Anytime-valid, Bayes-assisted, Prediction-Powered Inference

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

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

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

Abstract

Given a large pool of unlabelled data and a smaller amount of labels, prediction-powered inference (PPI) leverages machine learning predictions to increase the statistical efficiency of confidence interval procedures based solely on labelled data, while preserving fixed-time validity. In this paper, we extend the PPI framework to the sequential setting, where labelled and unlabelled datasets grow over time. Exploiting Ville's inequality and the method of mixtures, we propose prediction-powered confidence sequence procedures that are asymptotically valid uniformly over time and naturally accommodate prior knowledge on the quality of the predictions to further boost efficiency. We carefully illustrate the design choices behind our method and demonstrate its effectiveness in real and synthetic examples.

1 Introduction

Increasing the sample size of an experiment is arguably the single simplest way to improve the precision of the statistical conclusions drawn from it. However, in many fields – such as healthcare, finance, and social sciences – obtaining labelled data is often costly and time-consuming. In these settings, using machine learning (ML) models to impute additional labels represents a tempting alternative to expensive data collection, albeit at the risk of introducing bias. Prediction-powered inference (PPI) [1] is a recently introduced framework for valid statistical inference in the presence of a small labelled dataset and a large number of unlabelled examples paired with predictions from a black-box model.

Formally, given an input/output pair $(X, Y) \sim \mathbb{P} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$, consider the goal of estimating

$$\theta^\star = \underset{\theta \in \mathbb{R}}{\arg\min} \ \mathbb{E}[\ell_\theta(X, Y)], \tag{1}$$

where $\ell_\theta(x, y)$ is a convex loss function parameterised by $\theta \in \mathbb{R}$. As an example, the mean $\theta^\star = \mathbb{E}[Y]$ is the estimand induced by the squared loss $\ell_\theta(x, y) = (\theta - y)^2/2$. For $t = 1, 2, \ldots$, we observe a sequence of independent random variables Z_t, either drawn from $\mathbb{P}$ (labelled sample) or from $\mathbb{P}_X$ (unlabelled sample), and we are provided with a black-box prediction rule f that maps any input x to a prediction $f(x)$.

Let $(X_i, Y_i)_{i \geq 1}$ and $(\widetilde{X}_j)_{j \geq 1}$ denote the subsequence of labelled and unlabelled samples, respectively. For $n = 1, 2, \ldots$, let N_n denote the number of unlabelled samples observed before the nth labelled one, and assume that $N_n \geq n$, with $N_n \gg n$ in typical settings. PPI constructs an (asymptotic) $1 - \alpha$ confidence interval (CI) $\mathcal{C}^{\text{pp}}_{\alpha,n}$ for $\theta^\star$, that exploits the auxiliary information encoded in f. To this end, under mild assumptions, $\theta^\star$ can be expressed as the solution to

$$g_{\theta^\star} := \mathbb{E}[\ell'_{\theta^\star}(X, Y)] = 0, \tag{2}$$

[*]Equal contribution. Order decided by coin toss.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

where ℓ'_θ is a subgradient of ℓ_θ with respect to θ. The quantity g_θ in Equation (2) can be decomposed as $g_\theta = m_\theta + \Delta_\theta$, where

$$m_\theta := \mathbb{E}[\ell'_\theta(X, f(X))] \quad \text{and} \quad \Delta_\theta := \mathbb{E}[\ell'_\theta(X, Y) - \ell'_\theta(X, f(X))], \tag{3}$$

where m_θ represents a measure of fit of the predictor, while Δ_θ, the *rectifier*, accounts for the discrepancy between the predicted outputs $f(X)$ and the true labels Y. If $\mathcal{C}^g_{\alpha,\theta,n}$ is a $(1-\alpha)$ confidence interval for g_θ, then the PPI confidence interval $\mathcal{C}^{\text{pp}}_{\alpha,n}$, defined as

$$\mathcal{C}^{\text{pp}}_{\alpha,n} = \left\{\theta \mid 0 \in \mathcal{C}^g_{\alpha,\theta,n}\right\}, \tag{4}$$

also achieves the desired coverage, i.e. $\Pr(\theta^\star \in \mathcal{C}^{\text{pp}}_{\alpha,n}) \geq 1-\alpha$. Constructing $\mathcal{C}^g_{\alpha,\theta,n}$ naturally relies on estimating g_θ, for which PPI defines an estimator that leverages both the unlabelled data and the prediction rule f. The resulting method outperforms standard CI procedures based on the labelled data alone when f is sufficiently accurate and $N_n \gg n$. Intuitively, this is because, in this case, Δ_θ is close to zero, while m_θ can be estimated with low variance from the unlabelled data.

Crucially, coverage of the PPI CI (4) is guaranteed only at a fixed time, i.e., for a labelled sample size n fixed in advance. This is undesirable in many practical settings – such as online learning, real-time monitoring, or sequential decision-making – where it is essential to continuously draw conclusions as new data arrive. In this work, we address this by proposing an *anytime-valid* extension of the PPI CI (4). That is, we define a confidence sequence $(\mathcal{C}^{\text{avpp}}_{\alpha,n})_{n\geq 1}$, satisfying asymptotically the stronger coverage guarantee

$$\Pr(\theta^\star \in \mathcal{C}^{\text{avpp}}_{\alpha,n} \text{ for all } n \geq 1) \geq 1-\alpha,$$

while still taking advantage of the prediction rule f. Analogously to standard PPI, we construct a confidence sequence $(\mathcal{C}^g_{\alpha,\theta,n})_{n\geq 1}$ for g_θ and define $\mathcal{C}^{\text{avpp}}_{\alpha,n}$ through Equation (4) for $n \geq 1$. While our approach is agnostic to the specific form of the confidence sequence $(\mathcal{C}^g_{\alpha,\theta,n})_{n\geq 1}$, we mainly focus on asymptotic confidence sequences [15], as they provide a versatile time-uniform analogue of standard CLT-based CIs that applies to the PPI framework above in full generality. Moreover, being based on the method of mixtures [9, 4, 8], they can readily accommodate prior information on the quality of the prediction model f. In particular, by means of a zero-centred prior on the rectifier Δ_θ, we obtain tighter confidence sequences when the predictions are good, extending the fixed-time Bayes-assisted approach of Cortinovis and Caron [6].

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 provides background on (asymptotic) confidence sequences and discusses how prior information may be incorporated into their construction. Section 4 presents PPI in the context of control-variate estimators, whose asymptotic properties are crucial for our approach to anytime-valid, Bayes-assisted PPI, which is described in Section 5. Section 6 demonstrates the benefits of our method on synthetic and real data. Finally, Section 7 discusses limitations and further extensions of our approach. Proofs and additional experiments are provided in the Supplementary Material.

2 Related Work

PPI was introduced by Angelopoulos et al. [1] as a general framework for valid statistical inference with black-box machine learning predictors, and was later extended in Angelopoulos et al. [7]. Closely related ideas appear in the literatures on semi-supervised inference, missing-data methods, survey sampling, and double machine learning [8, 9, 10, 11, 12]. More recently, Cortinovis and Caron [6] proposed a Bayes-assisted variant of PPI. All of these contributions target fixed-time confidence intervals.

Confidence sequences were first introduced by Darling and Robbins [13] and developed further by Robbins and Siegmund [4] and Lai [8], building on earlier work by Ville [9] and Wald [11]. Interest has surged again in recent years [15, 15], motivated by applications such as A/B testing. The notion is closely linked to e-values [17, 17]. Building on the e-value framework, and on earlier work by Zrnic and Candès [18] and Waudby-Smith and Ramdas [15], Csillag et al. [19] proposed an exact, time-uniform PPI method that yields confidence sequences under stronger conditions (e.g., existence of bounded e-values) and does not leverage prior knowledge about the ML prediction quality. Furthermore, application of their method requires an active-learning setup in which, for each t, the observation Z_t can be labelled with strictly positive probability. In particular, it is not applicable to deterministic sequences of observations, such as those describing a large initial pool of unlabelled data followed by a stream of labelled data – the main focus of our experiments.

In the setting of double machine learning and semiparametric inference, Dalal et al. [20] and Waudby-Smith et al. [15] derive asymptotic confidence sequences for target parameters in the presence of high-dimensional nuisance components.

3 Asymptotic (Bayes-assisted) confidence sequences

In this section, we begin with background on (asymptotic) confidence sequences (CS). We then show how prior information can be incorporated into asymptotic CS procedures, leading to asymptotic Bayes-assisted confidence sequences.

3.1 Background

We start by defining an exact confidence sequence [13], a time-uniform analogue of classical CIs.

Definition 1 (Confidence sequence)**.** *Let $(\mathcal{C}_{\alpha,t})_{t\geq 1}$ be a sequence of random subsets of $\mathbb{R}$. For $\alpha \in (0,1)$, $(\mathcal{C}_{\alpha,t})_{t\geq 1}$ is a $1-\alpha$ confidence sequence for a fixed parameter $\mu \in \mathbb{R}$ if*

$$\Pr(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq 1) \geq 1-\alpha. \tag{5}$$

We now introduce the notion of an asymptotic confidence sequence (AsympCS) [15, 20].

Definition 2 (Asymptotic confidence sequence)**.** *Let $\alpha \in (0,1)$ and $(a_t)_{t\geq 1}$ be a real sequence such that $\lim_{t\to\infty} a_t = 0$. Let $(\widehat{\mu}_t)_{t\geq 1}$ be a consistent sequence of estimators of μ. The sequence of random intervals $(\mathcal{C}_{\alpha,t})_{t\geq 1}$, with $\mathcal{C}_{\alpha,t} = [\widehat{\mu}_t - L_t, \widehat{\mu}_t + U_t]$ and $L_t > 0$, $U_t > 0$, is said to be an asymptotic confidence sequence with (little-o) approximation rate a_t if there exists a (usually unknown) confidence sequence $(\mathcal{C}^\star_{\alpha,t})_{t\geq 1}$, with $\mathcal{C}^\star_{\alpha,t} = [\widehat{\mu}_t - L^\star_t, \widehat{\mu}_t + U^\star_t]$, such that*

$$\Pr(\mu \in \mathcal{C}^\star_{\alpha,t} \text{ for all } t \geq 1) \geq 1-\alpha$$

and, almost surely as $t \to \infty$, $\max\{L^\star_t - L_t, U^\star_t - U_t\} = o(a_t)$.

Thus, an asymptotic CS may be regarded as an approximation of an exact CS that becomes arbitrarily accurate in the limit. It is worth noting that, while classical fixed-sample asymptotic CIs rely on *convergence in distribution* of the scaled estimators, asymptotic confidence sequences rely on the *almost sure convergence at a given rate* of the centred lower and upper bounds relative to those of an underlying exact CS. The following is an example of an asymptotic CS that applies to i.i.d. data.

Theorem 1. *Let $(Y_t)_{t\geq 1}$ be a sequence of i.i.d. random variables with mean μ and such that $\mathbb{E}|Y_1|^{2+\delta} < \infty$ for some $\delta > 0$. For any $t \geq 1$, let $\overline{Y}_t$ be the sample mean, and $\widehat{\sigma}^2_t$ be the sample variance based on the first t observations. For any parameter $\rho > 0$, the sequence of intervals defined as*

$$\mathcal{C}^{\mathtt{NA}}_{\alpha,t}(\overline{Y}_t, \widehat{\sigma}_t; \rho) := \left[\overline{Y}_t \pm \frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\left(1+\frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)}\right] \tag{6}$$

forms a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t\log t}$ for μ.

For the sequel, it is useful to highlight some aspects of the proof of this theorem. First, if the random variables $(Y_t)_{t\geq 1}$ were Gaussian with variance σ^2, then $\mathcal{C}^{\mathtt{NA}}_{\alpha,t}(\overline{Y}_t, \sigma; \rho)$ would be an exact CS. This follows from combining the method of mixtures for nonnegative martingales with Ville's inequality [9, 4, 8, 14]. Second, the proof relies on KMT strong coupling [2, 3]: there exists i.i.d. Gaussian random variables $(W_t)_{t\geq 1}$ with mean μ and variance $\mathrm{var}(Y)$ such that

$$\frac{1}{t}\sum_{i=1}^t Y_i = \frac{1}{t}\sum_{i=1}^t W_i + o\left(\frac{1}{\sqrt{t\log t}}\right) \text{ a.s. as } t \to \infty.$$

Such a coupling plays a central role in constructing asymptotic confidence sequences, serving as a substitute for the CLT assumption underlying classical fixed-sample CIs. The construction in Theorem 1 extends beyond the i.i.d. case, provided a similar coupling exists.

Theorem 2. *Let $(\widehat{\mu}_t)_{t\geq 1}$ be a consistent sequence of estimators of μ. Assume that there exists a sequence of i.i.d. Gaussian random variables $(W_i)_{i\geq 1}$, with mean μ and variance σ^2, such that*

$$\widehat{\mu}_t = \frac{1}{t}\sum_{i=1}^t W_i + o\left(\frac{1}{\sqrt{t\log t}}\right) \text{ a.s. as } t \to \infty. \tag{7}$$

Let $(\widehat{\sigma}_t^2)_{t\geq 1}$ be a consistent sequence of estimators of σ^2 with $|\widehat{\sigma}_t - \sigma| = o\left(\frac{1}{\log t}\right)$ a.s. Then, for any parameter $\rho > 0$, the sequence of intervals $(\mathcal{C}^{\texttt{NA}}_{\alpha,t}(\widehat{\mu}_t, \widehat{\sigma}_t; \rho))_{t\geq 1}$ forms a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t \log t}$ for μ.

The asymptotic CS (6) includes a tuning parameter ρ, which can be chosen so as to minimise the width of the interval at a specified time t; see [15, Appendix B.2]. However, this method does not allow the incorporation of prior information about the parameter of interest to yield tighter intervals when the data align with those assumptions: the width of Equation (6) is indeed independent of $\overline{Y}_t$.

3.2 Asymptotic Bayes-assisted confidence sequences

To address this, we introduce a Bayes-assisted analogue of Theorem 1.

Theorem 3 (Bayes-assisted AsympCS – i.i.d. case)**.** *Let $(Y_t)_{t\geq 1}$ be a sequence of i.i.d. random variables with unknown mean μ and unknown variance σ^2, and such that $\mathbb{E}|Y_1|^{2+\delta} < \infty$ for some $\delta > 0$. For any $t \geq 1$, let $\overline{Y}_t$ be the sample mean, and $\widehat{\sigma}_t^2$ be the sample variance based on the first t observations. Let $\eta_t : \mathbb{R} \to (0, \sqrt{t/(2\pi)})$ be defined as*

$$\eta_t(z) = \int_{-\infty}^{\infty} \mathcal{N}\left(z; \zeta, 1/t\right) \pi(\zeta) d\zeta. \tag{8}$$

where π is a continuous and proper prior density on $\mathbb{R}$, strictly positive in a neighbourhood of μ/σ. Then

$$\mathcal{C}^{\texttt{BA}}_{\alpha,t}(\overline{Y}_t, \widehat{\sigma}_t; \pi) := \left[\overline{Y}_t \pm \frac{\widehat{\sigma}_t}{\sqrt{t}} \sqrt{\log\left(\frac{t}{2\pi\alpha^2 \eta_t(\overline{Y}_t/\widehat{\sigma}_t)^2}\right)}\right] \tag{9}$$

forms a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t \log t}$ for μ.

In Theorem 3, the density π encodes prior beliefs about the ratio μ/σ. Under this prior, η_t represents the marginal density of the standardised mean $\overline{Y}_t/\sigma$ that would arise if the observations $(Y_t)_{t\geq 1}$ were normally distributed. In contrast to the non-assisted AsympCS (6), the width of the Bayes-assisted AsympCS (9) varies with $\overline{Y}_t/\widehat{\sigma}_t$: when the data align with the prior, $\eta_t(\overline{Y}_t/\widehat{\sigma}_t)$ is large and the interval narrows; when they conflict, $\eta_t(\overline{Y}_t/\widehat{\sigma}_t)$ is small and the interval widens. It is worth emphasising that, even when the prior is strongly misspecified, the Bayes-assisted AsympCS (9) remains valid. In the case of a Gaussian prior π centred at μ_0 with variance τ^2, we obtain the following AsympCS :

$$\mathcal{C}^{\texttt{BA}}_{\alpha,t}(\overline{Y}_t, \widehat{\sigma}_t; \mathcal{N}(\cdot; \mu_0, \tau^2)) = \left[\overline{Y}_t \pm \frac{\widehat{\sigma}_t}{\sqrt{t}} \sqrt{\log\left(\frac{t\tau^2+1}{\alpha^2}\right) + \frac{(\overline{Y}_t/\widehat{\sigma}_t - \mu_0)^2}{\tau^2 + 1/t}}\right]. \tag{10}$$

Setting $\rho = \tau$ allows a direct comparison between (10) and its non-assisted counterpart (6). When the data agree with the prior – i.e., $\overline{Y}_t/\widehat{\sigma}_t - \mu_0 \simeq 0$ – the Bayes-assisted interval is narrower than the non-assisted one. Conversely, if the data conflict with the prior, $(\overline{Y}_t/\widehat{\sigma}_t - \mu_0)^2$ is large and the Bayes-assisted AsympCS becomes wider than (6).

The proof of Theorem 3 is similar to that of [15, Theorem 2.2]. First, note that $\mathcal{C}^{\texttt{BA}}_{\alpha,t}(\overline{Y}_t, \mathrm{var}(Y); \pi)$ would be an exact CS if the observations were normally distributed. This follows from an application of the method of mixtures for nonnegative martingales, using the prior π as mixing density, together with Ville's inequality. Second, we use KMT strong coupling to approximate in an almost sure sense $\overline{Y}_t$ by a sample average of i.i.d. Gaussian random variables. As in the non-assisted case, Theorem 3 can be extended to the non-i.i.d. setting, as long as one can find such a strong coupling.

Theorem 4 (Asymptotic Bayes-assisted CS – non-i.i.d. case)**.** *Consider the same notation and assumptions as in Theorem 2. Let π be a continuous and proper prior density on $\mathbb{R}$, strictly positive in a neighbourhood of μ/σ, and let η_t be the density* (8) *for any $t \geq 1$. Then, the sequence of intervals $(\mathcal{C}^{\texttt{BA}}_{\alpha,t}(\widehat{\mu}_t, \widehat{\sigma}_t; \pi))_{t\geq 1}$ forms a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t \log t}$ for μ.*

3.3 Asymptotic Type-I error control

The asymptotic confidence sequences defined above satisfy an asymptotic version of time-uniform Type-I error control (in the sense of [15, §2.5]; see also [24]).

Theorem 5 (Asymptotic Type-I error control)**.** *Assume the hypotheses of one of Theorems 1 to 4, and let* $(\mathcal{C}_{\alpha,t})$ *be the corresponding* $(1-\alpha)$*–AsympCS for* μ*. Then*

$$\liminf_{m\to\infty} \Pr\left(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq m\right) \geq 1-\alpha. \tag{11}$$

4 Control variates and PPI: background and strong coupling

Prediction-powered inference (PPI) closely relates to control variates, a standard variance-reduction method in Monte Carlo estimation [25, §4.1]. In fact, each PPI estimator can be expressed as a control-variate estimator. We begin with a review of control variates and derive a KMT-type strong-coupling result for these estimators, and then provide additional background on PPI.

4.1 Control variates: definitions and KMT strong coupling

Let (U, V) be real-valued random variables with finite variance, and consider the goal of estimating $\gamma = \mathbb{E}[V]$ from an i.i.d. sample $(U_i, V_i)_{i=1}^n$. If $\mu = \mathbb{E}[U]$ is known, the control-variate estimator (CVE) of γ is defined as

$$\widehat{\gamma}_\lambda^{\mathrm{icv}} = \overline{V} - \lambda(\overline{U} - \mu) = \frac{1}{n}\sum_{i=1}^n \left(V_i - \lambda(U_i - \mu)\right), \tag{12}$$

where $\overline{U}$ and $\overline{V}$ denote the empirical means of $(U_i)_{i=1}^n$ and $(V_i)_{i=1}^n$, respectively, $\lambda \in \mathbb{R}$ is a tunable coefficient, and the term $U_i - \mu$ acts as a control variate. The estimator $\widehat{\gamma}_\lambda^{\mathrm{icv}}$ is unbiased, consistent, and has variance $\mathrm{var}(\widehat{\gamma}_\lambda^{\mathrm{icv}}) = (\mathrm{var}(V) - 2\lambda\mathrm{cov}(U,V) + \lambda^2\mathrm{var}(U))/n$. Compared to the standard sample mean estimator $\overline{V}$, which attains variance $\mathrm{var}(\overline{V}) = \mathrm{var}(V)/n$, using $\widehat{\gamma}_\lambda^{\mathrm{icv}}$ results in variance reduction when $\lambda < 2\mathrm{cov}(U,V)/\mathrm{var}(U)$. The minimum variance is achieved at the optimal coefficient $\lambda^\star = \mathrm{cov}(U,V)/\mathrm{var}(U)$, for which $\mathrm{var}(\widehat{\gamma}_{\lambda^\star}^{\mathrm{icv}}) = (1-\rho_{U,V}^2)\mathrm{var}(\overline{V})$, where $\rho_{U,V}$ is the correlation between U and V. That is, stronger correlation leads to greater variance reduction.

In practice, both μ and $\lambda^\star$ are typically unknown. When this is the case, given an additional i.i.d. sample $(\widetilde{U}_j)_{j=1}^{N_n}$, independent of $(U_i, V_i)_{i=1}^n$, where $\widetilde{U}_1$ has the same distribution as U, one can estimate μ by $\widehat{\mu} = \frac{1}{N_n}\sum_{j=1}^{N_n}\widetilde{U}_j$ and plug it into Equation (12). For fixed λ, this gives

$$\widehat{\gamma}_\lambda^{\mathrm{cv}} = \overline{V} - \lambda(\overline{U} - \widehat{\mu}) = \frac{1}{n}\sum_{i=1}^n \left(V_i - \lambda(U_i - \widehat{\mu})\right). \tag{13}$$

Similarly, $\lambda^\star$ may be estimated from data as $\widehat{\lambda} = \widehat{\mathrm{cov}}((U_i,V_i)_{i=1}^n)/\widehat{\mathrm{var}}((U_i)_{i=1}^n)$, where $\widehat{\mathrm{var}}(\cdot)$ and $\widehat{\mathrm{cov}}(\cdot)$ denote the sample variance and covariance, respectively. Plugging $\widehat{\lambda}$ into (13) defines $\widehat{\gamma}^{\mathrm{cv}+} := \widehat{\gamma}_{\widehat{\lambda}}^{\mathrm{cv}}$, which is similar to the semi-supervised least squares estimator of Zhang et al. [11, Eq. (2.15)]. As discussed in Section 3, deriving an AsympCS requires a strong coupling between the estimator and a sequence of i.i.d. Gaussian random variables. We now establish this coupling, a key ingredient for AsympCS for CVEs (and, in particular, for PPI estimators).

Proposition 1 (Asymptotics for CVEs)**.** *Assume* $\mathbb{E}|U|^{2+\delta}$ *and* $\mathbb{E}|V|^{2+\delta} < \infty$ *for some* $0 < \delta < 1$*. Then, almost surely as* $n \to \infty$*,*

$$\widehat{\gamma}^{\mathrm{cv}+} = \widehat{\gamma}_{\lambda^\star}^{\mathrm{cv}} + o\left(\frac{1}{\sqrt{n\log n}}\right) = \overline{V} - \lambda^\star(\overline{U} - \widehat{\mu}) + o\left(\frac{1}{\sqrt{n\log n}}\right). \tag{14}$$

Proposition 2 (KMT coupling for CVEs)**.** *Assume* $\mathbb{E}|U|^{2+\delta}$ *and* $\mathbb{E}|V|^{2+\delta} < \infty$ *for some* $0 < \delta < 1$*. Assume additionally that* $|\frac{n}{N_n} - r| = O(1/n^{1-a})$ *with* $0 < a < 2/(2+\delta)$*, for some* $r \in [0,1]$*. Then, there exist i.i.d. Gaussian random variables* $(W_i^{\mathrm{cv}})_{i\geq 1}$ *with mean* γ *and variance*

$$\nu_\lambda^{\mathrm{cv}} := \mathrm{var}(V - \lambda U) + r\mathrm{var}(\lambda U) = \mathrm{var}(V) - 2\lambda\mathrm{cov}(U,V) + \lambda^2(1+r)\mathrm{var}(U)$$

such that, almost surely as $n \to \infty$*,*

$$\widehat{\gamma}_\lambda^{\mathrm{cv}} = \frac{1}{n}\sum_{i=1}^n W_i^{\mathrm{cv}} + o\left(\frac{1}{\sqrt{n\log n}}\right). \tag{15}$$

Similarly, there exist i.i.d. Gaussian random variables $(W_i^{\mathrm{cv}+})_{i\geq 1}$ *with mean* γ *and variance* $\nu^{\mathrm{cv}+} := \nu^{\mathrm{cv}}_{\lambda^\star} = \mathrm{var}(V)\left[1-(1-r)\rho^2_{U,V}\right]$ *such that, almost surely as* $n \to \infty$*,*

$$\widehat{\gamma}^{\mathrm{cv}+} = \frac{1}{n}\sum_{i=1}^{n} W_i^{\mathrm{cv}+} + o\left(\frac{1}{\sqrt{n\log n}}\right). \tag{16}$$

The estimators

$$\widehat{\nu}^{\mathrm{cv}}_{\lambda}((U_i,V_i)_{i=1}^n,(\widetilde{U}_j)_{j=1}^{N_n}) = \frac{1}{n-2}\sum_{i=1}^{n}(V_i-\overline{V}-\lambda(U_i-\overline{U}))^2 + \frac{n\lambda^2}{N_n(N_n-1)}\sum_{j=1}^{N_n}(\widetilde{U}_j-\widehat{\mu})^2 \tag{17}$$

$$\widehat{\nu}^{\mathrm{cv}+}((U_i,V_i)_{i=1}^n) = \frac{1-n/N_n}{n-2}\sum_{i=1}^{n}(V_i-\overline{V}-\widehat{\lambda}(U_i-\overline{U}))^2 + \frac{n/N_n}{n-1}\sum_{i=1}^{n}(V_i-\overline{V})^2 \tag{18}$$

are consistent estimators of $\nu^{\mathrm{cv}}_{\lambda}$ *and* $\nu^{\mathrm{cv}+}$*, respectively, where* $\widehat{\mu} = \frac{1}{N_n}\sum_{j=1}^{N_n}\widetilde{U}_j$*.*

4.2 PPI estimators: definitions and asymptotic properties

Owing to Equation (2), the PPI estimator $\widehat{\theta}_n$ is the value of θ that solves the equation $\widehat{g}_{\theta,n} = 0$, where $\widehat{g}_{\theta,n} = \widehat{m}_{\theta,n} + \widehat{\Delta}_{\theta,n}$ is an estimator of g_θ. Here, $\widehat{m}_{\theta,n}$ and $\widehat{\Delta}_{\theta,n}$ are estimators of m_θ and Δ_θ, respectively. A typical choice for $\widehat{m}_{\theta,n}$ is the sample mean of the unlabelled data,

$$\widehat{m}_{\theta,n} = \frac{1}{N_n}\sum_{j=1}^{N_n}\ell'_\theta(\widetilde{X}_j, f(\widetilde{X}_j)). \tag{19}$$

Different choices for $\widehat{\Delta}_{\theta,n}$ have been proposed in the literature, leading to different PPI estimators.

Standard PPI. Angelopoulos et al. [1] use the sample mean

$$\widehat{\Delta}^{\mathsf{PP}}_{\theta,n} = \frac{1}{n}\sum_{i=1}^{n}\left(\ell'_\theta(X_i,Y_i) - \ell'_\theta(X_i,f(X_i))\right) \tag{20}$$

as an estimator for Δ_θ. Combining Equation (20) with Equation (19),

$$\widehat{g}^{\mathsf{PP}}_{\theta,n} = \widehat{m}_{\theta,n} + \widehat{\Delta}^{\mathsf{PP}}_{\theta,n} = \left[\frac{1}{n}\sum_{i=1}^{n}\ell'_\theta(X_i,Y_i)\right] - \left(\left[\frac{1}{n}\sum_{i=1}^{n}\ell'_\theta(X_i,f(X_i))\right] - \widehat{m}_{\theta,n}\right) \tag{21}$$

is a CVE, with control variate $\ell'_\theta(X_i,f(X_i)) - \widehat{m}_{\theta,n}$ and control-variate parameter $\lambda = 1$. For the squared loss, the estimator $\widehat{\theta}^{\mathsf{PP}}_n$ solving $\widehat{g}^{\mathsf{PP}}_{\theta,n} = 0$ also takes the control-variate form

$$\widehat{\theta}^{\mathsf{PP}}_n = \frac{1}{n}\sum_{i=1}^{n}Y_i - \left(\frac{1}{n}\sum_{i=1}^{n}f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n}f(\widetilde{X}_j)\right), \tag{22}$$

with control variate $f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n}f(\widetilde{X}_j)$ and $\lambda = 1$.

PPI++. Angelopoulos et al. [7] extend the standard PPI estimator (21) by allowing the control-variate parameter λ, which they call *power-tuning* parameter, to take values other than 1. The resulting estimator is

$$\widehat{\Delta}^{\mathsf{PP+}}_{\theta,n} = \widehat{\Delta}^{\mathsf{PP}}_{\theta,n} - (\widehat{\lambda}_{\theta,n}-1)\left(\frac{1}{n}\left[\sum_{i=1}^{n}\ell'_\theta(X_i,f(X_i))\right] - \widehat{m}_{\theta,n}\right), \tag{23}$$

where $\widehat{\lambda}_{\theta,n}$ is the estimator $\widehat{\lambda}_{\theta,n} = \widehat{\mathrm{cov}}\left((\ell'_\theta(X_i,Y_i),\ell'_\theta(X_i,f(X_i)))_{i=1}^n\right)/\widehat{\mathrm{var}}\left((\ell'_\theta(X_i,f(X_i)))_{i=1}^n\right)$. In this case, $\widehat{\Delta}^{\mathsf{PP+}}_{\theta,n}$ is a CVE with centred control variate $\ell'_\theta(X_i,f(X_i)) - \widehat{m}_{\theta,n}$, which depends only on the black-box predictions. As a result of this choice,

$$\widehat{g}^{\mathsf{PP+}}_{\theta,n} = \widehat{m}_{\theta,n} + \widehat{\Delta}^{\mathsf{PP+}}_{\theta,n} = \left[\frac{1}{n}\sum_{i=1}^{n}\ell'_\theta(X_i,Y_i)\right] - \widehat{\lambda}_{\theta,n}\left(\left[\frac{1}{n}\sum_{i=1}^{n}\ell'_\theta(X_i,f(X_i))\right] - \widehat{m}_{\theta,n}\right) \tag{24}$$

is also a CVE. Under the squared loss, we obtain

$$\widehat{\theta}_n^{\mathtt{PP+}} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\lambda}_{0,n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j)\right), \tag{25}$$

where in this case $\widehat{\lambda}_{\theta,n} = \widehat{\lambda}_{0,n}$ for all θ.

Standard asymptotic confidence intervals for PPI and PPI++ rely on CLTs for the estimators $\widehat{g}_{\theta,n}$, $\widehat{m}_{\theta,n}$ and $\widehat{\Delta}_{\theta,n}$. In contrast, constructing asymptotic confidence sequences requires almost sure approximations by averages of i.i.d. Gaussian variables. Since the estimators for g_θ, m_θ and Δ_θ are all CVEs, the asymptotic results of Proposition 1 and the KMT coupling of Proposition 2 both apply.

5 Anytime-valid, Bayes-assisted, prediction-powered inference

In this section, we show how the results of Sections 3 and 4 can be combined in the context of PPI to obtain AsympCS for g_θ. For any $\theta \in \mathbb{R}$ and $i \geq 1$, let $U_{\theta,i} = \ell'_\theta(X_i, f(X_i))$, $\widetilde{U}_{\theta,i} = \ell'_\theta(\widetilde{X}_i, f(\widetilde{X}_i))$ and $V_{\theta,i} = \ell'_\theta(X_i, Y_i)$. Define $\overline{V}_{\theta,n} = \frac{1}{n}\sum_{i=1}^n V_{\theta,i}$ and $\overline{U}_{\theta,n} = \frac{1}{n}\sum_{i=1}^n U_{\theta,i}$. In the following, we assume $\mathbb{E}|U_{\theta,i}|^{2+\delta}$, $\mathbb{E}|\widetilde{U}_{\theta,i}|^{2+\delta}$ and $\mathbb{E}|V_{\theta,i}|^{2+\delta} < \infty$ for some $0 < \delta < 1$, and that $|\frac{n}{N_n} - r| = O(1/n^{1-a})$ with $0 < a < 2/(2+\delta)$ for some $r \in [0,1]$.

5.1 Anytime-valid PPI

We first derive AsympCS that do not incorporate prior information on the accuracy of the black-box predictor. The following result follows directly from Proposition 2 and Theorem 2, owing to the control-variate form of the PPI estimator $\widehat{g}^{\mathtt{PP}}_{\theta,n}$ (21) and the PPI++ estimator $\widehat{g}^{\mathtt{PP+}}_{\theta,n}$ (24).

Proposition 3. *Let $\widehat{g}_{\theta,n}$ be either the PPI* (21) *or the PPI++* (24) *estimator. For PPI, let $(\widehat{\sigma}^g_{\theta,n})^2 = \widehat{\nu}^{\mathrm{cv}}_1((U_{\theta,i}, V_{\theta,i})_{i=1}^n, (\widetilde{U}_{\theta,j})_{j=1}^{N_n})$ (see* (17)*). For PPI++, let $(\widehat{\sigma}^g_{\theta,n})^2 = \widehat{\nu}^{\mathrm{cv}+}((U_{\theta,i}, V_{\theta,i})_{i=1}^n)$ (see* (18)*). Then, for any $\rho > 0$, the sequence of intervals defined as $\mathcal{C}^g_{\alpha,\theta,n} = \mathcal{C}^{\mathtt{NA}}_{\alpha,n}(\widehat{g}_{\theta,n}, \widehat{\sigma}^g_{\theta,n}; \rho)$ forms a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{n\log n}$ for g_θ and asymptotic Type-I error control.*

5.2 Anytime-valid, Bayes-assisted, PPI

In many modern applications, extremely accurate black-box predictors are available (e.g., [26, 27, 28]). When this is the case, we can leverage this prior information to obtain tighter AsympCS for g_θ via a zero-mean prior on Δ_θ. Following the decomposition in Equation (3), we combine an AsympCS for m_θ (Proposition 4) with a Bayes-assisted AsympCS for Δ_θ (Proposition 5).

Proposition 4 (AsympCS for m_θ)**.** *Let $\widehat{m}_{\theta,n}$ and $(\widehat{\sigma}^f_{\theta,n})^2$ be the sample mean* (19) *and sample variance of $(\ell'_\theta(\widetilde{X}_j, f(\widetilde{X}_j)))_{j=1}^{N_n}$. Let $\delta \in (0,1)$. For any $\rho > 0$, $\mathcal{R}_{\delta,\theta,n} = \mathcal{C}^{\mathtt{NA}}_{\delta,n}(\widehat{m}_{\theta,n}, \widehat{\sigma}^f_{\theta,n}; \rho)$ forms a $(1-\delta)$–AsympCS with approximation rate $1/\sqrt{n\log n}$ for m_θ and asymptotic Type-I error control.*

Proposition 5 (Bayes-assisted AsympCS for Δ_θ)**.** *For PPI, let $\widehat{\Delta}_{\theta,n}$ and $(\widehat{\sigma}^\Delta_{\theta,n})^2$ be the sample mean* (20) *and sample variance of $(V_{\theta,i} - U_{\theta,i})_{i=1}^n$. For PPI++, let $\widehat{\Delta}_{\theta,n}$ be the control-variate estimator* (23) *and $(\widehat{\sigma}^\Delta_{\theta,n})^2 = \widehat{\nu}^{\mathrm{cv}+}((U_{\theta,i}, V_{\theta,i} - U_{\theta,i})_{i=1}^n)$ (see* (18)*). Let $\kappa \in (0,1)$. For any continuous proper prior π, the sequence of Bayes-assisted intervals $\mathcal{T}_{\kappa,\theta,n} = \mathcal{C}^{\mathtt{BA}}_{\kappa,n}(\widehat{\Delta}_{\theta,n}, \widehat{\sigma}^\Delta_{\theta,n}; \pi)$ forms a $(1-\kappa)$–AsympCS with approximation rate $1/\sqrt{n\log n}$ for Δ_θ and asymptotic Type-I error control.*

Finally, for both PPI and PPI++, the confidence sequences $\mathcal{R}_{\delta,\theta,n}$ and $\mathcal{T}_{\alpha-\delta,\theta,n}$ are combined via a Minkowski sum to obtain a $(1-\alpha)$–AsympCS for g_θ with approximation rate $1/\sqrt{n\log n}$ for μ and asymptotic Type-I error control, of the form

$$\mathcal{C}^g_{\alpha,\theta,n} = \left[\widehat{g}_{\theta,n} \pm \left\{\frac{\widehat{\sigma}^\Delta_{\theta,n}}{\sqrt{n}}\sqrt{\log\left(\frac{n(2\pi\kappa^2)^{-1}}{\eta_n(\widehat{\Delta}_{\theta,n}/\widehat{\sigma}^\Delta_{\theta,n})^2}\right)} + \frac{\widehat{\sigma}^f_{\theta,n}}{\sqrt{N_n}}\sqrt{\frac{1+N_n\rho^2}{N_n\rho^2}\log\left(\frac{N_n\rho^2+1}{\delta^2}\right)}\right\}\right] \tag{26}$$

where $\widehat{g}_{\theta,n}$ is either the PPI estimator (21) or the PPI++ estimator (24). Solving Equation (4) gives the confidence region for $\theta^\star$. In the case of the squared loss, $\mathcal{C}^{\text{avpp}}_{\alpha,n}$ is an interval, given by

$$\mathcal{C}^{\text{avpp}}_{\alpha,n} = \left[\widehat{\theta}_n \pm \left\{ \frac{\widehat{\sigma}^{\Delta}_{0,n}}{\sqrt{n}} \sqrt{\log\left(\frac{n}{2\pi\kappa^2\eta_n(\widehat{\Delta}_{0,n}/\widehat{\sigma}^{\Delta}_{0,n})^2}\right)} + \frac{\widehat{\sigma}^{f}_{0,n}}{\sqrt{N_n}} \sqrt{\frac{1+N_n\rho^2}{N_n\rho^2}\log\left(\frac{N_n\rho^2+1}{\delta^2}\right)} \right\}\right] \tag{27}$$

where $\widehat{\theta}_n$ is either the PPI estimator (22) or the PPI++ estimator (25).

6 Experiments

We compare the PPI and PPI++ AsympCS procedures introduced in Section 5 – with and without Bayes assistance – to the AsympCS relying solely on labelled data (obtained from Theorem 1 and referred to as "classical") on several estimation problems. Bayes-assisted methods are annotated with (G), (L), or (T) to indicate Gaussian, Laplace, or Student-t priors with mean zero and scale depending on the task and reported in the Supplementary Material. For the Student-t prior, we set the degrees of freedom to 2 in all experiments. Since PPI is motivated by settings with scarce labelled data and abundant unlabelled data, we consider the following experimental setting: labelled data arrive sequentially, i.e., $n = 1, 2, \ldots$, while a large unlabelled dataset is available from the start, i.e., $N_n = N$ for all n, with $N \gg n$ large enough to exclude any uncertainty on the measure of fit m_θ. As discussed by Cortinovis and Caron [6], this simplifies the comparison between non-assisted and Bayes-assisted PPI, as it rules out any potential loss of efficiency due to the Minkowski sum (26), thereby isolating the effect of the Bayes correction on the CS procedure. For synthetic data, we set $N = \infty$ to guarantee the simplification holds. For real data, we empirically verify that N is large enough to justify this assumption by confirming that anytime validity is preserved – specifically, that the cumulative miscoverage rate remains below the chosen threshold $\alpha = 0.1$ for all n. As with CLT-based CIs, the n at which one starts counting the cumulative miscoverage rate of an asymptotic CS is inherently arbitrary; unless otherwise stated, we choose $n = 40$, as we empirically find this to be a reasonably small labelled sample size at which the KMT coupling generally provides a good approximation.

6.1 Synthetic data

The synthetic experiments below follow a general structure: we start with $N = \infty$ unlabelled samples $\{\widetilde{X}_j\}_{j=1}^{N} \overset{\text{iid}}{\sim} \mathbb{P}_X$ and successively sample n labelled observations $\{(X_i, Y_i)\}_{i=1}^{n} \overset{\text{iid}}{\sim} \mathbb{P}$ with the goal of estimating the mean $\theta^\star = \mathbb{E}[Y]$. We compare methods in terms of the average interval volume as n increases across repetitions, and report the associated cumulative miscoverage rate in Appendix S7.1.

Noisy predictions. This experiment demonstrates that our method can adapt to varying correlation levels between predictions and true labels by using the PPI++ estimator (23). We sample $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, so that $\theta^\star = \mathbb{E}[Y] = 0$. The prediction rule is defined as $f(X_i) = Y_i + \epsilon_i$, where X_i is only used for indexing and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_Y^2)$, with the noise level $\sigma_Y \in \{0.1, 0.8, 3\}$. In this case, the optimal control-variate parameter is given by $\lambda^\star_\theta = \lambda^\star = \text{cov}(Y, f(X))/\text{var}(f(X)) = (1+\sigma_Y^2)^{-1}$, which decreases with σ_Y. Figure 1 compares the interval volume achieved by classical and non-assisted CS procedures as a function of n, while results under informative priors are reported in Appendix S7.1. For small noise levels, PPI and PPI++ achieve similar performance, and greatly outperform classical inference. As the noise level grows, the machine learning predictions become less informative and standard PPI loses ground to the classical CS. By contrast, PPI++ adapts to the noise level and always performs similarly to, or better than, the other baselines.

Biased predictions. This experiment illustrates the potential benefits of incorporating prior information into our method. We sample $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $Y_i = X_i + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} t_{\text{df}}(0,1)$, so that $\theta^\star = \mathbb{E}[Y] = 0$. The prediction rule is defined as $f(X_i) = X_i + \upsilon$, where $\upsilon \in \mathbb{R}$ controls its bias level. For all υ, $\lambda^\star = 1$, so that PPI and PPI++ coincide. We vary υ between -1.2 and 1.2, and $\text{df} \in \{5, 10, \infty\}$ to study the impact of bias level and noise distribution on the AsympCS procedures. Figure 2 compares the average interval volumes at $n = 100$ as a function of υ for each value of df. Classical inference and non-assisted PPI volumes remain essentially constant across bias levels,

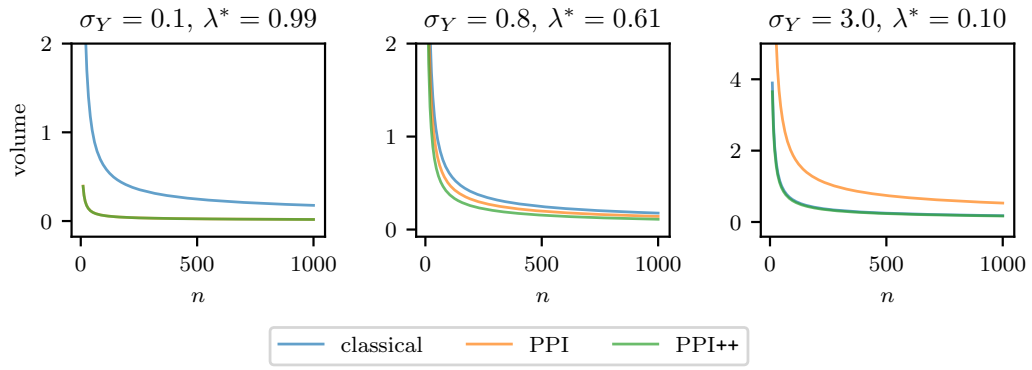

Figure 1: Noisy predictions study. The left, middle and right panels show average interval volume over 1000 repetitions as a function of the labelled sample size n for noise levels $\sigma_Y \in \{0.1, 0.8, 3.0\}$.

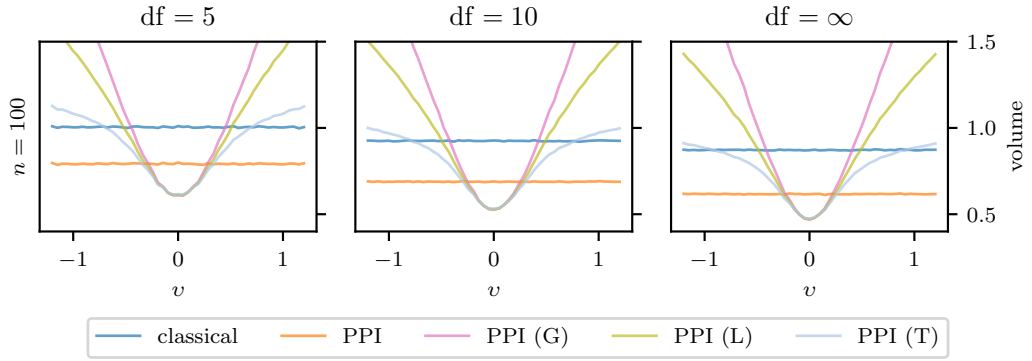

Figure 2: Biased predictions study. The left, middle and right panels show average interval volume over 100 repetitions as a function of the bias level υ for $\mathrm{df} = 5, 10, \infty$.

reflecting their lack of prior information, and with the latter consistently outperforming the former by leveraging imputed predictions. On the other hand, the volume of the Bayes-assisted procedures varies widely with the bias level υ: the volume is reduced for small υ, but grows with $|\upsilon|$ as the priors become increasingly misspecified. Notably, the volume under the Gaussian prior inflates the fastest with $|\upsilon|$, while heavier-tailed Laplace and Student-t priors offer comparatively greater robustness. These conclusions hold for all values of df, which controls the accuracy of the KMT coupling approximation for a given n. Coverage results in Appendix S7.1 show that, while smaller values of df lead to slightly worse coverage, the approximation quality is overall satisfactory in this example.

6.2 Real data

We evaluate our method on several real-world datasets, which are described in Appendix S6.2. While each dataset is, in principle, static (providing label/prediction pairs $\{(Y_i, f(X_i))\}_{i=1}^{N+n_1}$), we simulate an online setting akin to Section 6.1 by randomly splitting the data into a labelled set of size n_1, which serves as a labelled data stream, and an unlabelled set of size N.

Figure 3 compares classical and PPI++ AsympCS procedures on the FLIGHTS, FOREST, and GALAXIES datasets, where the goal is mean estimation. By taking advantage of the unlabelled data, PPI methods consistently yield smaller regions than the classical counterpart, while maintaining reliable coverage. Moreover, Bayes-assisted approaches further improve efficiency for moderate labelled sample sizes, as the quality of the predictions is generally high in these datasets.

Figure S11 reports results for three additional estimation tasks: linear regression (CENSUS), logistic regression (HEALTHCARE), and quantile estimation (GENES). For the first two tasks, the same

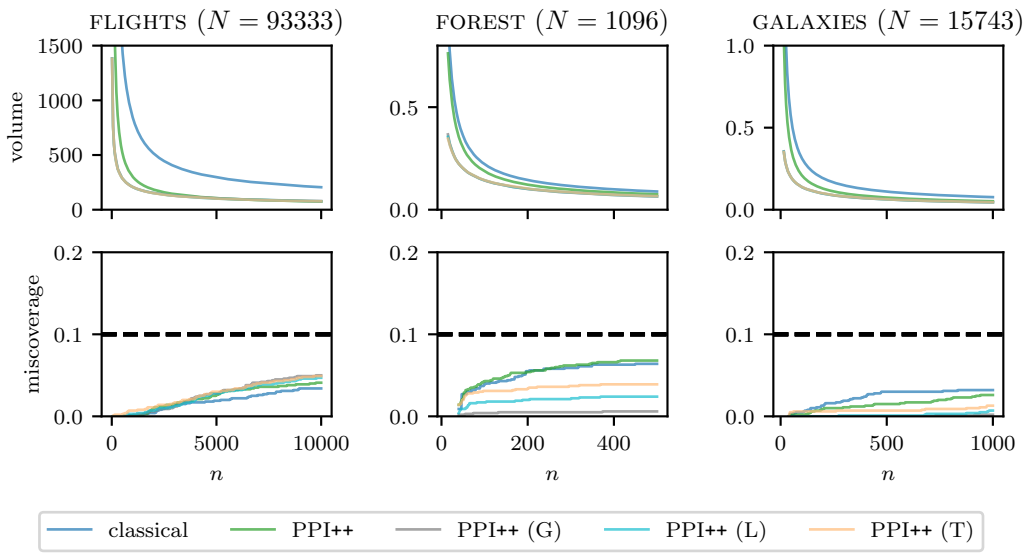

Figure 3: Mean estimation. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 1000 repetitions for the FLIGHTS, FOREST, and GALAXIES datasets.

conclusions as for mean estimation hold: PPI methods consistently outperform classical inference, with Bayes-assisted approaches providing an additional efficiency boost. For the quantile estimation task, non-assisted PPI still improves over classical inference by leveraging the machine learning predictions; however, the Bayes-assisted methods yield larger regions than the other approaches, reflecting lower prediction quality in this dataset.

7 Discussion

We extended the PPI framework to the sequential setting via asymptotic confidence sequences, which further allow for seamless integration of prior information about the quality of the auxiliary predictions. However, several directions merit further investigation. The results developed here are for scalar parameter values θ. Extensions to multivariate settings are discussed in Appendix S4, building on earlier work by Waudby-Smith et al. [15, §B.10]. In the non-assisted case, we focused on asymptotic confidence sequences of the form (6), but other options are possible. In particular, as discussed in Appendix S8, the parameter-free CS proposed by Wang and Ramdas [19], which is based on an improper prior, may be used as an exact reference CS in place of Equation (6).

The AsympCS derived in this paper are asymptotically valid for i.i.d. data under mild, nonparametric assumptions. Promising directions include extensions to non-i.i.d. observations, as well as the development of *nonasymptotic*, nonparametric Bayes-assisted confidence sequences under stricter assumptions (e.g., bounded means), building on the work of Waudby-Smith and Ramdas [15]. In the non-assisted case, the parameter ρ was assumed to be fixed. Waudby-Smith et al. [15, §2.5] considered delayed-start sequences $\mathcal{C}_{\alpha,t}(m)$ that may depend on the start time m; this includes allowing the tuning parameter ρ to depend on m. Their asymptotic Type-I error control result, derived under assumptions similar to those used here, also applies in our setting. Another interesting direction would be to adapt similar ideas to the Bayes-assisted construction.

PPI AsympCS procedures share the computational considerations of their fixed-time counterparts. Beyond mean estimation (e.g., Figure S11), they typically require constructing a grid over θ. When the marginal density η_t is not available in closed form (e.g., for the Student-t prior), the Bayes-assisted version involves numerical integration. If computation is a concern, the Laplace prior offers a good compromise: it has heavier tails than the Gaussian while still admitting a closed-form expression for η_t.

Acknowledgments and Disclosure of Funding

Valentin Kilian is supported by the Clarendon Funds Scholarship. Stefano Cortinovis is supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). The authors thank the reviewers for their time and valuable feedback, especially the suggestion to incorporate a discussion on Type-I error control.

References

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

[15] I. Waudby-Smith, D. Arbour, R. Sinha, E. Kennedy, and A. Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 52(6):2613–2640, 2024.

[9] J. Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

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

[8] T. L. Lai. On confidence sequences. *The Annals of Statistics*, pages 265–280, 1976.

[6] S. Cortinovis and F. Caron. FAB-PPI: Frequentist, assisted by Bayes, prediction-powered inference. In *International Conference on Machine Learning (ICML'2025)*, 2025.

[7] A. Angelopoulos, J. Duchi, and T. Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.

[8] J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[9] C.-E. Särndal, B. Swensson, and J. Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

[10] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[11] A. Zhang, L. D. Brown, and T. T. Cai. Semi-supervised inference: general theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

[12] Y. Zhang and J. Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2022.

[13] D. A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

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

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

[17] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[17] A. Ramdas and R. Wang. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1(1-2):1–390, 2025. ISSN 2978-4212. doi: 10.1561/3600000002.

[18] T. Zrnic and E. J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[19] D. Csillag, C. Jose Struchiner, and G. Tegoni Goedert. Prediction-powered e-values. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[20] A. Dalal, P. Blöbaum, S. Kasiviswanathan, and A. Ramdas. Anytime-Valid Inference for Double/Debiased Machine Learning of Causal Parameters. *arXiv:2408.09598*, 2024. doi: 10.48550/arXiv.2408.09598.

[14] S. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.

[2] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent RV'-s, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(1): 111–131, March 1975. ISSN 1432-2064. doi: 10.1007/BF00533093.

[3] P. Major. The approximation of partial sums of independent RV's. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35(3):213–220, September 1976. ISSN 1432-2064. doi: 10.1007/BF00532673.

[24] A. Bibaut, N. Kallus, and M. Lindon. Near-optimal non-parametric sequential tests and confidence sequences with possibly dependent observations. *arXiv preprint arXiv:2212.14411*, 2022.

[25] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

[26] A. Dal Pozzolo, O. Caelen, R. A Johnson, and G. Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE symposium series on computational intelligence*, pages 159–166. IEEE, 2015.

[27] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

[28] J. M. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A A Kohl, A. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

[19] H. Wang and A. Ramdas. The extended Ville's inequality for nonintegrable nonnegative supermartingales. *arXiv preprint arXiv:2304.01163*, 2023.

Supplement to Anytime-valid, Bayes-assisted, Prediction-Powered Inference

The supplementary material is organised as follows. Appendix S1 gives additional background on strong laws and couplings, and confidence sequences. Appendix S2 states secondary results and their proofs. Appendix S3 presents the proofs of the main theorems and propositions. Appendix S4 extends the results to the multivariate setting. Appendix S5 gives specific expressions for the case of prediction-powered mean estimation. Appendix S6 details the experimental setup used in the main text. Appendix S7 presents further experiments. Finally, Appendix S8 discusses a parameter-free, non-assisted, AsympCS, and provides some additional comparisons.

For clarity, all sections, theorems, propositions, and lemmas in the supplementary material are prefixed with "S" to distinguish them from those in the main text.

Contents

S1 Additional background

S1.1 Asymptotic theory of partial sums

S1.1.1 Iterated logarithm and Marcinkiewicz-Zygmund strong laws

Theorem S6. *(Iterated Logarithm Law [1, Theorem 8.5.2]) Let $(Y_t)_{t\geq 1}$ be i.i.d. random variables with zero mean and unit variance. Let $S_t = \sum_{i=1}^t Y_i$. Then,*

$$\limsup_{t\to\infty} \frac{|S_t|}{\sqrt{2t\log\log t}} = 1 \quad \textit{a.s.},$$

which implies

$$\left|\frac{S_t}{t}\right| = O\left(\sqrt{\frac{\log\log t}{t}}\right) \quad \textit{a.s. as } t\to\infty.$$

Theorem S7. *(Marcinkiewicz-Zygmund strong law of large numbers [1, Theorem 2.5.12]) Let $(Y_t)_{t\geq 1}$ be i.i.d. random variables with zero mean and $\mathbb{E}|Y_1|^p < \infty$ for some $1 < p < 2$. Let $S_t = \sum_{i=1}^t Y_i$. Then,*

$$\frac{S_t}{t^{1/p}} \to 0 \quad \textit{a.s. as } t\to\infty.$$

S1.1.2 Strong approximations

The following strong invariance result, attributed to Komlós, Major and Tusnady (KMT) [2, 3] shows that the partial sums of i.i.d. random variables can be approximated almost surely by a Brownian motion path. The following theorem is from Csörgö and Hall [4, Theorem 3.2].

Theorem S8 (KMT strong coupling [2, 3])**.** *Let $(Y_t)_{t\geq 1}$ be i.i.d. random variables with zero mean and unit variance such that $\mathbb{E}|Y_1|^q < \infty$ for some $q > 2$. Then, there exists a Brownian motion B such that, if we write $S_t = \sum_{i=1}^t Y_i$, we have*

$$S_t - B_t = o(t^{1/q}) \quad \textit{a.s. as } t\to\infty.$$

KMT strong coupling has been extended by Einmahl [5] to random vectors (see also [6, Section B11]).

Theorem S9. *(Multivariate KMT strong coupling [5]) Let $(Y_t)_{t\geq 1}$ be i.i.d. random vectors in $\mathbb{R}^d$ with zero mean, covariance matrix Σ, and such that $\mathbb{E}||Y_1||^q < \infty$ for some $q > 2$. Let $S_t = \sum_{i=1}^t Y_i$. Then, there exists a standard multivariate Brownian motion B such that,*

$$\Sigma^{-1/2} S_t - B_t = o(t^{1/q}) \quad \textit{a.s. as } t\to\infty.$$

S1.2 Confidence sequences

S1.2.1 Confidence intervals vs. confidence sequences

Let $(X_t)_{t\geq 1}$ be an observed data stream and let $\mu \in \mathbb{R}$ denote a fixed but unknown parameter (e.g., a mean). Write $\mathcal{F}_t = \sigma(X_{1:t})$ for the natural filtration, and let $\alpha \in (0,1)$ be a pre-specified error probability (so the confidence level is $1-\alpha$).

Fixed-time confidence intervals. A *(fixed-time) confidence interval* (CI) for μ at time t is an $\mathcal{F}_t$-measurable random set $\mathcal{C}_{\alpha,t} \subseteq \mathbb{R}$ such that

$$\Pr\big(\mu \in \mathcal{C}_{\alpha,t}\big) \;\geq\; 1-\alpha.$$

This guarantee is *marginal in t*: it holds for any chosen, deterministic t, but it need not be valid if t is selected after looking at the data (e.g., by continual monitoring or a data-dependent stopping rule). In particular, for a general $\mathcal{F}_t$-stopping time τ,

$$\Pr\big(\mu \in \mathcal{C}_{\alpha,\tau}\big) \quad \text{can be } \;<\; 1-\alpha,$$

unless the procedure is explicitly designed to be valid under optional stopping. Moreover, the family $(\mathcal{C}_{\alpha,t})_{t\geq 1}$ of fixed-time CIs need not be nested across t; disjoint intervals at different sample sizes can occur with positive probability (see Figure S4), illustrating the lack of any simultaneous-in-time guarantee.

Confidence sequences. A *confidence sequence* (CS) at level $1-\alpha$ is a sequence of $\mathcal{F}_t$-measurable random sets $(\mathcal{C}_{\alpha,t})_{t\geq 1}$ such that

$$\Pr\big(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq 1\big) \;\geq\; 1-\alpha.$$

Equivalently,

$$\Pr\Big(\sup_{t\geq 1} \mathbf{1}\{\mu \notin \mathcal{C}_{\alpha,t}\} = 1\Big) \;\leq\; \alpha.$$

The quantifier "for all t" lies *inside* the probability, yielding *uniform-in-time* (a.k.a. anytime-valid) coverage. A key consequence is validity under arbitrary data-dependent stopping: for every (a.s. finite) stopping time τ,

$$\Pr\big(\mu \in \mathcal{C}_{\alpha,\tau}\big) \;\geq\; 1-\alpha.$$

Thus CSs support continual monitoring and sequential decision-making without inflating error rates. Practically, CSs are typically wider than fixed-time CIs at the same t (especially early on) because they control the *maximum* over all times; widths often shrink with t and can approach classical rates up to iterated-logarithm factors.

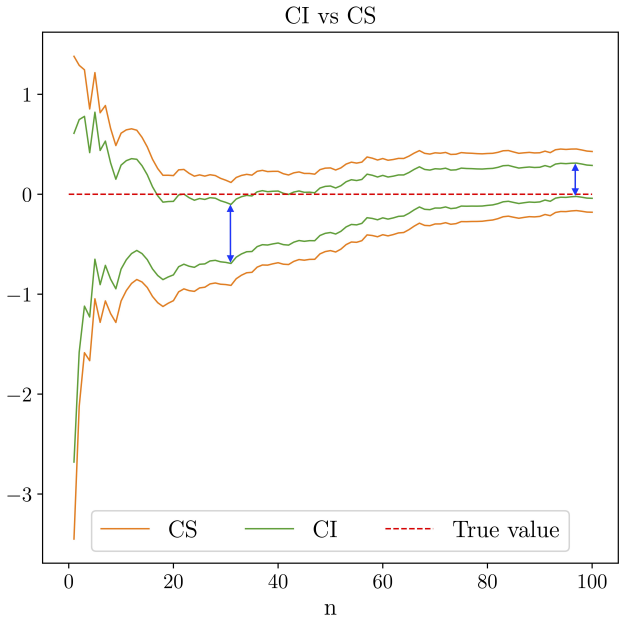

Figure S4: Comparison of fixed-time confidence intervals (CIs) and a confidence sequence (CS) for data from $\mathcal{N}(0,1)$. Two fixed-time CIs at different sample sizes happen to be disjoint (highlighted), illustrating that marginal coverage at each t does not imply simultaneous coverage over t. The CS is more conservative at small t, but its coverage holds uniformly over all t.

Early examples of CSs go back to sequential analysis [7, 8], and modern constructions often proceed via nonnegative supermartingales/test martingales and time-uniform concentration inequalities.

S1.2.2 Nonnegative supermartingale and Ville's inequality

Definition S3 (Nonnegative supermartingale)**.** *$M = (M_t)_{t\geq 1}$ is a nonnegative supermartingale (NSM) with respect to the filtration $(\mathcal{F}_t)_{t\geq 1}$ if $M_t \geq 0$ a.s, $\mathbb{E}[M_t] < \infty$ for all $t \geq 1$, and*

$$\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] \leq M_t \text{ a.s.}$$

If there is equality, then M is a nonnegative martingale.

Proposition S6 (Ville's inequality [9])**.** *Let $(M_t)_{t\geq 1}$ be a nonnegative supermartingale. For any constant $c > 0$,*

$$\Pr\left(\sup_{t\geq 1} M_t \geq c\right) \leq \frac{\mathbb{E}[M_1]}{c}$$

Ville's inequality can be seen as a generalisation of Markov's inequality. We have the following direct corollary.

Proposition S7. *Let $(M_t)_{t\geq 1}$ be a nonnegative supermartingale. For any $\alpha \in (0,1)$,*

$$\Pr\left(M_t \leq \frac{\mathbb{E}[M_1]}{\alpha} \text{ for all } t \geq 1\right) \geq 1-\alpha.$$

S1.2.3 Method of mixture

Under the appropriate conditions, mixtures of martingales remain martingales:

Proposition S8 (Lemma B1, [10])**.** *Let $\{(M_t(\mu'))_{t\in\mathbb{N}}, \mu' \in \mathbb{R}\}$ be a family of (super)martingales on a filtered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_t)_{t\in\mathbb{N}}, \Pr)$, indexed by μ' in a measurable space $(\mathbb{R}, \mathcal{B})$, such that*

1. *each $M_t(\mu')$ is $\mathcal{F}_t \otimes \mathcal{B}$-measurable; and*
2. *each $\mathbb{E}\left[M_t(\mu') \mid \mathcal{F}_{t-1}\right]$ is $\mathcal{F}_{t-1} \otimes \mathcal{B}$-measurable.*

Let π be a finite measure on $(\mathbb{R}, \mathcal{B})$ such that for all n,

$$\Pr \otimes \pi\text{-almost everywhere } M_t(\mu') \geq 0, \quad \text{or} \quad \mathbb{E}_{\mu'\sim\pi}\mathbb{E}\left[|M_t(\mu')|\right] < \infty$$

Then the mixture $\left(\tilde{M}_t\right)_{t\in\mathbb{N}}$, where $\tilde{M}_t = \mathbb{E}_{\mu'\sim\pi} M_t(\mu')$, is also a (super)martingale.

This is useful as it leads to the *method of mixtures* : if we have a family of nonnegative supermartingale (say) of the form $M_t(\mu')$ for $\mu' \in \mathbb{R}$ which satisfy conditions 1 and 2 above and a mixture distribution π satisfying the assumptions of Proposition S8, then we can conclude that $\int_{\mu'\in\mathbb{R}} M_t(\mu') d\pi(\mu')$ is also a supermartingale, and thus Ville's inequality gives for any $\alpha \in (0,1)$

$$\Pr\left(\int_{\mu'\in\mathbb{R}} M_t(\mu') d\pi(\mu') \leq \frac{1}{\alpha} \text{ for all } t \geq 1\right) \geq 1-\alpha. \tag{S28}$$

The method of mixtures dates back at least to Ville [9] and was developed in the context of sequential analysis by Wald [11]. It was then systematised and popularised by Darling and Robbins in the late 1960s, by Robbins and Siegmund in a series of papers culminating in [7], and by Lai [8]. The method of mixtures has found many applications, including confidence sequences [8, 12, 13, 14, 15], PAC-Bayes analysis [16, 10], anytime-valid testing [17], and A/B testing [18], to name but a few.

S2 Secondary results

S2.1 Strong coupling with i.i.d. Gaussian

The following proposition follows from KMT strong approximation (see Theorem S8). It will be used in the proofs of Theorem 3 and Proposition 2.

Proposition S9. *Let $\xi_1, \xi_2 \ldots$ be i.i.d. random variables with mean μ and variance σ^2 such that $\mathbb{E}|\xi_1|^q < \infty$ for some $q > 2$. Let $(N_n)_{n\geq 1}$ be a strictly increasing sequence of positive integers with $N_n \geq n$. Let $r \in (0,1]$ and assume $|\frac{n}{N_n} - r| = O(1/n^{1-a})$ with $0 < a < 2/q$. Then, there exists a sequence of i.i.d. Gaussian random variables $(W_i)_{i\geq 1}$ with mean μ and variance $r\sigma^2$ such that*

$$\frac{1}{N_n}\sum_{i=1}^{N_n} \xi_i = \frac{1}{n}\sum_{i=1}^{n} W_i + o\left(\frac{1}{n^{1-1/q}}\right) \text{ a.s. as } n \to \infty.$$

Proof. By Theorem S8, there exists a Brownian motion B such that, a.s. as $n \to \infty$,

$$\begin{aligned}\sum_{i=1}^{N_n} \frac{\xi_i - \mu}{\sigma} &= B_{N_n} + o\left(N_n^{1/q}\right) \\ &= B_{N_n} + o\left(n^{1/q}\right) \\ &= \frac{N_n}{n} r B_{n/r} + (B_{N_n} - \frac{N_n}{n} r B_{n/r}) + o\left(n^{1/q}\right). \end{aligned} \tag{S29}$$

We have

$$B_{N_n} - \frac{N_n}{n} r B_{n/r} = B_{N_n} - B_{n/r} + B_{n/r}(1 - \frac{N_n}{n} r).$$

$B_{N_n} - B_{n/r}$ is a zero-mean Gaussian random variable with variance

$$\begin{aligned}\mathrm{var}(B_{N_n} - B_{n/r}) &= |N_n - n/r| \\ &= \frac{N_n}{r}\left| r - \frac{n}{N_n} \right| \\ &= O(n^a).\end{aligned}$$

By an upper tail inequality for Gaussian random variables, for any $\epsilon > 0$,

$$\Pr(|B_{N_n} - B_{n/r}| > \epsilon n^{1/q}) \leq 2 \exp\left(- \frac{\epsilon^2 n^{2/q}}{\mathrm{var}(B_{N_n} - B_{n/r})} \right).$$

For $n_0 := n_0(\epsilon)$ large enough, for all $n > n_0$,

$$\exp\left(- \frac{\epsilon^2 n^{2/q}}{\mathrm{var}(B_{N_n} - B_{n/r})} \right) \leq \exp\left(-\epsilon^2 n^{2/q - a} \right) \leq \frac{1}{n^2}.$$

By comparison,

$$\sum_{n \geq 1} \Pr\left(|B_{N_n} - B_{n/r}| > \epsilon n^{1/q} \right) < \infty.$$

It follows from the Borel-Cantelli lemma that $|B_{N_n} - B_{n/r}| = o\left(n^{1/q}\right)$ a.s. as $n \to \infty$. Similarly, $B_{n/r}(1 - \frac{N_n}{n} r)$ is a zero-mean Gaussian random variable with variance $\frac{n}{r}(1 - \frac{N_n}{n} r)^2 = O(n^{2a-1}) = O(n^a)$. Using a similar proof, we obtain $B_{n/r}(1 - \frac{N_n}{n} r) = o\left(n^{1/q}\right)$ a.s. as $n \to \infty$. So, from Equation (S29), we obtain

$$\frac{1}{N_n} \sum_{i=1}^{N_n} \xi_i = \frac{1}{n}(n\mu + \sigma r B_{n/r}) + o\left(\frac{1}{n^{1 - 1/q}} \right).$$

We have,

$$n\mu + \sigma r B_{n/r} = \sum_{i=1}^{n} \left[\mu + \sigma r (B_{i/r} - B_{(i-1)/r}) \right] = \sum_{i=1}^{n} W_i,$$

where $W_i = \mu + \sigma r (B_{i/r} - B_{(i-1)/r})$ are i.i.d. Gaussian random variables with mean μ and variance $r\sigma^2$. This completes the proof. □

The following lemma will be useful in the proof of Proposition 1.

Lemma S1. *Let (U_i, V_i), $i = 1, \ldots, n$, be i.i.d. copies of a pair of random variables (U, V). Assume $\mathbb{E}|U|^{2+\delta}$ and $\mathbb{E}|V|^{2+\delta} < \infty$ for some $0 < \delta < 1$. Let $\lambda^\star = \frac{\mathrm{cov}(U,V)}{\mathrm{var}(U)}$ and $\widehat{\lambda} = \frac{\widehat{\mathrm{cov}}((U_i,V_i)_{i=1}^n)}{\widehat{\mathrm{var}}((U_i)_{i=1}^n)}$. Then*

$$|\lambda^\star - \widehat{\lambda}| = o(n^{-\frac{\delta}{2+\delta}}) \text{ a.s. as } n \to \infty.$$

Proof. Using the mean value theorem, we obtain

$$|\lambda^\star - \widehat{\lambda}| \leq \frac{1}{\mathrm{var}(U)} \left| \frac{1}{n} \sum_{i=1}^{n} (U_i V_i - \mathbb{E}(UV)) \right| + \frac{|\mathbb{E}(U)|}{\mathrm{var}(U)} \left| \overline{V} - \mathbb{E}V \right| + \frac{|\overline{V}|}{\mathrm{var}(U)} \left| \overline{U} - \mathbb{E}U \right|$$
$$+ \widehat{\mathrm{cov}}((U_i, V_i)_{i=1}^n) K_1 \left| \frac{1}{n} \sum_{i=1}^{n} (U_i^2 - \mathbb{E}(U^2)) \right| + \widehat{\mathrm{cov}}((U_i, V_i)_{i=1}^n) K_1 K_2 \left| \overline{U} - \mathbb{E}U \right|$$

where K_1 and K_2 are two constants, independent of n. By assumption, UV and U^2 have finite moments of order $1 + \delta/2$ and U and V have finite moments of order $2 + \delta$ (and so of order $1 + \delta/2$) thus we can apply Theorem S7 with $p = 1 + \delta/2$ to obtain the result. □

S2.2 Confidence sequence for i.i.d. Gaussian variables with known variance

An important step in the proof of all our results is the derivation of an exact confidence sequence for i.i.d. Gaussian variables with known variance. The non-assisted confidence sequence is a well-known result that can be found for instance in [7] or in the proof of Theorem 2.2 in [15]:

Theorem S10. *Let $W_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. For any parameter $\rho > 0$, the sequence of intervals defined as*

$$\mathcal{C}_{\alpha,t}^{\texttt{NA}}(\overline{W}_t, \sigma; \rho) := \left[\overline{W}_t \pm \frac{\sigma}{\sqrt{t}} \sqrt{\left(1 + \frac{1}{t\rho^2}\right) \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)} \right] \tag{S30}$$

is an exact $1 - \alpha$ confidence sequence for μ_0, that is

$$\Pr\left(\mu \in \mathcal{C}_{\alpha,t}^{\texttt{NA}}(\overline{W}_t, \sigma; \rho) \text{ for all } t \geq 1\right) \geq 1 - \alpha.$$

We also establish such a confidence sequence under a general prior on the mean. While Wang and Ramdas [19, Proposition C.1] propose an exact confidence sequence under a Gaussian prior, we extend their result to any continuous and proper prior.

Theorem S11. *Let $W_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and let π be a continuous and proper on μ/σ, then*

$$\mathcal{C}_{\alpha,t}^{\texttt{BA}}(\overline{W}_t, \sigma; \pi) = \left[\overline{W}_t \pm \frac{\sigma}{\sqrt{t}} \sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log \eta_t\left(\frac{\overline{W}_t}{\sigma}\right)} \right],$$

where η_t is defined in Equation (8)*, is an exact $1 - \alpha$ confidence sequence for μ_0, that is*

$$\Pr\left(\mu \in \mathcal{C}_{\alpha,t}^{\texttt{BA}}(\overline{W}_t, \sigma; \pi) \text{ for all } t \geq 1\right) \geq 1 - \alpha.$$

Of particular interest, we may consider the Gaussian prior $\mathcal{N}(\mu_0, \tau^2)$, which gives

$$\mathcal{C}_{\alpha,t}^{\texttt{BA}}(\overline{W}_t, \sigma; \mathcal{N}(\cdot; \mu_0, \tau^2)) = \left[\overline{W}_t \pm \frac{\sigma}{\sqrt{t}} \sqrt{\log\left(\frac{t\tau^2 + 1}{\alpha^2}\right) + \frac{\left(\overline{W}_t/\sigma - \mu_0\right)^2}{(\tau^2 + 1/t)}} \right].$$

Proof. The main idea is to apply the methods of mixture with the prior π. For each $t \geq 1$, let $p_{t,\mu}(w_{1:t})$ be the joint density of the $W_1, \ldots, W_t$ with respect to $\lambda^{\otimes t}$ where λ is the Lebesgue measure, for some unknown mean parameter $\mu \in \mathbb{R}$, where $w_{1:t} = (w_1, \ldots, w_t) \in \mathbb{R}^t$. To simplify notations, we drop the subscript t and simply write $p_\mu(w_{1:t})$ to denote this joint density. We have

$$p_\mu(w_1, \ldots, w_t) = C(\sigma, w_1, \ldots, w_t) \times \phi_t\left(\frac{\overline{w}_t - \mu}{\sigma}\right)$$

where $\overline{w}_t = \frac{1}{t}\sum_{i=1}^t w_i$, $C(\sigma, w_1, \ldots, w_t)$ does not depend on μ and $\phi_t(w)$ denotes the pdf of a zero-mean Gaussian random variable with variance $1/t$.

For any $\mu \in \mathbb{R}, w_{1:t} \in \mathbb{R}^n$, let

$$\tilde{M}_t(\overline{w}_t, \mu) = \int_{\mathbb{R}} \frac{p_{\mu'}(w_{1:t})}{p_{\mu}(w_{1:t})} \pi\left(\frac{\mu'}{\sigma}\right) \frac{d\mu'}{\sigma} \tag{S31}$$

$$= \int_{\mathbb{R}} \frac{\phi_t\left(\frac{\overline{w}_t - \mu'}{\sigma}\right)}{\phi_t\left(\frac{\overline{w}_t - \mu}{\sigma}\right)} \pi\left(\frac{\mu'}{\sigma}\right) \frac{d\mu'}{\sigma} \tag{S32}$$

$$= \frac{\eta_t(\overline{w}_t/\sigma)}{\phi_t((\overline{w}_t - \mu)/\sigma)}. \tag{S33}$$

We define $M_t(\mu') = \frac{p_{\mu'}(W_{1:t})}{p_{\mu}(W_{1:t})} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^t (W_i - \mu')^2 - \sum_{i=1}^t (W_i - \mu)^2\right)\right)$. We consider $(\mathcal{F}_t)_{t\in\mathbb{N}}$, the filtration adapted to the sequence of random variable $(W_t)_{t\in\mathbb{N}}$. For every $\mu' \in \mathbb{R}$ we have

$$M_t(\mu') = M_{t-1}(\mu') \times \exp\left(\frac{1}{2\sigma^2}(\mu^2 - {\mu'}^2)\right) \exp\left(-\frac{W_t}{\sigma^2}(\mu - \mu')\right),$$

and so

$$\begin{aligned}
\mathbb{E}\left[M_t(\mu') \mid \mathcal{F}_{t-1}\right] &= M_{t-1}(\mu') \times \exp\left(\frac{1}{2\sigma^2}(\mu^2 - {\mu'}^2)\right) \mathbb{E} \exp\left(-\frac{W_t}{\sigma^2}(\mu - \mu')\right) \\
&= M_{t-1}(\mu') \times \exp\left(\frac{1}{2\sigma^2}(\mu^2 - {\mu'}^2)\right) \exp\left(-\mu\frac{\mu - \mu'}{\sigma^2} + \frac{1}{2\sigma^2}(\mu - \mu')^2\right) \\
&= M_{t-1}(\mu').
\end{aligned}$$

Hence, $\{(M_t(\mu'))_{t\in\mathbb{N}}, \mu' \in \mathbb{R}\}$ is a family of martingale with respect to the adapted filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$.

$M_t(\mu')$ is clearly continuous in W_i for all $i \in \{1, ..., t\}$ and for $\mu' \in \mathbb{R}$. Hence, it is $\mathcal{F}_t \otimes \mathcal{B}(\mathbb{R})$-measurable. Similarly, $\mathbb{E}\left[M_t(\mu') \mid \mathcal{F}_{t-1}\right] = M_{t-1}(\mu')$ is $\mathcal{F}_{t-1} \otimes \mathcal{B}(\mathbb{R})$-measurable.

Finally, we have $M_t(\mu') \geq 0$ $\Pr \otimes \lambda$-almost everywhere, where λ is the Lebesgue measure on $\mathbb{R}$ (or any other measure dominated by the Lebesgue measure on $\mathbb{R}$). Then, by Proposition S8, $(\tilde{M}_t(\overline{W}_t, \mu))_{t\geq 1}$ is a nonnegative martingale with respect to the adapted filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. So by Ville's inequality, we have :

$$\Pr\left(\tilde{M}_t(\overline{W}_t, \mu) \leq \frac{1}{\alpha} \text{ for all } t \geq 1\right) \geq 1 - \alpha.$$

It follows that the sequence

$$\mathcal{C}^{\texttt{BA}}_{\alpha,t}(\overline{W}_t, \sigma; \pi) = \left\{\mu \mid \tilde{M}_t(\overline{W}_t, \mu) \leq \frac{1}{\alpha}\right\}$$

is an exact $1 - \alpha$ confidence sequence for μ. Finally,

$$\begin{aligned}
\tilde{M}_t(\overline{W}_t, \mu) \leq \frac{1}{\alpha} &\iff \exp\left(\frac{t}{2\sigma^2}(\mu - \overline{W}_t)^2\right) \eta_t\left(\frac{\overline{W}_t}{\sigma}\right) \frac{\sqrt{2\pi}}{\sqrt{t}} \leq \frac{1}{\alpha} \\
&\iff \frac{t}{2\sigma^2}(\mu - \overline{W}_t)^2 + \log\left(\eta_t\left(\frac{\overline{W}_t}{\sigma}\right) \frac{\sqrt{2\pi}}{\sqrt{t}}\right) \leq -\log(\alpha) \\
&\iff (\mu - \overline{W}_t)^2 \leq \frac{\sigma^2}{t}\left(-2\log\left(\eta_t\left(\frac{\overline{W}_t}{\sigma}\right)\right) - \log\left(\frac{2\pi\alpha^2}{t}\right)\right).
\end{aligned}$$

□

S2.3 Optimal control-variate parameter for PPI++

The next proposition follows from an application of Proposition 1 to the PPI++ estimators, identifying the value of the optimal control variate parameter in this case.

Proposition S10 (Asymptotics for PPI++)**.** *Assume that, for any $\theta \in \mathbb{R}$, $\mathbb{E}|\ell'_\theta(X_1, Y_1)|^2 < \infty$ and $\mathbb{E}|\ell'_\theta(X_1, f(X_1))|^2 < \infty$. Let*

$$\lambda^\star_\theta = \frac{\mathrm{cov}(\ell'_\theta(X_1, Y_1), \ell'_\theta(X_1, f(X_1)))}{\mathrm{var}(\ell'_\theta(X_1, f(X_1)))}. \tag{S34}$$

Then, for any θ, almost surely as $n \to \infty$,

$$\widehat{g}^{\texttt{PP+}}_{\theta,n} = \left[\frac{1}{n}\sum_{i=1}^n \ell'_\theta(X_i, Y_i)\right] - \lambda^\star_\theta \left(\left[\frac{1}{n}\sum_{i=1}^n \ell'_\theta(X_i, f(X_i))\right] - \widehat{m}_\theta\right) + o\left(\sqrt{\frac{\log\log n}{n}}\right), \tag{S35}$$

$$\widehat{\Delta}^{\texttt{PP+}}_{\theta,n} = \frac{1}{n}\sum_{i=1}^n (\ell'_\theta(X_i, Y_i) - \ell'_\theta(X_i, f(X_i))) - (\lambda^\star_\theta - 1)\left(\frac{1}{n}\left[\sum_{i=1}^n \ell'_\theta(X_i, f(X_i))\right] - \widehat{m}_\theta\right) + o\left(\sqrt{\frac{\log\log n}{n}}\right). \tag{S36}$$

Additionally, in the case of the squared loss,

$$\widehat{\theta}^{\texttt{PP+}}_n = \frac{1}{n}\sum_{i=1}^n Y_i - \lambda^\star_0 \left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j)\right) + o\left(\sqrt{\frac{\log\log n}{n}}\right), \tag{S37}$$

with $\lambda^\star_0 = \mathrm{cov}(Y_1, f(X_1))/\mathrm{var}(f(X_1))$.

S3 Proofs

S3.1 Proofs of Theorem 1 and Theorem 2

Theorem 1 is a corollary of Theorem 2, so we start by proving Theorem 2.

S3.1.1 Proof of Theorem 2

By assumption, we have, almost surely,

$$\widehat{\mu}_t = \overline{W}_t + \varepsilon_t,$$

where $\varepsilon_t = o\left(\frac{1}{\sqrt{t\log t}}\right)$ and $\overline{W}_t = \frac{1}{t}\sum_{i=1}^t W_i$.

Using Theorem S10, the sequence of intervals $\mathcal{C}^{\texttt{NA}}_{\alpha,t}(\overline{W}_t, \sigma; \rho) = \mathcal{C}^{\texttt{NA}}_{\alpha,t}(\widehat{\mu}_t - \varepsilon_t, \sigma; \rho) = [\widehat{\mu}_t - L^*_t, \widehat{\mu}_t + U^*_t]$, where

$$U^*_t = \frac{\sigma}{\sqrt{t}}\sqrt{\left(1 + \frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)} - \varepsilon_t, \text{ and}$$

$$L^*_t = \frac{\sigma}{\sqrt{t}}\sqrt{\left(1 + \frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)} + \varepsilon_t,$$

is an exact confidence sequence for μ. We have $\mathcal{C}^{\texttt{NA}}_{\alpha,t}(\widehat{\mu}_t, \widehat{\sigma}_t; \rho) = [\widehat{\mu}_t - L_t, \widehat{\mu}_t + U_t]$, where

$$U_t = L_t = \frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\left(1 + \frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)}.$$

Let $a_t = 1/\sqrt{t \log t}$. Then,

$$\frac{1}{a_t}\left(L_t - L_t^\star\right) = \frac{1}{a_t}\left[\frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\left(1+\frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)}\right.$$
$$\left. - \frac{\sigma}{\sqrt{t}}\sqrt{\left(1+\frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)}\right] + o(1).$$

We have

$$\frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\left(1+\frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)} - \frac{\sigma}{\sqrt{t}}\sqrt{\left(1+\frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)}$$
$$\sim (\widehat{\sigma}_t - \sigma) \times \sqrt{\frac{\log t}{t}} = o\left(\frac{1}{\sqrt{t\log t}}\right).$$

Hence $\frac{1}{a_t}\left(L_t - L_t^\star\right) = o(1)$. Similarly, $\frac{1}{a_t}\left(U_t - U_t^\star\right) = o(1)$. It follows that $\left(\mathcal{C}_{\alpha,t}^{\texttt{NA}}\right)$ is a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t\log t}$.

S3.1.2 Proof of Theorem 1

In order to apply Theorem 2 in this setting, we need Equation (7) to be satisfied for the i.i.d. sequence $(Y_t)_{t\geq 1}$. By KMT strong coupling (Theorem S8), there exists a sequence of i.i.d. Gaussian random variables $(W_i)_{i\geq 1}$ with mean μ and variance σ^2 such that, a.s.,

$$\overline{Y}_t = \frac{1}{t}\sum_{i=1}^{t} W_i + \varepsilon_t \quad \text{where} \quad \varepsilon_t = o\left(\frac{1}{t^{1-1/(2+\delta)}}\right) = o\left(\frac{1}{\sqrt{t\log t}}\right).$$

We also need to satisfied the condition on the variance. Under $\mathbb{E}[|Y|^{2+\delta}] < \infty$, the Marcinkiewicz-Zygmund strong law of large numbers (Theorem S7) with $p = 1 + \delta/2 \in (1,2)$ yields a polynomial a.s. rate for $\overline{Y}_t$ and $\overline{Y_t^2}$; consequently $|\widehat{\sigma}_t - \sigma| = o(t^{-\gamma})$ for some $\gamma > 0$, which implies

$$|\widehat{\sigma}_t - \sigma| = o\left(\frac{1}{\log t}\right) \text{ a.s. as } t \to \infty.$$

The result follows.

S3.2 Proofs of Theorem 3 and Theorem 4

The following lemma is a direct consequence of Theorem 8.14(b) p. 242 [20].

Lemma S2. *Let π be a proper and continuous probability density function on $\mathbb{R}^d$. Let $(Z_t)_{t\geq 1}$ be a sequence of random vectors in $\mathbb{R}^d$, with $Z_t \to c$ a.s. as $t \to \infty$. Let*

$$\eta_t(z) = \int_{\mathbb{R}^d} \mathcal{N}(z;\zeta, I_d/t)\pi(\zeta)d\zeta.$$

Then

$$\eta_t(Z_t) \to \pi(c) \text{ almost surely as } t \to \infty.$$

Theorem 3 is a corollary of Theorem 4, so we start by proving Theorem 4.

S3.2.1 Proof of Theorem 4

By assumption, we have, almost surely,

$$\widehat{\mu}_t = \overline{W}_t + \varepsilon_t,$$

where $\varepsilon_t = o\left(\frac{1}{\sqrt{t\log t}}\right)$ and $\overline{W}_t = \frac{1}{t}\sum_{i=1}^{t} W_i$.

Using Theorem S11, the sequence of intervals $\mathcal{C}^{\texttt{BA}}_{\alpha,t}(\overline{W}_t, \sigma; \pi) = \mathcal{C}^{\texttt{BA}}_{\alpha,t}(\widehat{\mu}_t - \varepsilon_t, \sigma; \pi) = [\widehat{\mu}_t - L^*_t, \widehat{\mu}_t + U^*_t]$, where

$$U^*_t = \frac{\sigma}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)} - \varepsilon_t, \text{ and}$$
$$L^*_t = \frac{\sigma}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)} + \varepsilon_t,$$

is an exact confidence sequence for μ. We have $\mathcal{C}^{\texttt{BA}}_{\alpha,t}(\widehat{\mu}_t, \widehat{\sigma}_t; \pi) = [\widehat{\mu}_t - L_t, \widehat{\mu}_t + U_t]$, where

$$U_t = L_t = \frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t}{\widehat{\sigma}_t}\right)}.$$

Let $a_t = 1/\sqrt{t\log t}$. Then,

$$\frac{1}{a_t}\left(L_t - L^\star_t\right) = \frac{1}{a_t}\left[\frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t}{\widehat{\sigma}_t}\right)}\right.$$
$$\left. - \frac{\sigma}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)}\right] + o(1)$$

We have

$$\frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t}{\widehat{\sigma}_t}\right)} - \frac{\sigma}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)}$$
$$= \frac{\frac{\widehat{\sigma}^2_t}{t}\left[\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t}{\widehat{\sigma}_t}\right)\right] - \frac{\sigma^2}{t}\left[\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)\right]}{\frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t}{\widehat{\sigma}_t}\right)} + \frac{\sigma}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)}}$$
$$\sim (\widehat{\sigma}_t - \sigma) \times \sqrt{\frac{\log t}{t}} = o\left(\frac{1}{\sqrt{t\log t}}\right)$$

as, by Lemma S2, we have

$$\eta_t\left(\frac{\widehat{\mu}_t}{\widehat{\sigma}_t}\right) \to \pi\left(\frac{\mu}{\sigma}\right) \text{ and } \eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right) \to \pi\left(\frac{\mu}{\sigma}\right) \text{ a.s. as } t \to \infty.$$

It follows that $\left(\mathcal{C}^{\texttt{BA}}_{\alpha,t}\right)$ is a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t\log t}$.

S3.2.2 Proof of Theorem 3

The result follows from Theorem 4 by applying the same reasoning as for the proof of Theorem 1.

S3.3 Proof of Theorem 5

The idea of the proof is as follows. Recall that the intervals $(\mathcal{C}_{\alpha,t})$ of interest are approximations of an exact CS $(\mathcal{C}^\star_{\alpha,t})$. For any $\alpha' > \alpha$, the narrower exact CS $(\mathcal{C}^\star_{\alpha',t})$ is eventually (a.s.) contained in $(\mathcal{C}_{\alpha,t})$ for all large t. A standard sandwiching argument using this eventual containment yields the desired asymptotic Type-I control.

Let $a_t = 1/\sqrt{t\log t}$. For each construction (non-assisted/Bayes-assisted), the sequence of intervals of interest are of the form $\mathcal{C}_{\alpha,t} = [\widehat{\mu}_t \pm U_{\alpha,t}]$ where

$$U_{\alpha,t} = \frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\left(1 + \frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)}$$

for the non-assisted case, and

$$U_{\alpha,t} = \frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t}{\widehat{\sigma}_t}\right)}$$

for the Bayes-assisted case. $\mathcal{C}_{\alpha,t}$ approximates a reference exact CS of the form $\mathcal{C}^\star_{\alpha,t} = [\widehat{\mu}_t - L^\star_{\alpha,t}, \widehat{\mu}_t + U^\star_{\alpha,t}]$, where

$$U^*_{\alpha,t} = \frac{\sigma}{\sqrt{t}}\sqrt{\left(1+\frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)} - \varepsilon_t, \text{ and}$$
$$L^*_{\alpha,t} = \frac{\sigma}{\sqrt{t}}\sqrt{\left(1+\frac{1}{t\rho^2}\right)\log\left(\frac{t\rho^2+1}{\alpha^2}\right)} + \varepsilon_t,$$

for the non-assisted case, and

$$U^*_{\alpha,t} = \frac{\sigma}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)} - \varepsilon_t, \text{ and}$$
$$L^*_{\alpha,t} = \frac{\sigma}{\sqrt{t}}\sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\widehat{\mu}_t - \varepsilon_t}{\sigma}\right)} + \varepsilon_t,$$

for the Bayes-assisted case, where $\varepsilon_t = \widehat{\mu}_t - \overline{W}_t = o(a_t)$ a.s. does not depend on α. In both cases, $(\mathcal{C}^\star_{\alpha,t})_{t\geq 1}$ is an exact CS (Theorems S10 and S11); hence, for $E^\star_\alpha = \{\mu \in \mathcal{C}^\star_{\alpha,t} \text{ for all } t \geq 1\}$,

$$\Pr(E^\star_\alpha) \geq 1-\alpha.$$

Additionally,

$$U_{\alpha,t} \sim U^\star_{\alpha,t} \sim L^\star_{\alpha,t} \sim \sigma\sqrt{\frac{\log t}{t}} \text{ a.s. as } t \to \infty,$$

and, as shown in the proofs of Theorems 2 and 4, a.s.,

$$U_{\alpha,t} - U^\star_{\alpha,t} = o(a_t) \text{ and } U_{\alpha,t} - L^\star_{\alpha,t} = o(a_t). \tag{S38}$$

Let $\alpha' \in (\alpha, 1)$. We now aim to show that, for some random, finite time $T_{\alpha'}$, $\mathcal{C}^\star_{\alpha',t} \subseteq \mathcal{C}_{\alpha,t}$ for all $t \geq T_{\alpha'}$. We have

$$U^\star_{\alpha,t} - U^\star_{\alpha',t} = \frac{(U^\star_{\alpha,t})^2 - (U^\star_{\alpha',t})^2}{U^\star_{\alpha,t} + U^\star_{\alpha',t}} \sim \frac{\sigma^2/t\log\frac{(\alpha')^2}{\alpha^2}}{2\sigma\sqrt{\log t/t}} \sim c(\alpha,\alpha')a_t \tag{S39}$$

a.s. as $t \to \infty$, where $c(\alpha,\alpha') = \sigma\log(\alpha'/\alpha) > 0$. Similarly,

$$L^\star_{\alpha,t} - L^\star_{\alpha',t} \sim c(\alpha,\alpha')a_t \text{ a.s. as } t \to \infty. \tag{S40}$$

Combining Equations (S39) and (S40) with Equation (S38), we obtain, a.s.

$$U_{\alpha,t} - U^\star_{\alpha',t} \sim c(\alpha,\alpha')a_t \tag{S41}$$
$$L_{\alpha,t} - L^\star_{\alpha',t} \sim c(\alpha,\alpha')a_t. \tag{S42}$$

So, there exists a collection of events Ω_0 with $\Pr(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$, there is a finite $T_{\alpha'}(\omega)$ with

$$U_{\alpha,t}(\omega) - U^\star_{\alpha',t}(\omega) \geq \frac{1}{2}c(\alpha,\alpha')a_t \tag{S43}$$
$$L_{\alpha,t}(\omega) - L^\star_{\alpha',t}(\omega) \geq \frac{1}{2}c(\alpha,\alpha')a_t. \tag{S44}$$

for all $t \geq T_{\alpha'}(\omega)$. Therefore, $\mathcal{C}^\star_{\alpha',t} \subseteq \mathcal{C}_{\alpha,t}$ for all $t \geq T_{\alpha'}$. It follows, that for every $m \geq 1$,

$$\Pr(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq m) \geq \Pr(E^\star_{\alpha'} \cap \{T_{\alpha'} \leq m\}) \longrightarrow_{m\to\infty} \Pr(E^\star_{\alpha'}) \geq 1-\alpha'.$$

Hence, for any $\alpha' \in (\alpha, 1)$, $\liminf_{m\to\infty}\Pr(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq m) \geq 1-\alpha'$ thus

$$\liminf_{m\to\infty}\Pr(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq m) \geq 1-\alpha.$$

S3.4 Proof of Proposition 1

Let

$$\begin{aligned}\epsilon_n &= \widehat{\gamma}^{\mathrm{cv}+} - \widehat{\gamma}^{\mathrm{cv}}_{\lambda^\star} \\ &= \widehat{\gamma}^{\mathrm{cv}+} - (\overline{V} - \lambda^\star(\overline{U} - \widehat{\mu})) \\ &= (\lambda^\star - \widehat{\lambda})(\overline{U} - \mu) - (\lambda^\star - \widehat{\lambda})(\widehat{\mu} - \mu).\end{aligned}$$

By the triangle inequality,

$$|\epsilon_n| \leq |\lambda^\star - \widehat{\lambda}|(|\overline{U} - \mu| + |\widehat{\mu} - \mu|).$$

By the Lemma S1 we have

$$|\widehat{\lambda} - \lambda^\star| = o\left(n^{-2/(2+\delta)}\right) \text{ a.s. as } n \to \infty$$

and, by the law of the iterated logarithm

$$|\overline{U} - \mu| = O\left(\sqrt{\frac{\log\log n}{n}}\right), \quad |\widehat{\mu} - \mu| = O\left(\sqrt{\frac{\log\log n}{n}}\right) \quad \text{a.s. as } n \to \infty. \tag{S45}$$

It follows that

$$|\epsilon_n| = O\left(\frac{1}{n^{2/(2+\delta)}}\sqrt{\frac{\log\log n}{n}}\right) = o\left(\frac{1}{\sqrt{n\log n}}\right) \quad \text{a.s. as } n \to \infty. \tag{S46}$$

S3.5 Proof of Proposition 2

We have

$$\widehat{\gamma}^{\mathrm{cv}}_{\lambda} = \overline{V} - \lambda(\overline{U} - \mu) + \lambda(\widehat{\mu} - \mu). \tag{S47}$$

The random variables $\overline{V} - \lambda(\overline{U} - \mu)$ and $\lambda(\widehat{\mu} - \mu)$ are independent and are both sample average of i.i.d. random variables with finite moment of order q for some $q = 2 + \delta > 2$. Note that $1 - 1/q > 1/2$. By KMT strong coupling (Theorem S8), there exist i.i.d. Gaussian random variables $(G_i^{(1)})_{i\geq 1}$ with mean γ and variance $\mathrm{var}(V - \lambda U)$ such that

$$\overline{V} - \lambda(\overline{U} - \mu) = \frac{1}{n}\sum_{i=1}^{n} G_i^{(1)} + o\left(\frac{1}{n^{1-1/q}}\right) \text{ a.s. as } n \to \infty.$$

If $r = 0$, then $n/N_n = O(1/n^{1-a})$. By the law of the iterated logarithm,

$$|\lambda(\widehat{\mu} - \mu)| = O\left(\sqrt{\frac{\log\log N_n}{N_n}}\right) = o\left(\frac{1}{\sqrt{n\log n}}\right) \text{ a.s. as } n \to \infty.$$

If $r > 0$, then, by Proposition S9, there exist i.i.d. Gaussian random variables $(G_i^{(2)})_{i\geq 1}$, independent of $(G_i^{(1)})_{i\geq 1}$, with mean 0 and variance $r\lambda^2\mathrm{var}(U)$ such that,

$$\lambda(\widehat{\mu} - \mu) = \frac{1}{n}\sum_{i=1}^{n} G_i^{(2)} + o\left(\frac{1}{n^{1-1/q}}\right) \text{ a.s. as } n \to \infty.$$

Setting $W_i^{\mathrm{cv}} = G_i^{(1)}$ if $r = 0$ and $W_i^{\mathrm{cv}} = G_i^{(1)} + G_i^{(2)}$ if $r > 0$ gives the Gaussian coupling (15) for $\nu_\lambda^{\mathrm{cv}}$, as $\frac{1}{n^{1-1/q}} = o\left(\frac{1}{\sqrt{n\log n}}\right)$. From this, using Equation (14), we deduce the coupling (16) for $\widehat{\gamma}^{\mathrm{cv}+}$, noting that

$$\nu^{\mathrm{cv}+} := \nu^{\mathrm{cv}}_{\lambda^\star} = \mathrm{var}(V)\left[1 - (1 - r)\rho^2_{U,V}\right]. \tag{S48}$$

We have

$$\frac{1}{n-2}\sum_{i=1}^{n}(V_i - \overline{V} - \lambda(U_i - \overline{U}))^2 \to \mathrm{var}(V - \lambda U) \text{ a.s.}$$

and

$$\frac{n\lambda^2}{N_n(N_n-1)}\sum_{j=1}^{N_n}(\widetilde{U}_j-\widehat{\mu})^2 \to r\lambda^2\mathrm{var}(U) \text{ a.s.}$$

therefore $\widehat{\nu}^{\mathrm{cv}}_{\lambda} \to \nu^{\mathrm{cv}}_{\lambda}$ a.s. Finally, we show that $\widehat{\nu}^{\mathrm{cv}+}$ is a consistent estimator of $\nu^{\mathrm{cv}+}$. Set $\delta_i = V_i - \overline{V} - \widehat{\lambda}(U_i - \overline{U})$. We have $\delta_i = V_i - \widehat{\alpha} - \widehat{\beta}U_i$ where $\widehat{\alpha} = \overline{V} - \widehat{\lambda}\overline{U}$ and $\widehat{\beta} = \widehat{\lambda}$ are the least squares estimates, minimising $\sum_{i=1}^{n}(V_i - \alpha - \beta U_i)^2$. It is well known [21, Theorem 2] that, for

$$(\alpha^\star, \beta^\star) = \arg\min_{\alpha,\beta}\mathbb{E}[(V-\alpha-\beta U)^2] = \left(\gamma - \mu\frac{\mathrm{cov}(U,V)}{\mathrm{var}(U)}, \frac{\mathrm{cov}(U,V)}{\mathrm{var}(U)}\right)$$

we have

$$\frac{1}{n-2}\sum_{i=1}^{n}\delta_i^2 \to \mathbb{E}[(V-\alpha^\star-\beta^\star U)^2] = \mathrm{var}(V)(1-\rho^2_{U,V}) \quad \text{a.s. as } n\to\infty. \tag{S49}$$

Additionally, by the strong law of large numbers,

$$\frac{n/N_n}{n-1}\sum_{i=1}^{n}(V_i-\overline{V})^2 \to r\mathrm{var}(V) \quad \text{a.s. as } n\to\infty. \tag{S50}$$

Hence,

$$\widehat{\nu}^{\mathrm{cv}+} \to (1-r)\mathrm{var}(V)(1-\rho^2_{U,V}) + r\mathrm{var}(V) = \mathrm{var}(V)(1-(1-r)\rho^2_{U,V})$$

almost surely as $n\to\infty$.

S3.6 Proof of Proposition 4

We apply Theorem 1 to the i.i.d. sequence $(\ell'_\theta(\widetilde{X}_i, f(\widetilde{X}_i)))_{i\geq 1}$, to obtain an AsympCS $(\widetilde{\mathcal{R}}_{\delta,\theta,i})_{i\geq 1}$ for m_θ with approximation rate $1/\sqrt{i\log i}$. The subsequence $(\mathcal{R}_{\delta,\theta,n})_{n\geq 1}$ with $\mathcal{R}_{\delta,\theta,n} = \widetilde{\mathcal{R}}_{\delta,\theta,N_n}$ is also an AsympCS for m_θ with approximation rate $1/\sqrt{n\log n}$. Asymptotic Type-I error control follows directly from Theorem 5.

S3.7 Proof of Proposition 5

In the PPI case, the proof follows from a direct application of Theorem 1 (non-assisted) or Theorem 3 (Bayes-assisted) to the sequence of i.i.d. random variables $(V_{\theta,i} - U_{\theta,i})_{i=1}^n$. In the PPI++ case, it follows from an application of Theorem 2 (non-assisted) or Theorem 4 (Bayes-assisted), together with Proposition 2, to the control variate estimator (23). Asymptotic Type-I error control follows directly from Theorem 5.

S4 Multivariate AsympCS

In this section, we discuss how the results developed in this paper for scalar θ can be extended to obtain asymptotic confidence regions for $\theta \in \mathbb{R}^d$.

We first provide the definitions of a multivariate confidence sequence and of an asymptotic spherical confidence sequence. Let $B(x,r) \subset \mathbb{R}^d$ denote the ball of radius r centered at x.

S4.1 Definitions

Definition S4. *(Confidence Sequence) Let $(\mathcal{C}_{\alpha,t})_{t\geq 1}$ be a sequence of random subsets of $\mathbb{R}^d$. For $\alpha \in (0,1)$, $(\mathcal{C}_{\alpha,t})_{t\geq 1}$ is a $1-\alpha$ confidence sequence for a fixed parameter $\mu \in \mathbb{R}^d$ if*

$$\Pr(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq 1) \geq 1-\alpha.$$

We now introduce the notion of an asymptotic spherical confidence sequence, inspired by [15, Section B.10].

Definition S5. *(Asymptotic Spherical Confidence Sequence) Let $\alpha \in (0,1)$ and $(a_t)_{t\geq 0}$ a real sequence such as $\lim_{t\to\infty} a_t = 0$. Let $(\widehat{\mu}_t)_{t\geq 1}$ be a consistent sequence of estimators of μ. The sequence of random balls $(\mathcal{C}_{\alpha,t})_{t\geq 1}$, with $\mathcal{C}_{\alpha,t} = B(\widehat{\mu}_t, R_t)$ and $R_t > 0$, is said to be an asymptotic spherical confidence sequence with (little-o) approximation rate a_t, if there exists a (usually unknown) confidence sequence $(\mathcal{C}^\star_{\alpha,t})_{t\geq 1}$, with $\mathcal{C}^\star_{\alpha,t} = B(\widehat{\mu}_t, R^\star_t)$, such that*

$$\Pr(\mu \in \mathcal{C}^\star_{\alpha,t} \text{ for all } t \geq 1) \geq 1-\alpha$$

and

$$|R_t - R^\star_t| = o(a_t) \text{ a.s. as } t \to +\infty.$$

S4.2 Nonasymptotic Bayes-assisted CS for i.i.d. Gaussian random vectors

Let $Y_1, Y_2, \ldots,$ be i.i.d. Gaussian random vectors with mean $\mu \in \mathbb{R}^d$ and known d-by-d positive definite covariance matrix Σ. Let π be some prior on $\Sigma^{-1/2}\mu$ and define $\eta_t(z) = \int \mathcal{N}(z;\zeta, I_d/t)\pi(\zeta)d\zeta$ where I_d denotes the d-by-d identity matrix. By the method of mixtures and Ville's inequality (similarly as in Theorem S11), the sequence of ellipsoid regions defined by

$$\mathcal{C}_{\alpha,t}(\overline{Y}_t, \Sigma; \pi) = \left\{ \mu \in \mathbb{R}^d \mid \|\Sigma^{-1/2}(\mu - \overline{Y}_t)\| \leq \frac{1}{\sqrt{t}} \sqrt{\log\left(\frac{t^d}{(2\pi)^d \alpha^2 \eta_t(\Sigma^{-1/2}\overline{Y}_t)^2}\right)} \right\}$$

forms a $(1-\alpha)$ confidence sequence for μ. One could also consider spherical confidence intervals using $\Lambda_{\max}(\Sigma)$, the maximum eigenvalue of Σ, similarly to what is done in [15, Section B10] for non-assisted confidence regions. In this case, the corresponding (more conservative) $(1-\alpha)$ confidence sequence for μ is

$$\mathcal{C}^{\texttt{mBA}}_{\alpha,t}(\overline{Y}_t, \Sigma; \pi) = \left\{ \mu \in \mathbb{R}^d \mid \|\mu - \overline{Y}_t\| \leq \frac{\sqrt{\Lambda_{\max}(\Sigma)}}{\sqrt{t}} \sqrt{\log\left(\frac{t^d}{(2\pi)^d \alpha^2 \eta_t(\Sigma^{-1/2}\overline{Y}_t)^2}\right)} \right\}. \tag{S51}$$

S4.3 Multivariate extension of Theorems 3 and 4

To illustrate that our results extend to the multivariate case, we provide multivariate (and tight) versions of Theorem 4 and Theorem 3.

Theorem S12. *(Multivariate version of Theorem 4) Let $(\widehat{\mu}_t)_{t\geq 1}$ be a consistent sequence of estimators of μ. Assume that there exists a sequence of i.i.d. Gaussian vectors with mean μ and positive definite covariance matrix Σ such that, a.s. as $t \to \infty$,*

$$\widehat{\mu}_t = \frac{1}{t}\sum_{i=1}^t W_i + \varepsilon_t \quad \text{where} \quad \varepsilon_t = o\left(\frac{1}{\sqrt{t\log t}}\right). \tag{S52}$$

Let $(\widehat{\Sigma}_t)_{t\geq 1}$ be a consistent sequence of estimators of Σ such that $\|\widehat{\Sigma}_t - \Sigma\| = o(1/\log t)$ a.s. where $\|\cdot\|$ is the spectral norm, and π be a continuous and proper prior density on $\mathbb{R}^d$. Then, $\mathcal{C}^{\texttt{mBA}}_{\alpha,t}(\widehat{\mu}_t, \widehat{\Sigma}_t; \pi)$ forms a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t\log t}$.

Proof. (Theorem S12) By hypothesis, we have, almost surely,

$$\widehat{\mu}_t = \overline{W}_t + \varepsilon_t,$$

where $\varepsilon_t = o\left(\frac{1}{\sqrt{t\log t}}\right)$ and $\overline{W}_t = \frac{1}{t}\sum_{i=1}^t W_i$. As stated in Appendix S4.2, the sequence of balls $\mathcal{C}^{\texttt{mBA}}_{\alpha,t}(\overline{W}_t, \Sigma; \pi) = \mathcal{C}^{\texttt{mBA}}_{\alpha,t}(\widehat{\mu}_t - \varepsilon_t, \Sigma; \pi)$ forms an exact $(1-\alpha)$ CS for μ. Let

$$R^\star_t = \frac{\sqrt{\Lambda_{\max}(\Sigma)}}{\sqrt{t}} \sqrt{\log\left(\frac{t^d}{(2\pi)^d \alpha^2 \eta_t(\Sigma^{-1/2}(\widehat{\mu}_t - \varepsilon_t))^2}\right)} + \|\varepsilon_t\|.$$

As $B(\widehat{\mu}_t, R^\star_t) \supseteq \mathcal{C}^{\texttt{mBA}}_{\alpha,t}(\widehat{\mu}_t - \varepsilon_t, \Sigma; \pi)$, the sequence of random balls $B(\widehat{\mu}_t, R^\star_t)$ also forms an exact $(1-\alpha)$ CS for μ. Define $\mathcal{C}^{\texttt{mBA}}_{\alpha,t}(\widehat{\mu}_t, \widehat{\Sigma}_t; \pi) = B(\widehat{\mu}_t, R_t)$, where

$$R_t = \frac{\sqrt{\Lambda_{\max}(\widehat{\Sigma}_t)}}{\sqrt{t}} \sqrt{\log\left(\frac{t^d}{(2\pi)^d \alpha^2 \eta_t(\widehat{\Sigma}_t^{-1/2}\widehat{\mu}_t)^2}\right)}.$$

By the Courant-Fischer theorem, we have,

$$|\Lambda_{\max}\left(\widehat{\Sigma}_t\right) - \Lambda_{\max}(\Sigma)| \leq \|\widehat{\Sigma}_t - \Sigma\| = o(1/\log t) \text{ a.s.}$$

Additionally, by Lemma S2, we have

$$\eta_t(\widehat{\Sigma}_t^{-1/2}\widehat{\mu}_t) \to \pi\left(\Sigma^{-1/2}\mu\right) \text{ and } \eta_t(\Sigma^{-1/2}(\widehat{\mu}_t - \varepsilon_t)) \to \pi\left(\Sigma^{-1/2}\mu\right) \text{ a.s. as } t \to \infty.$$

It follows that

$$\begin{aligned} R_t - (R_t^\star - \|\varepsilon_t\|) &= \frac{R_t^2 - (R_t^\star - \|\varepsilon_t\|)^2}{R_t + (R_t^\star - \|\varepsilon_t\|)} \\ &\sim \frac{(\Lambda_{\max}(\widehat{\Sigma}_t) - \Lambda_{\max}(\Sigma))\frac{\log t^d}{t}}{\sqrt{\Lambda_{\max}(\Sigma)\frac{\log t^d}{t}}} \\ &= o(1/\sqrt{t\log t}) \end{aligned}$$

a.s. Then $|R_t - R_t^\star| = o(1/\sqrt{t\log t})$ and so $\mathcal{C}_{\alpha,t}^{\texttt{mBA}}(\widehat{\mu}_t, \widehat{\Sigma}_t; \pi)$ is a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t\log t}$. □

Theorem S13. *(Multivariate version of Theorem 3) Let $(Y_t)_{t\geq 1}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$ with mean μ and such that $\mathbb{E}\|Y_1\|^{2+\delta} < \infty$ for some $\delta > 0$. Then, $\mathcal{C}_{\alpha,t}^{\texttt{mBA}}(\overline{Y}_t, \widehat{\Sigma}_t; \pi)$ is a $(1-\alpha)$–AsympCS with approximation rate $1/\sqrt{t\log t}$, where $\overline{Y}_t$ is the sample mean and $\widehat{\Sigma}_t$ the sample covariance.*

Proof. (Theorem S13) By the strong law of large numbers, $\overline{Y}_t$ and $\widehat{\Sigma}_t$ are consistent estimators of μ and Σ, respectively. By the multivariate KMT coupling due to Einmahl [5] (see Theorem S9), there exists a sequence of i.i.d. Gaussian random vectors $(W_i)_{i\geq 1}$ with mean μ and covariance matrix Σ such that

$$\overline{Y}_t = \frac{1}{t}\sum_{i=1}^{t} W_i + \varepsilon_t \quad \text{where} \quad \varepsilon_t = o\left(\frac{1}{t^{1-1/(2+\delta)}}\right) = o\left(\frac{1}{\sqrt{t\log t}}\right).$$

The result then follows from Theorem S12. □

All other results can be extended to the multivariate case in a similar manner. In the case of Theorem 1 and Theorem 2, we require a multivariate, non-assisted, exact confidence sequence for i.i.d. Gaussian random variables with known variance. For any $\rho > 0$, Waudby-Smith et al. [6, Equation (29)] propose to use

$$\mathcal{C}_{\alpha,t}^{\texttt{mNA}} := \left\{\mu \in \mathbb{R}^d : \left\|\overline{Y}_t - \mu\right\| < \sqrt{\frac{\Lambda_{\max}(\widehat{\Sigma}_t)\cdot 9d}{2}\cdot\frac{1+t\rho^2}{t^2\rho^2}\cdot\left[2+\log\left(\frac{\sqrt{1+t\rho^2}}{\alpha}\right)\right]}\right\},$$

although alternative constructions are possible. To extend our results on control variates and PPI to the multivariate setting, the proofs can be adapted accordingly. In doing so, we will require multivariate versions of the Marcinkiewicz-Zygmund strong law of large numbers (see Ledoux and Talagrand [22, Theorem 7.9]) and of the law of the iterated logarithm (see Koval [23, Corollary 1]).

S5 Derivations for prediction-powered mean estimation

Throughout the main text, we report expressions of quantities related to the construction of prediction-powered AsympCS for mean estimation. Here, we explicitly derive those expressions.

The convex loss associated with the estimand $\theta^\star = \mathbb{E}[Y]$ is the squared loss $\ell_\theta(x,y) = (\theta - y)^2/2$, whose subgradient with respect to θ is given by $\ell'_\theta(x,y) = \theta - y$. As a result of this, the measure of fit m_θ takes the form

$$m_\theta = \theta - \mathbb{E}\left[f(X)\right], \tag{S53}$$

whereas the rectifier Δ_θ is given by

$$\Delta_\theta = \mathbb{E}\left[f(X) - Y\right]. \tag{S54}$$

In particular, notice that, in the case of mean estimation, the rectifier is independent of θ, i.e. $\Delta_\theta = \Delta_0$ for all θ.

As discussed in Section 4, PPI uses the sample mean as an estimator of m_θ, which in this case is given by

$$\widehat{m}_{\theta,n} = \theta - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j). \tag{S55}$$

For Δ_0, either the PPI estimator $\widehat{\Delta}^{\mathsf{PP}}_{0,n}$ (20) or the PPI++ estimator $\widehat{\Delta}^{\mathsf{PP+}}_{0,n}$ (23) may be used. In the case of mean estimation, these are given by

$$\widehat{\Delta}^{\mathsf{PP}}_{0,n} = \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - Y_i\right), \tag{S56}$$

$$\widehat{\Delta}^{\mathsf{PP+}}_{0,n} = \widehat{\Delta}^{\mathsf{PP}}_{n} - (\widehat{\lambda}_{\theta,n} - 1)\left(\frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) \tag{S57}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\lambda}_{\theta,n} f(X_i) - Y_i\right) - (\widehat{\lambda}_{\theta,n} - 1)\frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j), \tag{S58}$$

where

$$\widehat{\lambda}_{\theta,n} = \frac{\widehat{\mathrm{cov}}\left((\ell'_\theta(X_i, Y_i), \ell'_\theta(X_i, f(X_i)))_{i=1}^n\right)}{\widehat{\mathrm{var}}\left((\ell'_\theta(X_i, f(X_i)))_{i=1}^n\right)} = \frac{\widehat{\mathrm{cov}}((Y_i, f(X_i))_{i=1}^n)}{\widehat{\mathrm{var}}((f(X_i))_{i=1}^n)}. \tag{S59}$$

Again, the control-variate parameter $\widehat{\lambda}_{\theta,n}$ does not depend on θ, i.e. $\widehat{\lambda}_{\theta,n} = \widehat{\lambda}_{0,n}$ for all θ.

Given $\widehat{m}_{\theta,n}$ and an estimator $\widehat{\Delta}_{0,n}$ of Δ_0, the associated prediction-powered estimator of $\theta^\star$ is found by solving, in θ, the equation

$$\widehat{g}_{\theta,n} = \widehat{m}_{\theta,n} + \widehat{\Delta}_{0,n}.$$

For the two estimators of Δ_0 discussed above, this quantity takes the form

$$\widehat{g}^{\mathsf{PP}}_{\theta,n} = \theta - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j) + \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - Y_i\right) \tag{S60}$$

$$= \theta - \frac{1}{n}\sum_{i=1}^{n} Y_i + \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j)\right), \tag{S61}$$

$$\widehat{g}^{\mathsf{PP+}}_{\theta,n} = \theta - \widehat{\lambda}_{0,n}\frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j) + \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\lambda}_{0,n} f(X_i) - Y_i\right) \tag{S62}$$

$$= \theta - \frac{1}{n}\sum_{i=1}^{n} Y_i + \widehat{\lambda}_{0,n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j)\right), \tag{S63}$$

whose zeroes are given by

$$\widehat{\theta}^{\mathsf{PP}}_{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j)\right), \tag{S64}$$

$$\widehat{\theta}^{\mathsf{PP+}}_{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\lambda}_{0,n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j)\right), \tag{S65}$$

which match the expressions in Equations (22) and (25), respectively.

As discussed in Section 1, a prediction-powered $(1-\alpha)$ AsympCS $(\mathcal{C}^{\text{avpp}}_{\alpha,n})_{n\geq 1}$ for θ is defined through Equation (4) by first constructing a valid AsympCS $(\mathcal{C}^{g}_{\alpha,\theta,n})_{n\geq 1}$ for g_θ.

Section 5.1 defines valid AsympCS for g_θ that incorporate no prior information. In particular, $\mathcal{C}^{g}_{\alpha,\theta,n}$ is constructed as

$$\mathcal{C}^{g}_{\alpha,\theta,n} = \mathcal{C}^{\texttt{NA}}_{\alpha,t}(\widehat{g}_{\theta,n}, \widehat{\sigma}^{g}_{\theta,n}; \rho) \tag{S66}$$

$$= \left[\widehat{g}_{\theta,n} \pm \frac{\widehat{\sigma}^{g}_{\theta,n}}{\sqrt{n}}\sqrt{\left(1+\frac{1}{n\rho^2}\right)\log\left(\frac{n\rho^2+1}{\alpha^2}\right)}\right], \tag{S67}$$

where $\mathcal{C}^{\texttt{NA}}_{\alpha,n}$ is defined in Theorem 1, $\widehat{g}_{\theta,n}$ is either the PPI estimator $\widehat{g}^{\texttt{PP}}_{\theta,n}$ or the PPI++ estimator $\widehat{g}^{\texttt{PP+}}_{\theta,n}$, and $(\widehat{\sigma}^{g}_{\theta,n})^2$ is the corresponding variance estimator, as defined in Proposition 3. More specifically, under the squared loss, for $\widehat{g}^{\texttt{PP}}_{\theta,n}$, we have

$$(\widehat{\sigma}^{g}_{\theta,n})^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - f(X_i) - \frac{1}{n}\sum_{k=1}^{n}(Y_k - f(X_k))\right)^2 \tag{S68}$$

$$+ \frac{n/N_n}{N_n-1}\sum_{j=1}^{N_n}\left(f(\widetilde{X}_j) - \frac{1}{N_n}\sum_{k=1}^{N_n} f(\widetilde{X}_k)\right)^2, \tag{S69}$$

while, for $\widehat{g}^{\texttt{PP+}}_{\theta,n}$, we have

$$(\widehat{\sigma}^{g}_{\theta,n})^2 = \frac{1-n/N_n}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{\lambda}_{0,n}f(X_i) - \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \widehat{\lambda}_{0,n}f(X_k)\right)\right)^2 \tag{S70}$$

$$+ \frac{n/N_n}{n-1}\sum_{i=1}^{n}\left(Y_i - \frac{1}{n}\sum_{k=1}^{n}Y_k\right)^2. \tag{S71}$$

Given the specific form of $\widehat{g}_{\theta,n}$ under the squared loss, $\mathcal{C}^{\text{avpp}}_{\alpha,n}$ can be explicitly expressed as

$$\mathcal{C}^{\text{avpp}}_{\alpha,n} = \left\{\theta \mid 0 \in \mathcal{C}^{g}_{\alpha,\theta,n}\right\} \tag{S72}$$

$$= \left[\widehat{\theta}_n \pm \frac{\widehat{\sigma}^{g}_{0,n}}{\sqrt{n}}\sqrt{\left(1+\frac{1}{n\rho^2}\right)\log\left(\frac{n\rho^2+1}{\alpha^2}\right)}\right], \tag{S73}$$

which is an interval, and where $\widehat{\theta}_n$ is either the PPI estimator $\widehat{\theta}^{\texttt{PP}}_n$ or the PPI++ estimator $\widehat{\theta}^{\texttt{PP+}}_n$.

Similarly, Section 5.2 defines valid AsympCS for g_θ that incorporate prior information by means of a zero-mean prior π on Δ_θ. In particular, for $\delta \in (0,\alpha)$, Proposition 4 first construct a standard $(1-\delta)$ AsympCS $\mathcal{R}_{\delta,\theta,n}$ for m_θ, which in the case of mean estimation takes the form

$$\mathcal{R}_{\delta,\theta,n} = \mathcal{C}^{\texttt{NA}}_{\delta,n}(\widehat{m}_{\theta,n}, \widehat{\sigma}^{f}_{\theta,n}; \rho) \tag{S74}$$

$$= \left[\theta - \frac{1}{N_n}\sum_{j=1}^{N_n} f(\widetilde{X}_j) \pm \frac{\widehat{\sigma}^{f}_{\theta,n}}{\sqrt{N_n}}\sqrt{\left(1+\frac{1}{N_n\rho^2}\right)\log\left(\frac{N_n\rho^2+1}{\delta^2}\right)}\right], \tag{S75}$$

where $(\widehat{\sigma}^{f}_{\theta,n})^2$ is the sample variance of $(\ell'_\theta(\widetilde{X}_i, f(\widetilde{X}_i)))_{i=1}^{N_n}$, that is

$$(\widehat{\sigma}^{f}_{\theta,n})^2 = \widehat{\text{var}}((\ell'_\theta(\widetilde{X}_i, f(\widetilde{X}_i)))_{i=1}^{N_n}) \tag{S76}$$

$$= \frac{1}{N_n-1}\sum_{j=1}^{N_n}\left(f(\widetilde{X}_j) - \frac{1}{N_n}\sum_{k=1}^{N_n} f(\widetilde{X}_k)\right)^2. \tag{S77}$$

Next, Proposition 5 constructs a Bayes-assisted $(1-(\alpha-\delta))$ AsympCS $\mathcal{T}_{\alpha-\delta,\theta,n}$ for Δ_θ, which under the squared loss takes the form

$$\mathcal{T}_{\alpha-\delta,\theta,n} = \mathcal{C}^{\texttt{BA}}_{\alpha-\delta,n}(\widehat{\Delta}_{0,n}, \widehat{\sigma}^{\Delta}_{\theta,n}; \pi) \tag{S78}$$

$$= \left[\widehat{\Delta}_{0,n} \pm \frac{\widehat{\sigma}^{\Delta}_{\theta,n}}{\sqrt{n}}\sqrt{\log\left(\frac{n(2\pi(\alpha-\delta)^2)^{-1}}{\eta_n(\widehat{\Delta}_{0,n}/\widehat{\sigma}^{\Delta}_{\theta,n})^2}\right)}\right], \tag{S79}$$

where $\mathcal{C}^{\texttt{BA}}_{\alpha-\delta,n}$ and η_n are defined in Theorem 3, $\widehat{\Delta}_{0,n}$ is either the PPI estimator $\widehat{\Delta}^{\texttt{PP}}_{0,n}$ or the PPI++ estimator $\widehat{\Delta}^{\texttt{PP+}}_{0,n}$, and $(\widehat{\sigma}^{\Delta}_{\theta,n})^2$ is the corresponding variance estimator, as defined in Proposition 5. In the case of mean estimation, $(\widehat{\sigma}^{\Delta}_{\theta,n})^2$ takes the form

$$(\widehat{\sigma}^{\Delta}_{\theta,n})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - f(X_i) - \frac{1}{n}\sum_{k=1}^{n}(Y_k - f(X_k))\right)^2, \tag{S80}$$

for PPI, and is given by

$$(\widehat{\sigma}^{\Delta}_{\theta,n})^2 = \frac{1-n/N_n}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{\lambda}_{0,n}f(X_i) - \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \widehat{\lambda}_{0,n}f(X_k)\right)\right)^2 \tag{S81}$$

$$+ \frac{n/N_n}{n-1}\sum_{i=1}^{n}\left(Y_i - f(X_i) - \frac{1}{n}\sum_{k=1}^{n}(Y_k - f(X_k))\right)^2 \tag{S82}$$

for PPI++. Finally, $\mathcal{T}_{\alpha-\delta,\theta,n}$ and $\mathcal{R}_{\delta,\theta,n}$ are combined via a Minkowski sum to construct a valid $(1-\alpha)$ AsympCS for g_θ as

$$\mathcal{C}^{g}_{\alpha,\theta,n} = \mathcal{T}_{\alpha-\delta,\theta,n} + \mathcal{R}_{\delta,\theta,n} \tag{S83}$$

$$= \left[\widehat{g}_{\theta,n} \pm \left\{\frac{\widehat{\sigma}^{\Delta}_{\theta,n}}{\sqrt{n}}\sqrt{\log\left(\frac{n(2\pi(\alpha-\delta)^2)^{-1}}{\eta_n(\widehat{\Delta}_{0,n}/\widehat{\sigma}^{\Delta}_{\theta,n})^2}\right)} + \frac{\widehat{\sigma}^{f}_{\theta,n}}{\sqrt{N_n}}\sqrt{\left(1+\frac{1}{N_n\rho^2}\right)\log\left(\frac{N_n\rho^2+1}{\delta^2}\right)}\right\}\right], \tag{S84}$$

where $\widehat{g}_{\theta,n}$ is either the PPI estimator $\widehat{g}^{\texttt{PP}}_{\theta,n}$ or the PPI++ estimator $\widehat{g}^{\texttt{PP+}}_{\theta,n}$. As above, the form of $\widehat{g}_{\theta,n}$ for mean estimation allows expressing $\mathcal{C}^{\text{avpp}}_{\alpha,n}$ explicitly as

$$\mathcal{C}^{\text{avpp}}_{\alpha,n} = \left\{\theta \mid 0 \in \mathcal{C}^{g}_{\alpha,\theta,n}\right\} \tag{S85}$$

$$= \left[\widehat{\theta}_n \pm \left\{\frac{\widehat{\sigma}^{\Delta}_{\theta,n}}{\sqrt{n}}\sqrt{\log\left(\frac{n(2\pi(\alpha-\delta)^2)^{-1}}{\eta_n(\widehat{\Delta}_{0,n}/\widehat{\sigma}^{\Delta}_{\theta,n})^2}\right)} + \frac{\widehat{\sigma}^{f}_{\theta,n}}{\sqrt{N_n}}\sqrt{\left(1+\frac{1}{N_n\rho^2}\right)\log\left(\frac{N_n\rho^2+1}{\delta^2}\right)}\right\}\right], \tag{S86}$$

which matches the expression in Equation (27), and where $\widehat{\theta}_n$ is either the PPI estimator $\widehat{\theta}^{\texttt{PP}}_n$ or the PPI++ estimator $\widehat{\theta}^{\texttt{PP+}}_n$.

S6 Experimental details

S6.1 Implementation

Code implementing our method is written in Python and made available at `https://github.com/stefanocortinovis/ppi-cs`. All experiments were run locally on an Apple Silicon M4 Pro CPU with 24GB of memory.

S6.2 Datasets

Here we briefly describe each dataset used for the real data experiments in Section 6.2. The FLIGHTS dataset was downloaded from Kaggle[2], while all the others are available as part of the `ppi-python` package[3].

Flights. For each of 103333 economy class flight tickets, the FLIGHTS dataset reports the ticket price ($Y_i \in \mathbb{R}$), as well as the prediction of a gradient-boosted tree for Y_i ($f(X_i) \in \mathbb{R}$). The goal is to estimate the average price of a flight, i.e. $\theta^\star = \mathbb{E}[Y] \in \mathbb{R}$.

[2] `https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction`

[3] `https://pypi.org/project/ppi-python/`

Forest. For each of 1596 parcels of land in the Amazon rainforest [24], the FOREST dataset reports whether the parcel has been subject to deforestation ($Y_i \in \{0, 1\}$), as well as the prediction of a gradient-boosted tree model for the probability of Y_i being equal to one ($f(X_i) \in [0, 1]$). The goal is to estimate the fraction of Amazon rainforest lost to deforestation, i.e. $\theta^\star = \mathbb{E}[Y] \in [0, 1]$.

Galaxies. For each of 16743 images from the Galaxy Zoo 2 initiative [25], the GALAXIES dataset reports whether the galaxy has spiral arms ($Y_i \in \{0, 1\}$), as well as the prediction of a ResNet50 model [26] for the probability of Y_i being equal to one ($f(X_i) \in [0, 1]$). The goal is to estimate the fraction of galaxies with spiral arms, i.e. $\theta^\star = \mathbb{E}[Y] \in [0, 1]$.

Census. For each of 380091 individuals from the 2019 California census, the CENSUS dataset reports the individual's age ($X_i \in \mathbb{R}$) and yearly income ($Y_i \in \mathbb{R}$), as well as the prediction of a gradient-boosted tree model trained on the previous year's data for Y_i ($f(X_i) \in \mathbb{R}$). The goal is to estimate the ordinary least squares (OLS) regression coefficient when regressing income on age. We preprocess the data by excluding non-positive incomes ($Y_i \leq 0$ or $f(X_i) \leq 0$), and applying a log-transformation to both the response Y_i and the prediction $f(X_i)$. This results in a dataset of 268118 individuals.

Healthcare. For each of 318215 individuals from the 2019 California census, the HEALTHCARE dataset reports the individual's yearly income ($X_i \in \mathbb{R}$) and whether they have health insurance ($Y_i \in \{0, 1\}$), as well as the prediction of a gradient-boosted tree model trained on the previous year's data for the probability of Y_i being equal to one ($f(X_i) \in [0, 1]$). The goal is to estimate the logistic regression coefficient when regressing health insurance status on income. As above, we preprocess the data by excluding non-positive incomes ($X_i \leq 0$), and applying a log-transformation to the covariate X_i. This results in a dataset of 270214 individuals.

Genes. For each of 61150 gene promoter sequences [27], the GENES dataset reports the expression level of the gene induced by the promoter ($Y_i \in \mathbb{R}$), as well as the prediction of a transformer model for Y_i ($f(X_i) \in \mathbb{R}$). The goal is to estimate the median expression level across sequences. We preprocess the data by applying a log-transformation to the response Y_i.

S6.3 Predictor performance

Table S1 reports the performance of the predictors used for each real data dataset above, measured in terms of normalised root mean squared error (NRMSE) for regression (R) tasks (FLIGHTS, CENSUS, GENES) and cross-entropy (CE) for the binary classification (C) tasks (FOREST, GALAXIES, HEALTHCARE). While we report these for completeness, we emphasise that non-assisted PPI improves

Table S1: Predictor performance on real data datasets.

Dataset	Flights	Forest	Galaxies	Census	Healthcare	Genes
Task	R	C	C	R	C	R
Performance	0.20	0.31	0.29	0.11	0.36	0.33

over classical inference in the presence of correlation between the predictions and the true labels, regardless of the absolute predictive performance (see e.g. Figure 2). On the other hand, while knowledge on the predictive performance can be used to choose a suitable prior for Bayes-assisted PPI, the latter is placed on the rectifier Δ_θ, which depends on the downstream inference task, thereby making the relationship between predictive performance and efficiency gains less direct.

S6.4 AsympCS hyperparameters

Here we discuss the hyperparameters of the PPI and PPI++ AsympCS procedures defined in Section 5.

The non-assisted prediction-powered AsympCS discussed in Section 5.1 requires the specification of the parameter ρ for Equation (6) in Theorem 1. As mentioned at the end of Section 3.1, ρ can be chosen so as to minimise the width of the interval at a specified time. In particular, as shown in

Waudby-Smith et al. [6, Appendix B.2], setting

$$\rho = \sqrt{\frac{-W_{-1}(-\alpha^2 \exp(-1)) - 1}{t^\star}}, \tag{S87}$$

where W_{-1} is the lower branch of the Lambert W function, minimises the width of the interval at time $t^\star$.

The Bayes-assisted prediction-powered AsympCS discussed in Section 5.2 requires the specification of both the parameter ρ used for the non-assisted AsympCS for the measure of fit m_θ (Proposition 4) and the scale parameter τ of the prior π for the Bayes-assisted AsympCS for the rectifier Δ_θ (Proposition 5). While the former can be chosen as above, the same approach does not work for the latter, as the prior scale that minimises the width of the Bayes-assisted interval at a specified time t depends on the observed value of $\overline{Y}_t/\widehat{\sigma}_t$ in Equation (9) of Theorem 3. Instead, we propose the following heuristic for choosing τ. If $Z_1, Z_2, \ldots | \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, then the posterior mean $\mathbb{E}[\mu | Z_1, \ldots, Z_t]$ after t observations under a Gaussian prior $\mu \sim \mathcal{N}(\mu_0, \tau^2)$ is given by

$$\mathbb{E}[\mu | Z_1, \ldots, Z_t] = \frac{1}{1 + t\tau^2}\mu_0 + \frac{t\tau^2}{1 + t\tau^2}\overline{Z}_t, \tag{S88}$$

where the two terms on the right-hand side measure the influence of the prior and the data on the posterior mean, respectively. Noticing that the Gaussian likelihood leading to the posterior mean (S88) is of the same form as the one implicitly used for the construction of Bayes-assisted AsympCS (see e.g. Equation (8)), we choose τ so that the prior and the data have the same influence on the posterior mean (S88) at time $t^\star$, i.e.

$$\tau = \frac{1}{\sqrt{t^\star}}. \tag{S89}$$

Appendix S7 reports the hyperparameter value used for each experiment in terms of $t^\star$. Notice that, as discussed in Section 6, the initial N_n is set large enough to rule out any uncertainty on the measure of fit m_θ. As a result of this, the only hyperparameter that needs to be chosen for the Bayes-assisted procedure is the prior scale τ.

S7 Additional experimental results

Additional experimental results to complement Section 6 are presented here. Legend names are as in Section 6.

S7.1 Synthetic data

S7.1.1 Noisy predictions

For this experiment, we set $t^\star = 500$ (see Appendix S6.4) for all methods.

Figure S5 reports the average cumulative miscoverage rate for the results shown in Figure 1. As desired, the cumulative miscoverage rate lies below the threshold α for all n.

Figure S6 shows the performance of the Bayes-assisted prediction-powered AsympCS procedures under a Gaussian prior on the noisy predictions experiment in Section 6.1. The results are consistent with those presented in Section 6.1. In particular, also with Bayes-assistance, PPI++ easily adapts to increasing noise levels, while standard PPI fails to do so. Moreover, in this case, Bayes-assisted PPI++ outperforms the non-assisted version across all noise levels. This is due to the fact that, in this experiment, the predictions from f, while noisy, are unbiased for all values of σ_Y. As a result of this, the zero-mean prior used by the Bayes-assisted procedures is well specified, and any additional shrinkage performed by the latter is beneficial.

S7.1.2 Biased predictions

For this experiment, we set $t^\star = 500$ (see Appendix S6.4) for all methods shown in Figures 2 and S7. The values of $t^\star$ used for Figure S8 are reported in the figure legend.

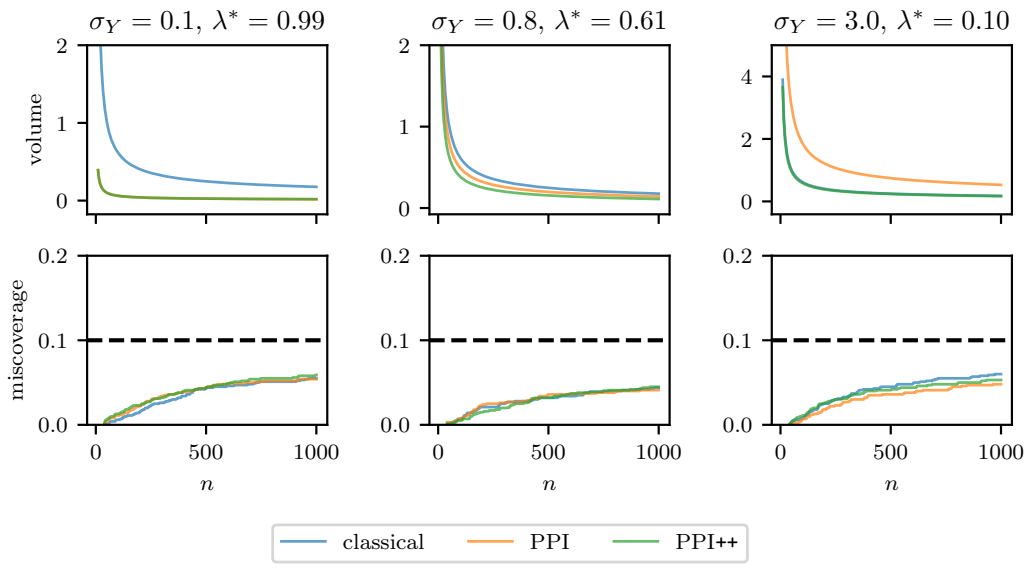

Figure S5: Noisy predictions study. The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for noise levels $\sigma_Y = 0.1, 0.8, 3.0$.

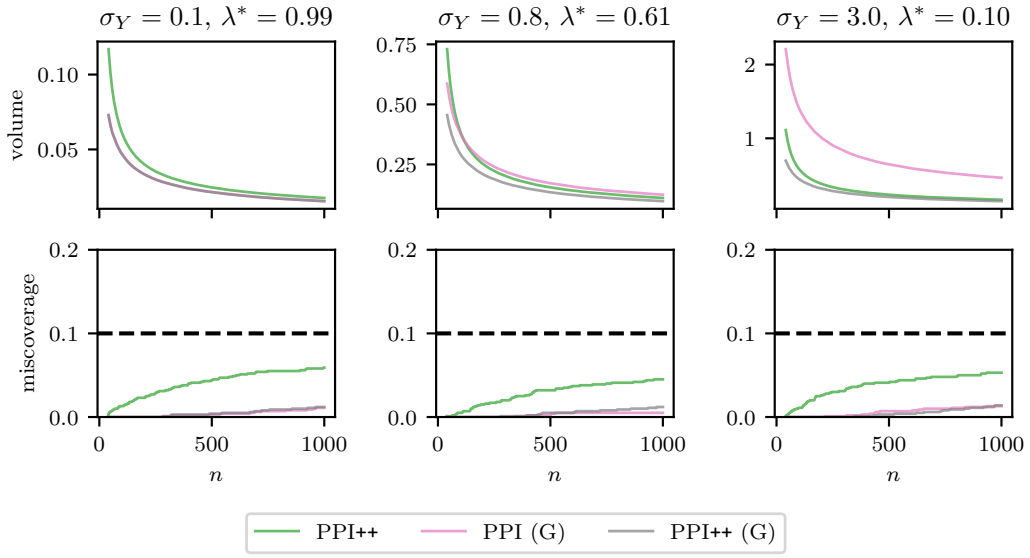

Figure S6: Noisy predictions study with Gaussian prior. The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for noise levels $\sigma_Y = 0.1, 0.8, 3.0$. Results for non-assisted PPI++ are shown for reference.

Figure S7 reports the average cumulative miscoverage rate for the results shown in Figure 2 at $\upsilon = 0$. As discussed in Section 6.1, the cumulative miscoverage rate increases slightly as we decrease df, but remains below the threshold α for all n, as desired.

Figure S8 repeats the simulation of Figure S7 for different values of $t^\star$, which affects the procedure-specific hyperparameters as discussed in Appendix S6.4. For non-assisted PPI, $t^\star$ represents the time at which the procedure's interval width is minimised. Therefore, as expected, increasing $t^\star$ above 100 leads to larger intervals for $n = 100$. However, the qualitative behaviour of the non-assisted methods as υ varies is the same across all values of $t^\star$: their volume remain constant across bias

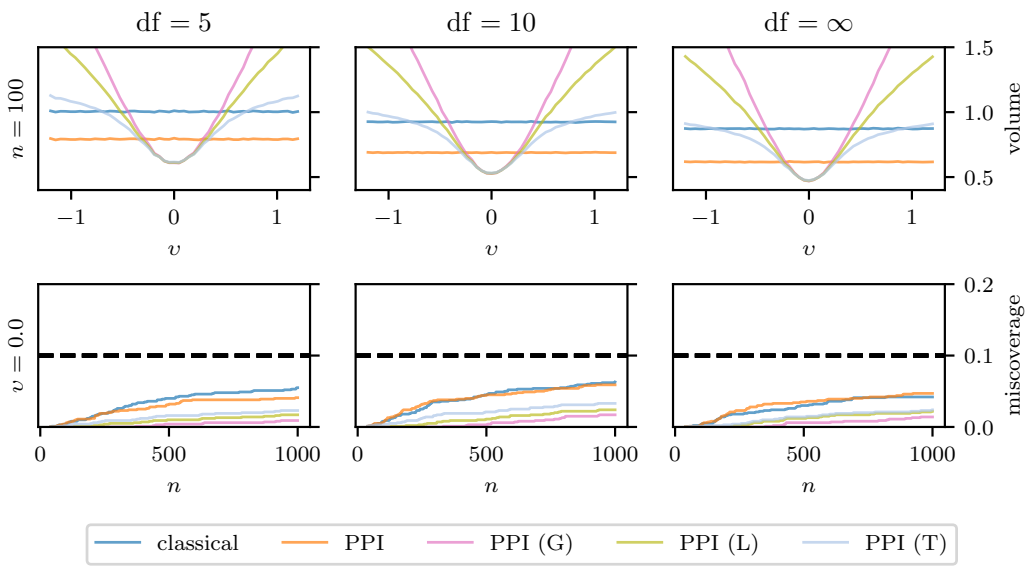

Figure S7: Biased predictions study. The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for $\mathrm{df} = 5, 10, \infty$.

levels, reflecting the lack of prior information. On the other hand, for Bayes-assisted methods, a larger $t^\star$ implies a smaller prior scale τ. As a result of this, increasing $t^\star$ above 100 leads to stronger prior influence at $n = 100$. In particular, a large $t^\star$ results in slightly smaller intervals for $v \approx 0$, but larger intervals for $|v| \gg 0$. When comparing the results across different priors, the results are consistent with those in Figure 2: the interval volume under heavier-tailed priors, such as the Laplace and Student-t priors, grow at a lower rate with $|v|$ compared to the Gaussian prior, thereby offering greater robustness to prior misspecification.

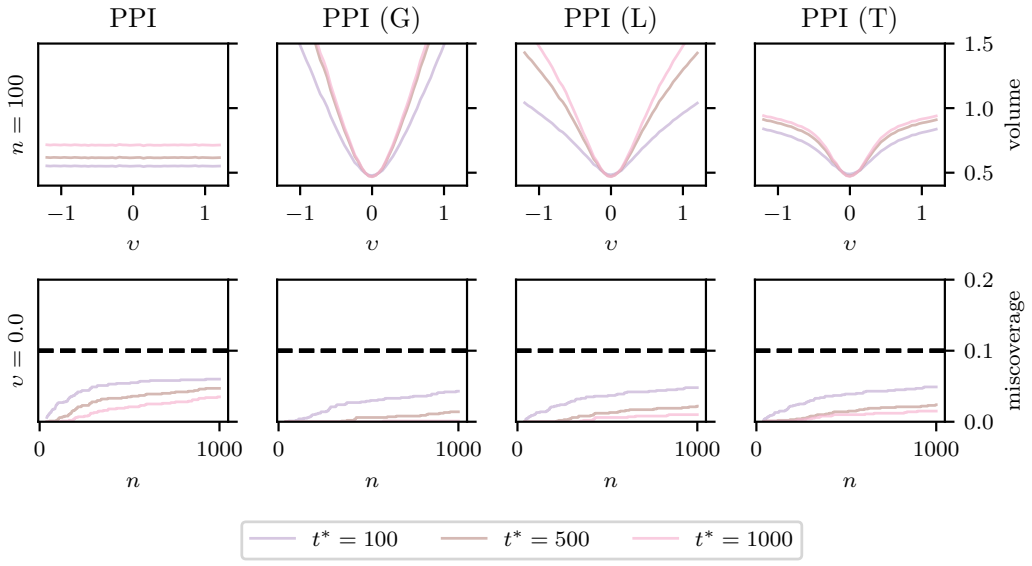

Figure S8: Biased predictions study with different hyperparameters. Each column correspond to one of the prediction-powered methods in Figure 2 for $t^\star = 100, 500, 1000$. The top and bottom column show average interval volume and cumulative miscoverage rate over 1000 repetitions with $\mathrm{df} = \infty$.

S7.1.3 Multivariate biased predictions

Here, we illustrate the multivariate AsympCS procedure described in Appendix S4 in the context of PPI. To do this, we study a simple multivariate version of the mean estimation task with biased prediction described in Section 6.1. In particular, for $d = 5$, we sample d-dimensional observations $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is a Toeplitz covariance matrix with entries $\Sigma_{ij} = 0.5^{|i-j|}$, and define $Y_i = f(X_i) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d)$, so that $\theta^\star = \mathbb{E}[Y] = \mathbf{0}$. Then, we proceed as in Section 6.1 and define biased predictions $f(X_i) = X_i + \upsilon$, where $\upsilon \in \mathbb{R}$ controls the bias level of the predictor. In this setup, we compare classical inference, non-assisted PPI, and Bayes-assisted PPI under a Gaussian prior with mean zero and isotropic covariance matrix using the spherical multivariate AsympCS procedures described in Appendix S4. The AsympCS hyperparameters are set using natural extensions of the rules in Appendix S6.4 to the multivariate case, with $t^\star = 1000$ for all methods. Figure 2 shows the average spherical interval volumes of each method as a function of υ, which we vary between -6 and 6, at $n \in \{100, 250, 500\}$. The results are consistent with those

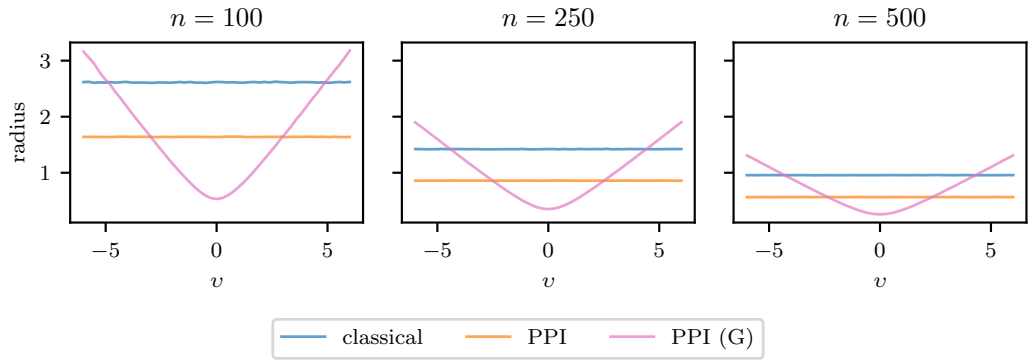

Figure S9: Multivariate biased predictions study. The left, middle and right panels show average spherical region radius over 1000 repetitions for $n = 100, 250, 500$.

in Figure 2. In particular, non-assisted PPI consistently outperforms classical inference, with both methods yielding constant interval radii across bias levels. On the other hand, Bayes-assisted PPI achieves smaller radii than the other baselines for small values of υ, but its radius grows quickly with $|\upsilon|$ as the prior becomes increasingly misspecified. For this example, we do not report coverage results, as we find that all methods achieve near perfect coverage across the values of υ considered, with cumulative miscoverage rates close to zero, likely due to the conservative spherical construction mentioned in Appendix S4.

S7.2 Real data

S7.2.1 Mean estimation

For each of the mean estimation experiments, we set $t^\star$ (see Appendix S6.4) equal to the largest n considered in the experiment. In particular, for the FLIGHTS, FOREST, and GALAXIES datasets, we set $t^\star = 10000$, 500, and 1000, respectively.

Figure S10 adds the results of the standard PPI procedures to the ones shown in Figure 3. For these experiments, the improvement of PPI++ over standard PPI is small, and the results remain consistent with those in Figure 3. That is, PPI methods consistently improve over classical inference, with Bayes-assisted methods providing an additional efficiency boost for moderate labelled sample sizes.

S7.2.2 Other estimation tasks

The estimation tasks considered here involve linear regression (CENSUS dataset), logistic regression (HEALTHCARE dataset), and median estimation (GENES dataset). As above, for each estimation task, we set $t^\star = 2000$ (see Appendix S6.4), as that is the largest n considered in all experiments.

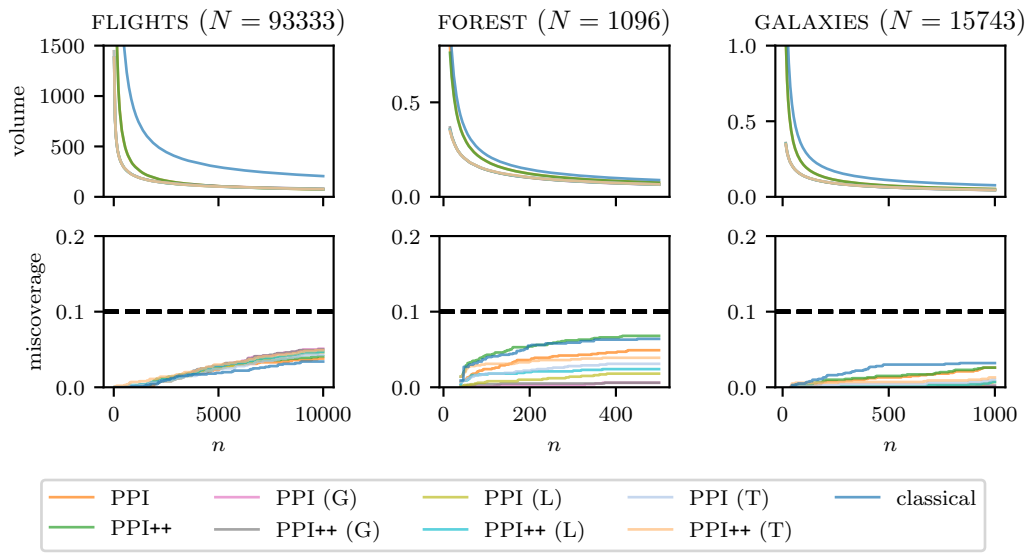

Figure S10: Mean estimation. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 1000 repetitions for the FLIGHTS, FOREST, and GALAXIES datasets.

For these, AsympCS procedures relying on classical inference (obtained from Theorem 1) and PPI, both non-assisted (Proposition 3) and Bayes-assisted (Equation (26)), require constructing a grid over θ through Equation (4). To initialise the grid, we use the first n_0 labelled data points to compute a preliminary estimate of $\theta^\star$, which we then use to centre the grid. The same n_0 is also used as the starting point to evaluate the AsympCS procedures and compute their cumulative miscoverage rate reported in the figures below. We set $n_0 = 100$ for the CENSUS and HEALTHCARE datasets, and $n_0 = 40$ for the GENES dataset.

Furthermore, some priors, including the Student-t prior, require numerical integration to compute the marginal density η_t used in Equation (26). As a result, when Bayes-assisted PPI under such priors is used, the computational cost grows significantly when Equation (26) is evaluated across many n and θ values simultaneously. Because of this, we only report results for the Gaussian and Laplace priors, which admit closed-form expressions for η_t.

Figure S11 compares classical and PPI AsympCS procedures on the three estimations tasks above in terms of average interval volume and cumulative miscoverage rate as n increases. As discussed in Section 6.2, PPI methods outperform classical inference for the linear and logistic regression tasks, with Bayes-assisted methods further improving efficiency when n is moderate. For the median estimation task, on the other hand, non-assisted PPI still improves over classical inference, while Bayes-assisted PPI yield larger regions than the other methods due to the higher bias of the predictions in this dataset. In all cases, coverage remains satisfactory.

As discussed in Appendix S6.2, the CENSUS, HEALTHCARE, and GENES datasets are preprocessed by applying a log-transformation to relevant positive skewed variables, as it is commonly done in the literature. For instance, this is the case for the income variable Y_i in the CENSUS dataset. In practice, such a transformation improves the accuracy of the KMT coupling approximation for a given n, essentially lowering the effective labelled sample size necessary to achieve satisfactory coverage. To see this, we repeat the linear regression experiment on the CENSUS dataset without applying any preprocessing to Y_i. As shown in Figure S12, the results are strikingly different: all methods yield significantly larger cumulative miscoverage rates compared to the preprocessed case in Figure S11, with non-assisted PPI violating the nominal guarantee around $n \approx 500$, in turn invalidating the efficiency comparison. Without preprocessing, a substantially larger starting labelled sample size n_0 is needed before the KMT coupling approximation is accurate enough for satisfactory uniform-time coverage by the AsympCS procedures. This example highlights the importance of knowledge of the data distribution when using AsympCS procedures in practice. A possible way to obtain such knowledge in practice is to estimate the third moment of the data distribution, if it exists, from a

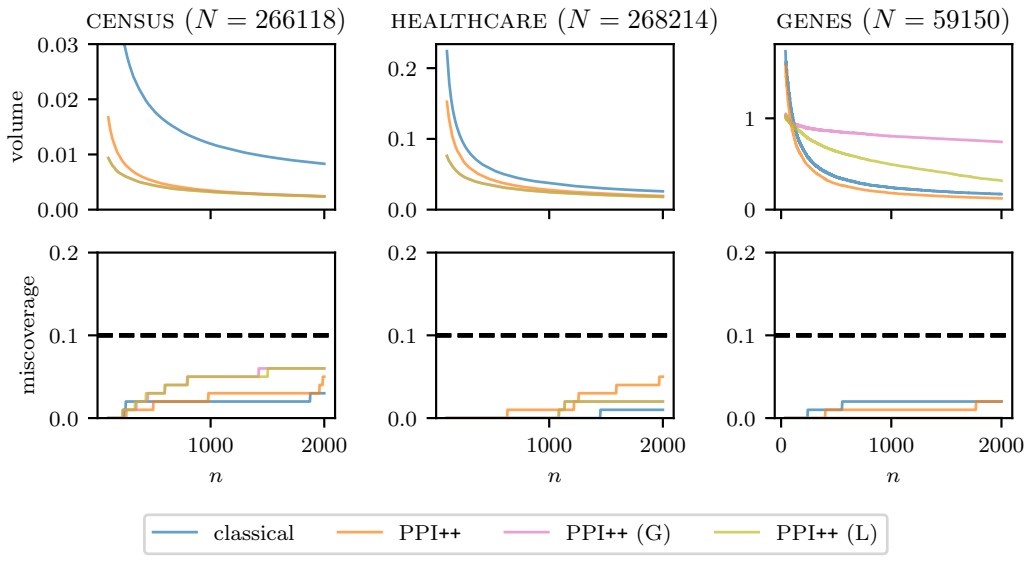

Figure S11: Other estimation tasks. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 100 repetitions for the CENSUS, HEALTHCARE, and GENES datasets.

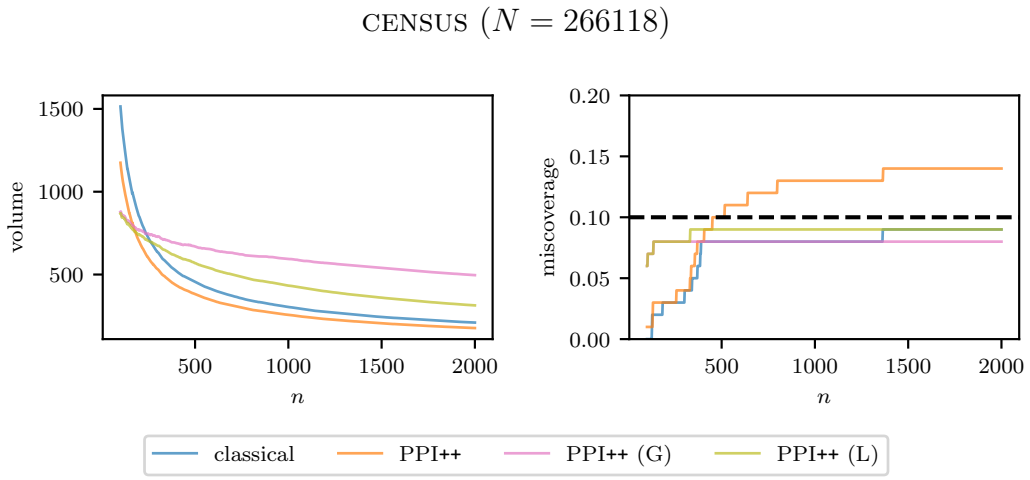

Figure S12: Linear regression on the CENSUS dataset without preprocessing. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 100 repetitions.

held-out validation set and use a Berry-Esseen-type bound [28, Theorem 2.1.4] to choose a starting labelled sample size n_0 at which the KMT coupling provides a good approximation.

S8 Alternative non-assisted AsympCS

S8.1 Parameter-free AsympCS via improper prior

As discussed in Section 3.1, the non-assisted asymptotic confidence sequence $\mathcal{C}^{\mathtt{NA}}_{\alpha,t}(\overline{Y}_t, \widehat{\sigma}_t; \rho)$ in Equation (6) approximates the exact CS in Equation (S30) and becomes arbitrarily accurate as in the limit. This suggests constructing alternative non-assisted AsympCS by approximating other exact CSs for which adaptations of Theorems 1 and 2 apply.

One example is the parameter-free non-assisted CS of Wang and Ramdas [19, Corollary 5.9]. Define the continuous, strictly decreasing bijection $g\colon [1,\infty) \to (0,1]$ by

$$g(x) := 2\left[1 - \Phi\left(\sqrt{\log(x^2)}\right)\right] + 2\sqrt{\log(x^2)}\phi\left(\sqrt{\log(x^2)}\right),$$

with Φ and ϕ the standard normal CDF and PDF, and let $z_\alpha := g^{-1}(\alpha)$, which is well-defined. The corresponding AsympCS is given by

$$\mathcal{C}^{\mathtt{NA}'}_{\alpha,t}(\overline{Y}_t, \widehat{\sigma}_t) := \left[\overline{Y}_t \pm \frac{\widehat{\sigma}_t}{\sqrt{t}}\sqrt{\log\left(tz_\alpha^2\right)}\right]. \tag{S90}$$

Notably, for i.i.d. Gaussian observations with known variance, the exact (nonasymptotic) counterpart of (S90) is obtained by applying the method of mixtures for extended nonnegative martingales [19, Def. 3.1] together with extended Ville's inequality [19, Theorem 4.1], using a non-informative improper prior as the mixing density.

S8.2 Experiments

We use $\mathcal{C}^{\mathtt{NA}'}_{\alpha,t}$ from (S90) as a parameter-free drop-in replacement for $\mathcal{C}^{\mathtt{NA}}_{\alpha,t}$ in the classical and prediction-powered AsympCS procedures from the main text and repeat some of the experiments from Section 6. In figures, runs that use the alternative AsympCS are annotated (I) for "improper". Compared with

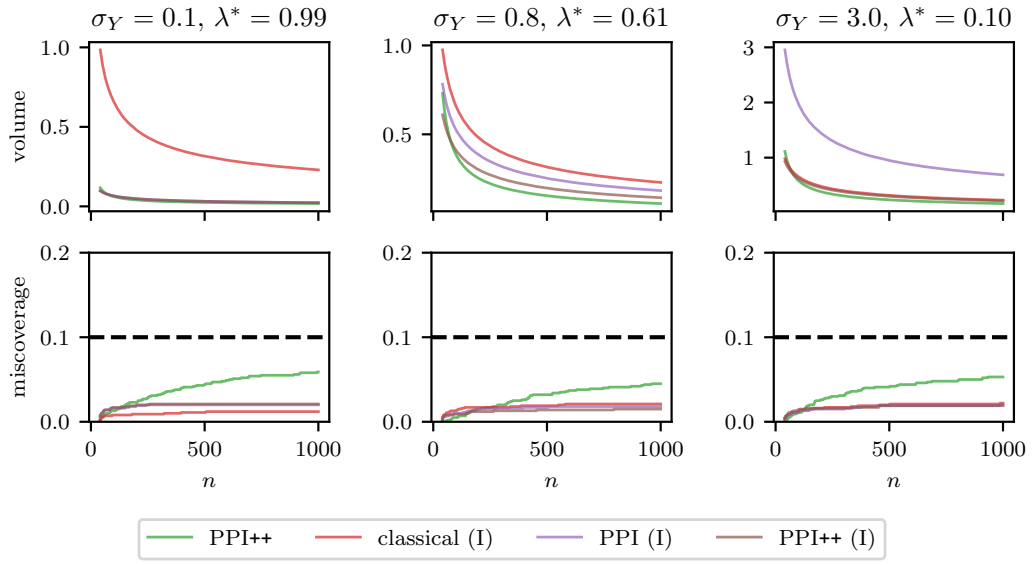

Figure S13: Noisy predictions study with alternative AsympCS . The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for noise levels $\sigma_Y = 0.1, 0.8, 3.0$. Results for non-assisted PPI++ based on Equation (6) are shown for reference.

the standard non-assisted AsympCS used in the main text, which depends on the hyperparameter ρ, the parameter-free alternative typically performs slightly worse under our default choice of ρ (see Appendix S6.4). Nonetheless, $\mathcal{C}^{\mathtt{NA}'}_{\alpha,t}$ can represent an attractive choice when selecting ρ is problematic, precisely because it avoids any tuning.

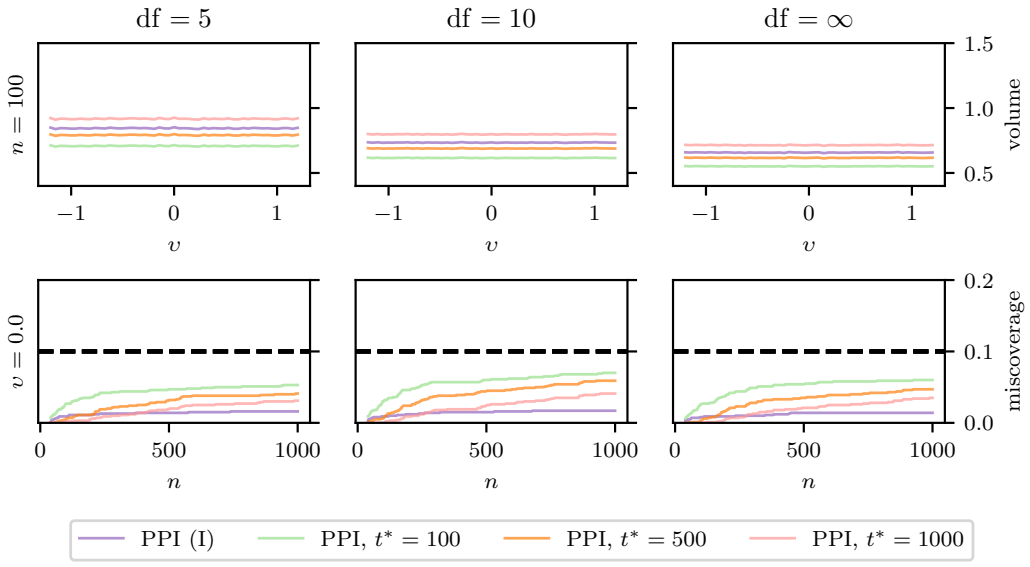

Figure S14: Biased predictions study with alternative AsympCS . The left, middle and right panels show average interval volume and cumulative miscoverage rate over 100 repetitions for $\mathrm{df} = 5, 10, \infty$. Results for non-assisted PPI based on Equation (6) are shown for reference.

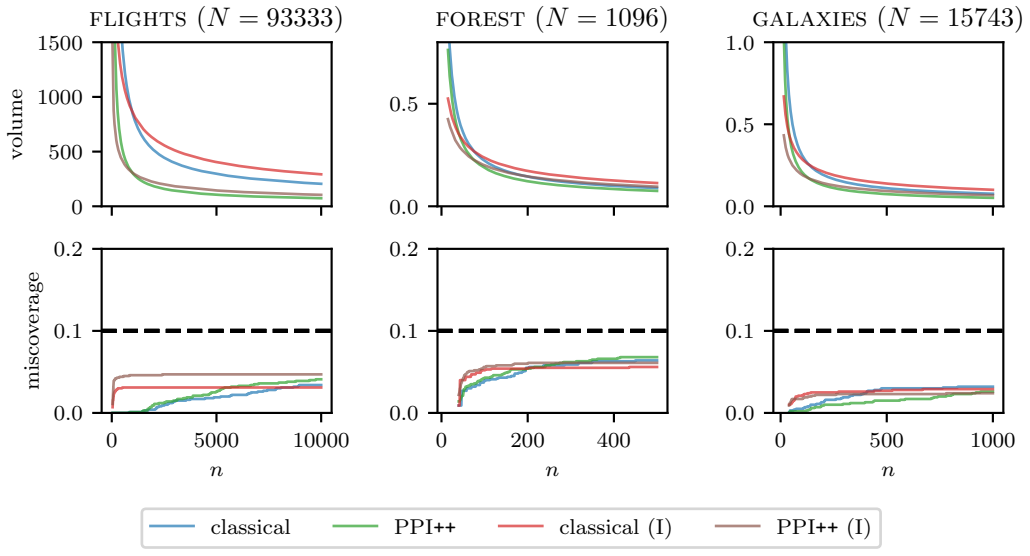

Figure S15: Real data study with alternative AsympCS . The top and bottom rows show the average interval volume and cumulative miscoverage rate over 1000 repetitions for the FLIGHTS, FOREST, and GALAXIES datasets. Results for classical inference and non-assisted PPI++ based on Equation (6) are shown for reference.

References

[1] R. Durrett. *Probability: theory and examples*. Duxbury Press, fifth edition, 2019. ISBN 0-534-24318-5.

[2] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent RV'-s, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(1): 111–131, March 1975. ISSN 1432-2064. doi: 10.1007/BF00533093.

[3] P. Major. The approximation of partial sums of independent RV's. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35(3):213–220, September 1976. ISSN 1432-2064. doi: 10.1007/BF00532673.

[4] S. Csörgö and P. Hall. The Komlós-Major-Tusnády Approximations and Their Applications. *Australian Journal of Statistics*, 26(2):189–218, 1984. ISSN 1467-842X. doi: 10.1111/j.1467-842X.1984.tb01233.x.

[5] U. Einmahl. Strong invariance principles for partial sums of independent random vectors. *The Annals of Probability*, pages 1419–1440, 1987.

[6] I. Waudby-Smith, D. Arbour, R. Sinha, E. Kennedy, and A. Ramdas. Supplement to "time-uniform central limit theory and asymptotic confidence sequences". *The Annals of Statistics*, 52 (6):2613–2640, 2024.

[7] H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

[8] T. L. Lai. On confidence sequences. *The Annals of Statistics*, pages 265–280, 1976.

[9] J. Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

[10] B. Chugg, H. Wang, and A. Ramdas. A unified recipe for deriving (time-uniform) pac-bayes bounds. *Journal of Machine Learning Research*, 24(372):1–61, 2023.

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

[12] A. Balsubramani and A. Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. *arXiv preprint arXiv:1506.03486*, 2015.

[13] E. Kaufmann and W. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

[14] S. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.

[15] I. Waudby-Smith, D. Arbour, R. Sinha, E. Kennedy, and A. Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 52(6):2613–2640, 2024.

[16] M. Haddouche and B. Guedj. Pac-bayes generalisation bounds for heavy-tailed losses through supermartingales. *arXiv preprint arXiv:2210.00928*, 2022.

[17] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[18] R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[19] H. Wang and A. Ramdas. The extended Ville's inequality for nonintegrable nonnegative supermartingales. *arXiv preprint arXiv:2304.01163*, 2023.

[20] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. ISBN 978-1-118-62639-9.

[21] H. White. Using least squares to approximate unknown regression functions. *International economic review*, pages 149–170, 1980.

[22] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 1991. edition, 1991. ISBN 978-3-642-08087-6. doi: 10.1007/978-3-642-20212-4.

[23] V. Koval. A New Law of the Iterated Logarithm in Rd with Application to Matrix-Normalized Sums of Random Vectors. *Journal of Theoretical Probability*, 15(1):249–257, January 2002. ISSN 1572-9230. doi: 10.1023/A:1013851720494.

[24] E. Bullock, C. Woodcock, C. Souza Jr, and P. Olofsson. Satellite-based estimates reveal widespread forest degradation in the amazon. *Global Change Biology*, 26(5):2956–2969, 2020.

[25] K. Willett, C. Lintott, S. Bamford, K. Masters, B. Simmons, K. Casteels, E. Edmondson, L. Fortson, S. Kaviraj, W. Keel, T. Melvin, Nichol R., Raddick M., Schawinski K., Simpson R., Skibba R., Smith A., and Thomas D. 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.

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

[27] E. D. Vaishnav, C. G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis, D. A. Thompson, J. Z. Levin, F. A. Cubillos, and A. Regev. The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, 603(7901):455–463, 2022.

[28] Roman Vershynin. *High-dimensional probability*. Cambridge University Press, 2009.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

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

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification:The limitations are discussed in our Discussion section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Proofs of all results are included in full detail in the supplementary material, except for some very classical results for which we provide only references. When relevant, we also provide a sketch of the proof in the main text.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: As stated in the supplementary material, the code used to perform our experiments is made available online under a permissive licence.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: As stated in the supplementary material, the code used to perform our experiments is made available online under a permissive licence.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All details necessary to understand the results are provided in the paper.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: For all experiments, results are averaged over many repetitions (100 or 1000 repetitions, depending on the experiment). Variations from the mean are negligible. Statistical guarantees (i.e. asymptotic time-uniform coverage) are checked empirically computing the average cumulative miscoverage rate for all experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: All experiments were run locally on an Apple Silicon M4 Pro CPU with 24GB of memory, and implementation details are provided in the supplementary material.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

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

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [NA]

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

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

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

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets have permissive licenses and are properly credited in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: As stated in the supplementary material, the code used to perform our experiments is made available online under a permissive licence.

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.