal{W}_1(\widetilde{\mathbb{P}}_{n,\mu}, \mathbb{P}_n) \le 2|\gamma_{n,\mu}| + \frac{\tau_n}{n+\tau_n} + \frac{\tau_n}{n} e^{2|\gamma_{n,\mu}|}. \tag{32}$$

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 (32). □

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 $[0,1]$, (32) 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 (26), 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. 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 (23) 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 (23) 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 $\{0, 1\}$:

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

C. Implementation details

The methods proposed in this paper were implemented in Python 3.12. All experiments were run on an Apple Silicon M4 Pro with 24GB of RAM.

C.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 a grid size G, 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

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

which enforces nested intervals while preserving anytime-validity (see, e.g., Waudby-Smith & Ramdas, 2024).

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.

C.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,

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

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,

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

Posterior inference is approximated using the IBIS/waste-free SMC sampler from the `particles` package (Chopin et al., 2020). With N_{smc} particles and Markov chains of length L_{smc}, each time point stores

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

posterior samples. Conditional on these samples, the predictive score

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

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 (24) 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 (24) 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 B. 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.

C.3. Computational complexity

Let n denote the sequence length and let G be the number of candidate means used for test inversion. All CS methods based on (5), like the ones above, evaluate wealth processes over the $n \times G$ 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.

C.4. Default numerical parameters

Unless otherwise specified, all experiments use confidence level $\alpha = 0.1$, truncation parameter $c = 0.95$, and $G = 500$ grid points for test inversion. 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 $\kappa = 50$.

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 requires a prior distribution over the beta working model mean–concentration parameters ρ and ν in Equation (39). This usually takes the form

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

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

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

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

D. Experimental details

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

D.1. Synthetic experiments

D.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: $\rho \sim \text{Beta}(200\mu^\star, 200(1-\mu^\star)), \quad \nu \sim \text{Gamma}(a_\nu, b_\nu)$, where $(a_\nu, b_\nu) = (7.5, 1), (2, 0.1), (2, 1)$ in the three cases.

- Non-informative prior: $\rho \sim \mathrm{Beta}(1,1), \quad \nu \sim \mathrm{Gamma}(1.5,1)$.
- Misspecified prior: $\rho \sim \mathrm{Beta}(200\mu_{\text{bad}}, 200(1-\mu_{\text{bad}})), \quad \nu \sim \mathrm{Gamma}(a_\nu, b_\nu)$, where $\mu_{\text{bad}} = 0.9$ and $(a_\nu, b_\nu) = (7.5,1), (2,0.1), (2,1)$ 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 $0.25\,\mathrm{Beta}(5,15) + 0.75\,\mathrm{Beta}(15,5)$, which gives $\mu^\star = 0.625$. For Parametric/MDP, we set:

- Informative prior: $\rho \sim \mathrm{Beta}(500\mu^\star, 500(1-\mu^\star)), \quad \nu \sim \mathrm{Gamma}(2,2)$.
- Non-informative prior: $\rho \sim \mathrm{Beta}(1,1), \quad \nu \sim \mathrm{Gamma}(1.5,1)$.
- Misspecified prior: $\rho \sim \mathrm{Beta}(200\mu_{\text{bad}}, 200(1-\mu_{\text{bad}})), \quad \nu \sim \mathrm{Gamma}(2,2)$, where $\mu_{\text{bad}} = 0.1$.

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

D.2. Best-arm identification for LLM evaluation

Appendix D.2.1 provides a detailed description of the LUCB algorithm (Kalyanakrishnan et al., 2012; Kaufmann & Kalyanakrishnan, 2013), whereas Appendix D.2.2 details the implementation of the LUCB algorithm for best-arm identification in LLM evaluation, and explains our experimental setup.

D.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 (Kalyanakrishnan et al., 2012) and refined by (Kaufmann & Kalyanakrishnan, 2013). 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.
- $U_a(t)$, a $(1-\alpha)$-upper confidence bound for the true mean of a.
- $L_a(t)$, a $(1-\alpha)$-lower confidence bound 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 $(1-\alpha)$-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* $l_t = \arg\max_{a \notin J_t} U_a(t)$.
  $t \leftarrow t + 1$.
end while
return J_t

D.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` (https://huggingface.co/lmstudio-community/gemma-3-1B-it-qat-GGUF *license gemma*)
- **Qwen 3 1.7B:** `qwen/qwen3-1.7b`: (https://huggingface.co/lmstudio-community/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 0.4.11+1 locally on an Apple Silicon M4 Pro CPU with 24 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 $\alpha = 0.1$ 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 repetitions, 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.

D.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 ($n \in \{4, 6, 8, 10, 12, 14, 16, 18\}$) and fit a sigmoid function to extrapolate the predicted performance at $n = 20$, denoted by $\hat{\mu}_{\text{sigmoid}}$. With this, we then set the following priors for the Bayes-assisted CS methods:

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

D.3. Prediction-powered inference

D.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 (Angelopoulos et al., 2023). 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 (Kilian et al., 2025) or by building prediction-powered e-values in more general sequential settings (Csillag et al., 2025). 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. (2025), 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 $C^Z_{\alpha,n}$ denotes the resulting confidence sequence for $\mu_Z^\star$, then the corresponding confidence sequence for the rectifier is

$$C^\Delta_{\alpha,n} = \ell + (u-\ell)C^Z_{\alpha,n},$$

and the reported PPI confidence sequence for the original target is

$$C^\theta_{\alpha,n} = \widehat{m}_N + C^\Delta_{\alpha,n}.$$

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 $\widehat{m}_N$.

D.3.2. Datasets

Galaxies. The GALAXIES dataset contains a total of 16743 images from the Galaxy Zoo 2 initiative (Willett et al., 2013). 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 (He et al., 2016) 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 (Bohnet et al., 2022). 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 $Z_i = (R_i+2)/4 \in [0,1]$.

D.3.3. Prior specification

We choose the PPI priors to match the heuristic of Kilian et al. (2025), 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 C.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.

E. Additional experimental results

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

E.1. Synthetic data

E.1.1. Additional baselines

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.

E.1.2. Additional priors

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 D.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).

E.2. Best-arm identification for LLM evaluation

E.2.1. Additional baselines

Table 2 reports the full results for the best-arm identification experiment in Section 6. Overall, the results presented in the main text stand, with the robust 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, and the parametric Bayesian predictive slightly underperforms the empirical predictive, possibly suggesting that the beta working model may not be a good fit for this example.

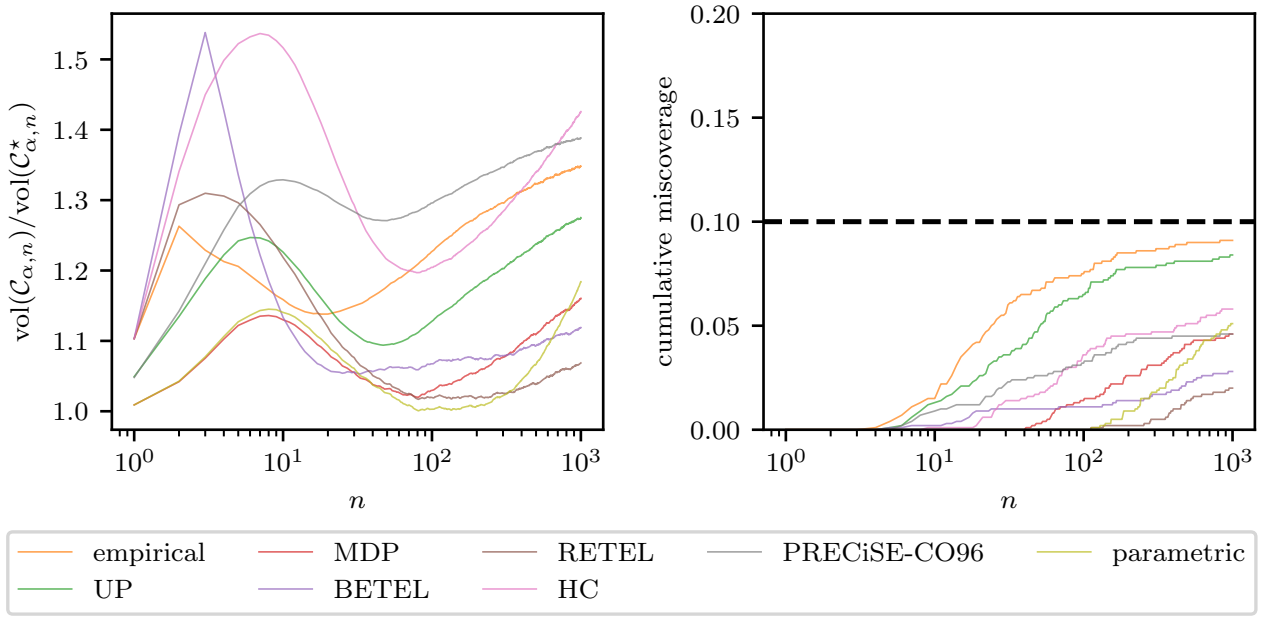

Figure 3. Full-baseline version of the synthetic beta-mixture experiment in Figure 1. The left and right panels show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 repetitions; the dashed line marks $\alpha = 0.1$.

Table 2. Full-baseline version of the LLM best-arm identification experiment with Spearman score in Table 1a.

Method	Avg Pulls (s.e.)	Avg rank (s.e.)	W/T/L vs Empirical
BETEL	85.28 (3.94)	1.78 (0.15)	35/2/13
RETEL	93.12 (3.65)	2.64 (0.21)	27/2/21
MDP	97.64 (4.30)	3.04 (0.13)	24/4/22
Empirical	98.48 (5.67)	3.56 (0.32)	–
Parametric	102.32 (4.63)	4.30 (0.20)	22/1/27
UP	106.68 (4.71)	4.96 (0.15)	15/5/30
PRECiSE-CO96	130.72 (5.45)	7.08 (0.10)	2/1/47
HC	137.16 (4.89)	7.58 (0.09)	3/1/46

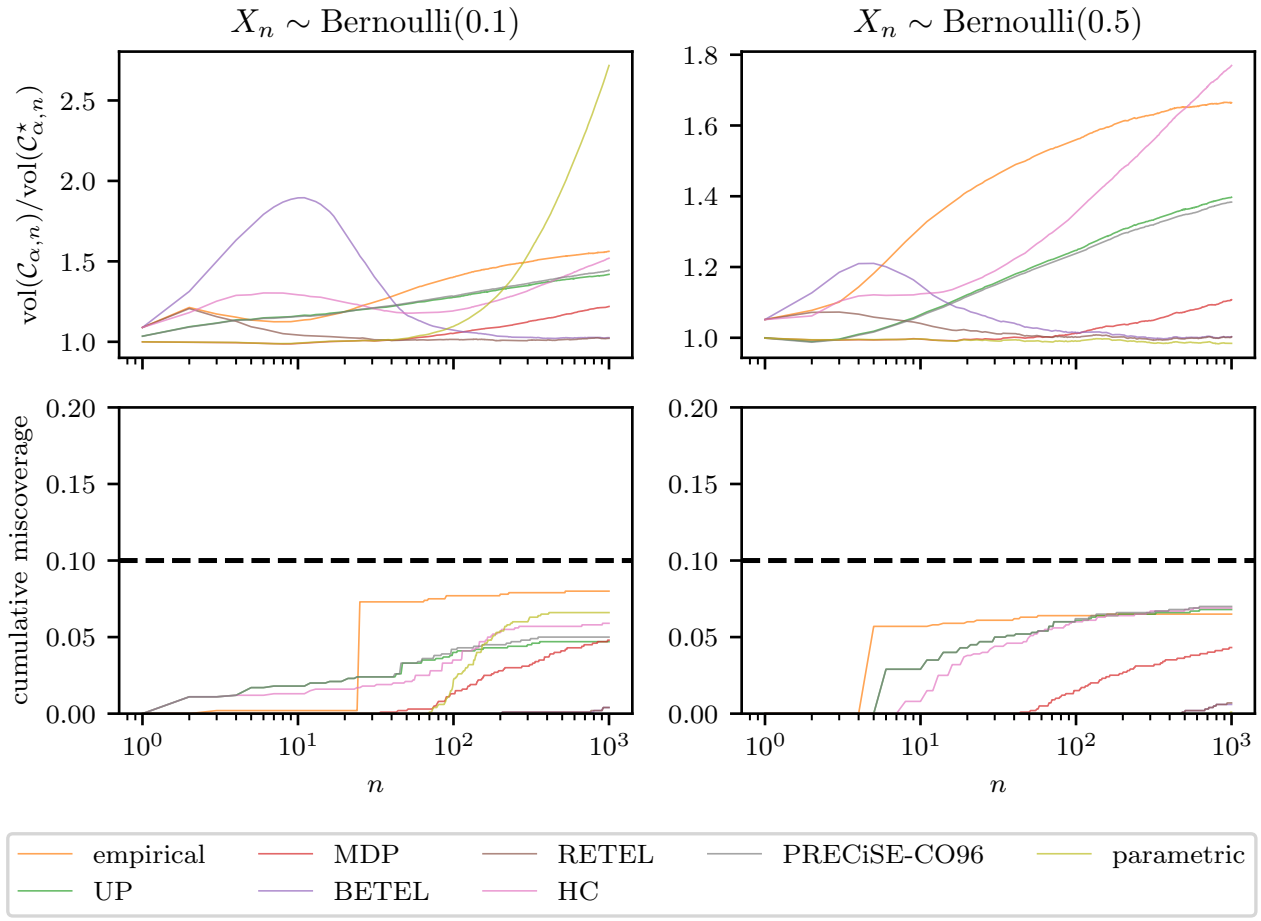

Figure 4. Synthetic Bernoulli experiments with all baselines. The left and right columns correspond to $X_n \sim \mathrm{Bernoulli}(0.1)$ and $X_n \sim \mathrm{Bernoulli}(0.5)$, respectively. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 repetitions; the dashed line marks $\alpha = 0.1$.

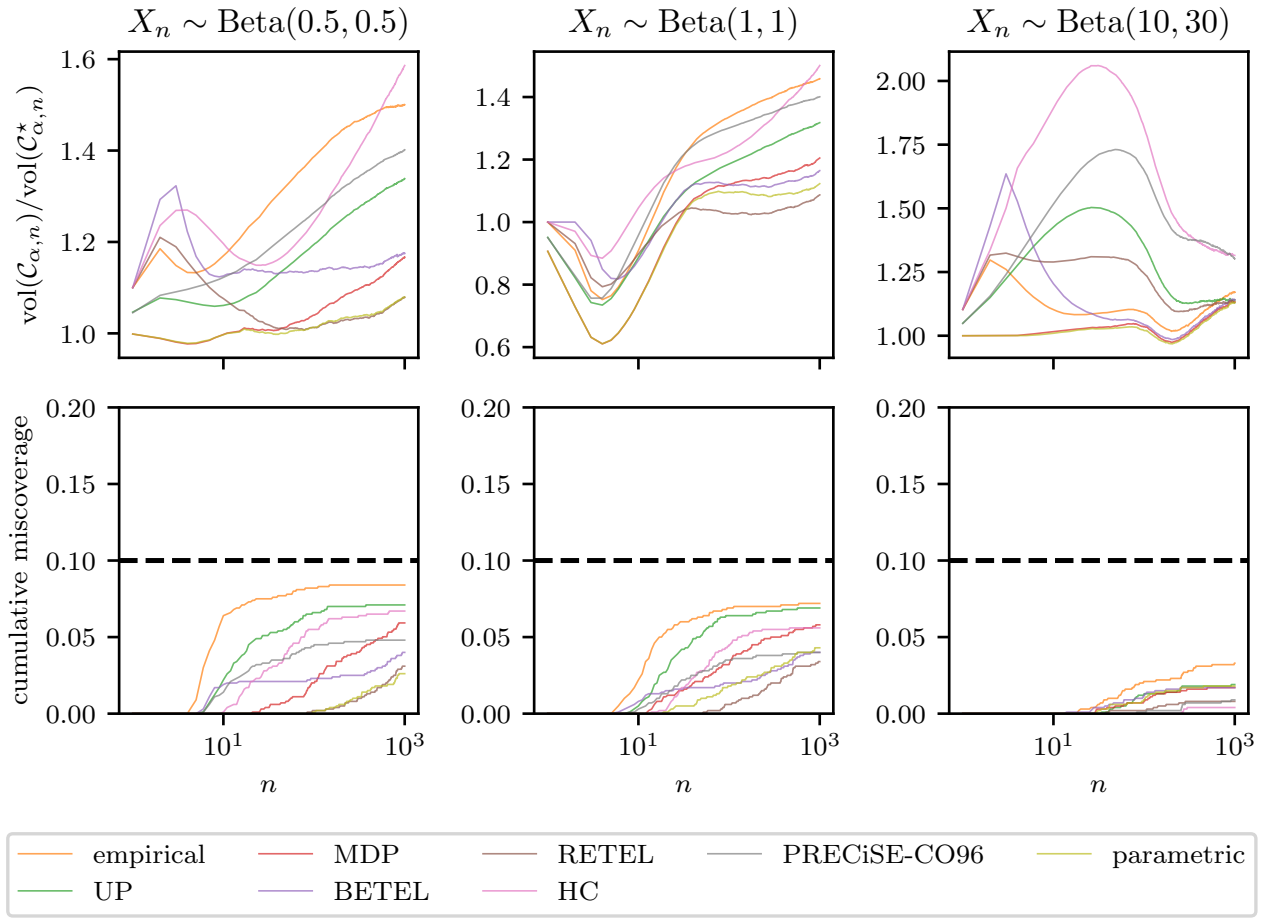

Figure 5. Synthetic beta experiments with all baselines. The left, centre, and right columns correspond to $X_n \sim \text{Beta}(0.5, 0.5)$, $X_n \sim \text{Beta}(1, 1)$, and $X_n \sim \text{Beta}(10, 30)$, respectively. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 repetitions; the dashed line marks $\alpha = 0.1$.

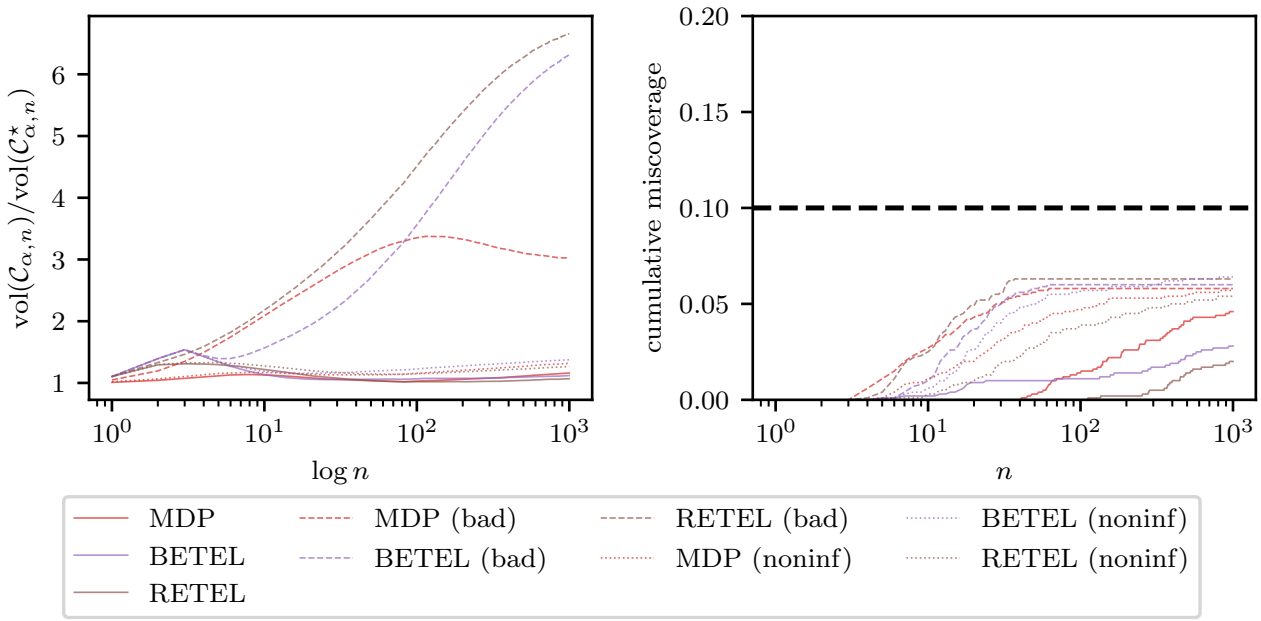

Figure 6. Synthetic beta-mixture experiment with different prior specifications for the same setting as Figures 1 and 3. The left and right panels show average CS length to the oracle and empirical cumulative miscoverage, respectively, over 1000 repetitions. Solid, dotted, and faint solid curves correspond to informative, misspecified, and non-informative priors, respectively; the dashed horizontal line marks $\alpha = 0.1$.

E.2.2. Statistical significance of the best-arm identification results

To establish the statistical significance of the performance disparities among the evaluated methods in Tables 1a and 2, 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 (Wilcoxon, 1945) 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 (Holm, 1979), 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 3.

Table 3. Pairwise statistical comparison for the Spearman score: Is Method A (row) significantly faster than Method B (column)?

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

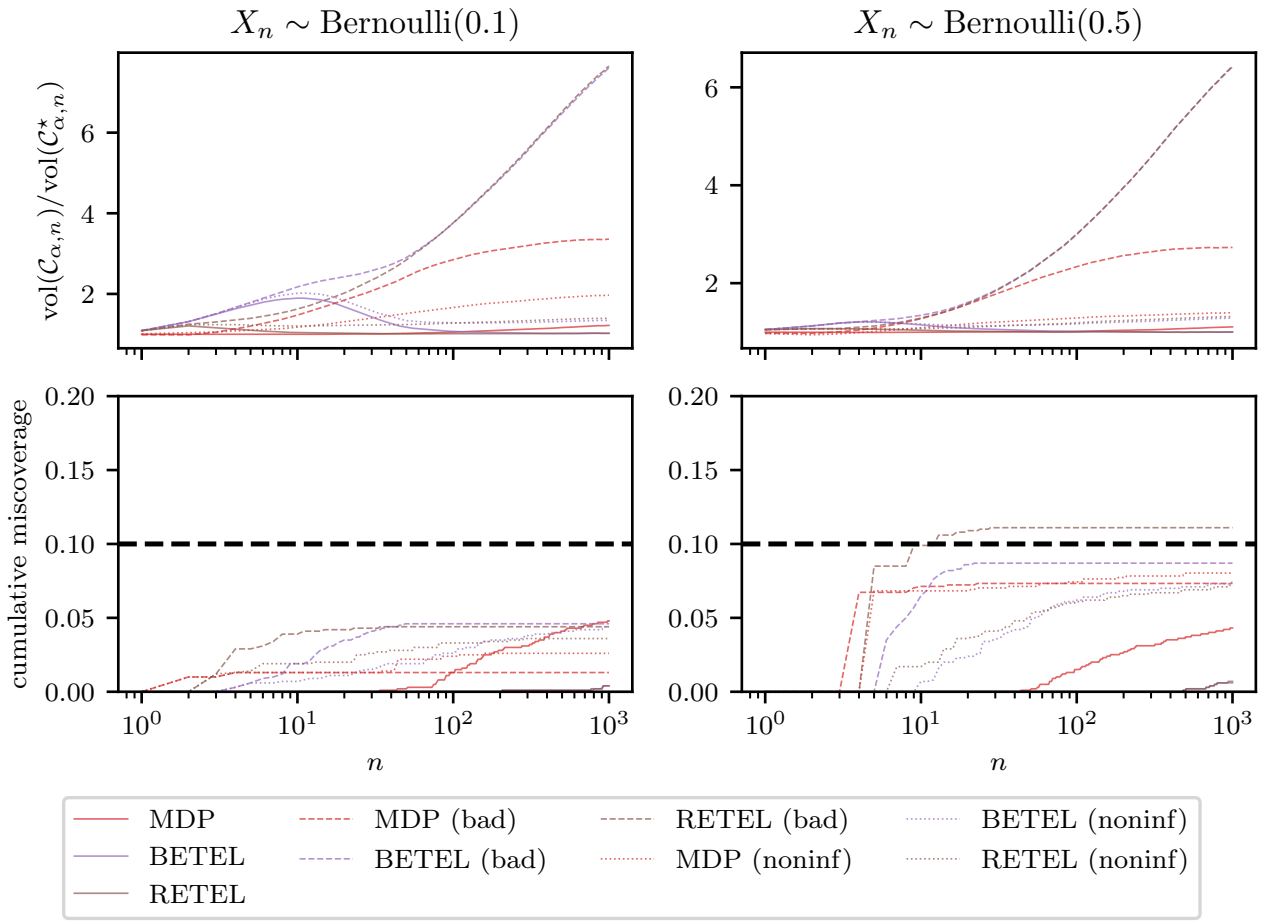

Figure 7. Synthetic Bernoulli experiment with different prior specifications for the same setting as in Figure 4. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 repetitions. Solid, dotted, and faint solid curves correspond to informative, misspecified, and non-informative priors, respectively; the dashed horizontal line marks $\alpha = 0.1$.

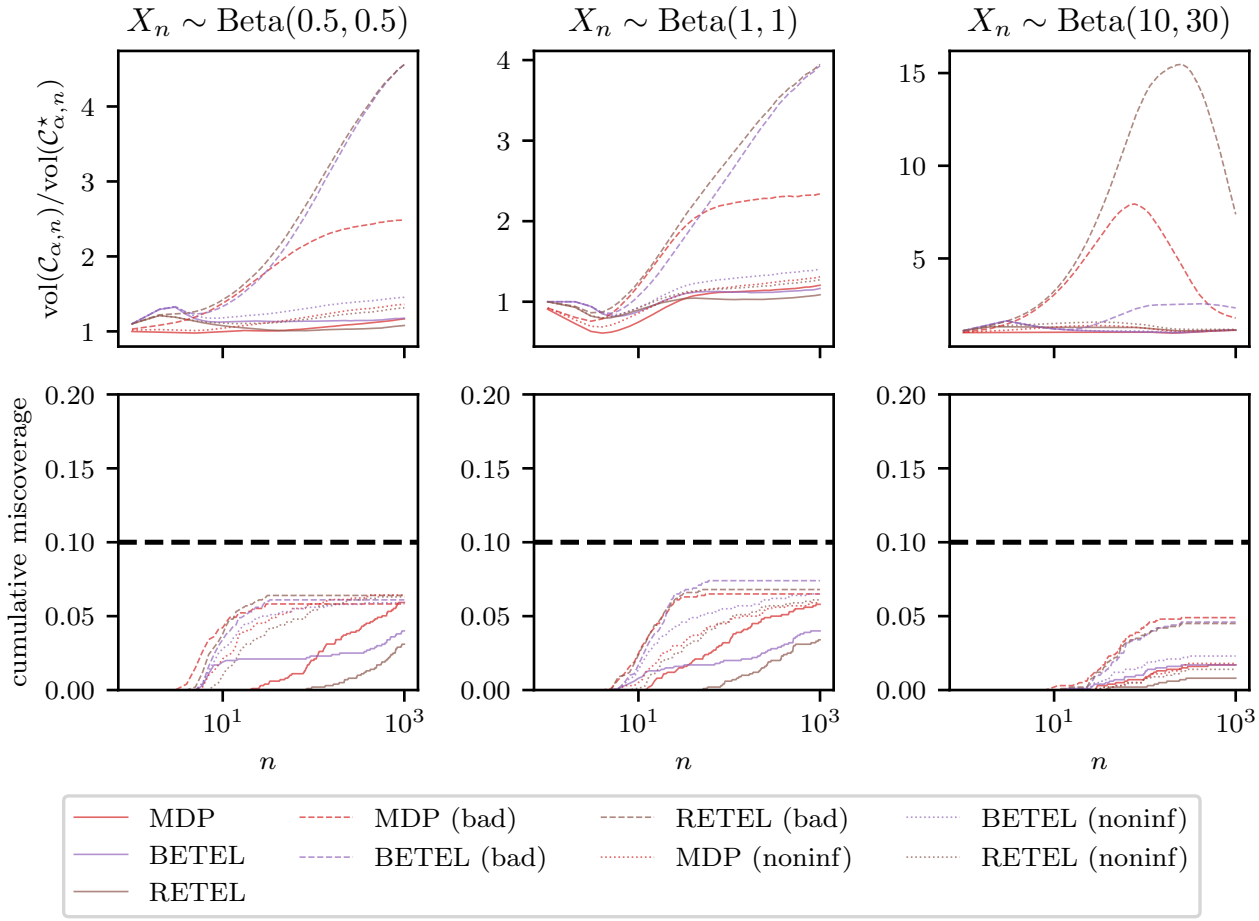

Figure 8. Synthetic beta experiment with different prior specifications for the same setting as in Figure 5. The top and bottom rows show average CS length relative to the oracle and empirical cumulative miscoverage, respectively, over 1000 repetitions. Solid, dotted, and faint solid curves correspond to informative, misspecified, and non-informative priors, respectively; the dashed horizontal line marks $\alpha = 0.1$.

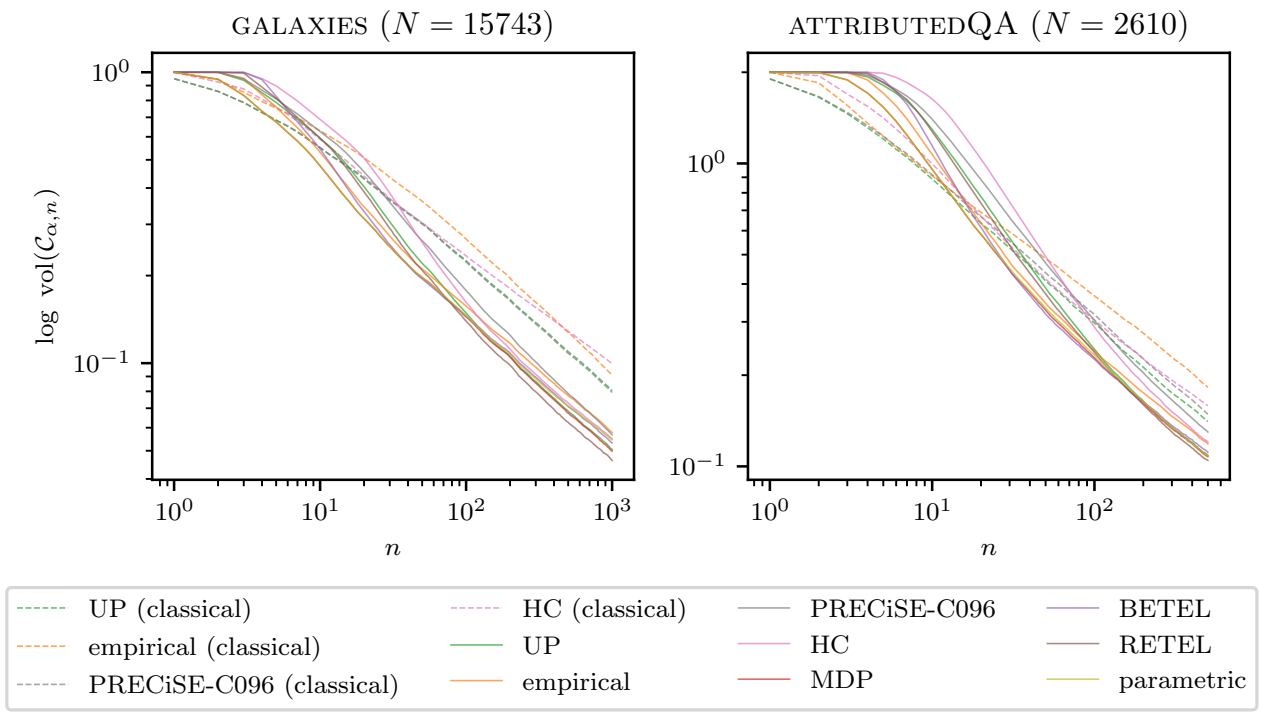

Figure 9. Full-baseline version of prediction-powered mean estimation experiment in Figure 2. The left and right panels show average CS length over 1000 repetitions for the GALAXIES and ATTRIBUTEDQA datasets. Curves labelled "classical" use only labelled observations, while the remaining curves use the PPI construction.

E.2.3. PREDICTION-POWERED INFERENCE

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 4 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 4. 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 repetitions. Avg rank ranks methods by stopping time within each repetition. W/T/L gives paired wins, ties, and losses against the PPI empirical predictive baseline.

Method	Avg labels (s.e.)	Avg rank (s.e.)	W/T/L vs Empirical
Parametric	52.23 (3.46)	2.62 (0.15)	64 / 29 / 7
MDP	52.79 (3.65)	3.58 (0.16)	65 / 31 / 4
BETEL	54.60 (3.98)	3.95 (0.23)	63 / 13 / 24
RETEL	58.24 (3.12)	4.87 (0.25)	53 / 3 / 44
Empirical	65.17 (4.77)	4.98 (0.29)	–
UP	66.95 (3.61)	6.27 (0.18)	33 / 11 / 56
HC	84.85 (3.71)	9.04 (0.20)	14 / 9 / 77
PRECISE-CO96	84.99 (4.34)	9.01 (0.18)	8 / 3 / 89
UP (classical)	91.44 (7.01)	7.01 (0.32)	25 / 3 / 72
HC (classical)	97.91 (8.19)	7.61 (0.28)	23 / 2 / 75
PRECISE-CO96 (classical)	100.39 (7.25)	8.28 (0.32)	17 / 2 / 81
Empirical (classical)	182.69 (11.38)	10.78 (0.31)	12 / 0 / 88