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 subsequences 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 $\mathcal{C}^g_{\alpha,\theta,n}$ relies on estimating g_θ, for which PPI defines an estimator leveraging 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 [2], 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 [3, 4, 5], 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 of our approach and further extensions. 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 ML predictors, and later extended in Angelopoulos et al. [7]. Closely related ideas appear in the literature 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 (CS) were introduced by Darling and Robbins [13] and further developed by Robbins and Siegmund [4] and Lai [5], building on earlier work by Ville [3] and Wald [14]. Interest has surged again in recent years [15, 2], motivated by applications such as A/B testing. The notion is closely linked to e-values [16, 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 CS under stronger conditions (e.g., existence of bounded e-values) but does not leverage prior knowledge about the quality of the ML predictions. Furthermore, applying their method requires an active-learning setup in which, at each time 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, which are 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. [2] 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 first review background on (asymptotic) confidence sequences (CS), and 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) [2, 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 scaled estimators, asymptotic confidence sequences rely on *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 form an exact CS. This follows from combining the method of mixtures for nonnegative martingales with Ville's inequality [3, 4, 5, 21]. Second, the proof relies on KMT strong coupling [22, 23]: 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 in 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 [2, 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 such assumptions: the width of Equation (6) is 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 [2, 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 [2, §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, before providing 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 $\widehat{\gamma}_\lambda^{\mathrm{icv}}$ yields 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, 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 constructing 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* $0 < \delta < 1$*. Additionally, assume 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 $n \to \infty$*,*

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

Likewise, 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) and 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,

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

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 combine the results of Sections 3 and 4 within the PPI framework 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 about the black-box predictor's accuracy. 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 of 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 $1/\sqrt{n\log n}$ 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 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 $\theta^\star = \mathbb{E}[Y]$.

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 Section 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 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, reflecting their lack of prior information, and with the latter consistently outperforming the former

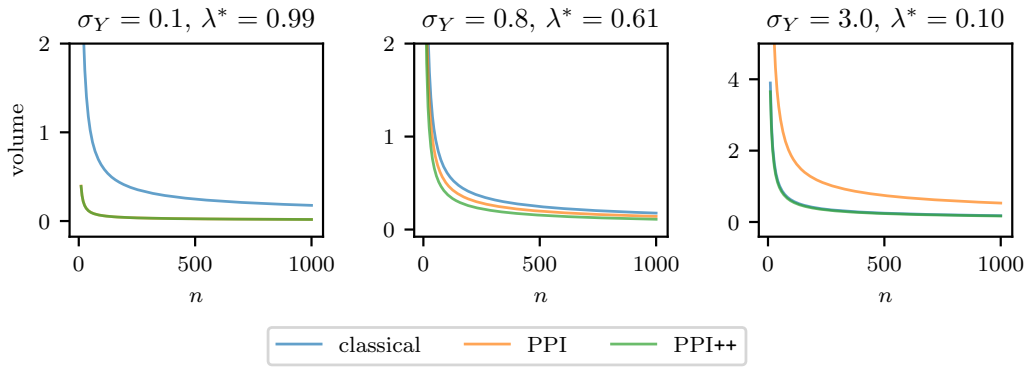

Figure 1: Noisy predictions study. The left, middle and right panels show average interval volume over 1000 repetitions as a function of the labelled sample size n for noise levels $\sigma_Y \in \{0.1, 0.8, 3.0\}$.

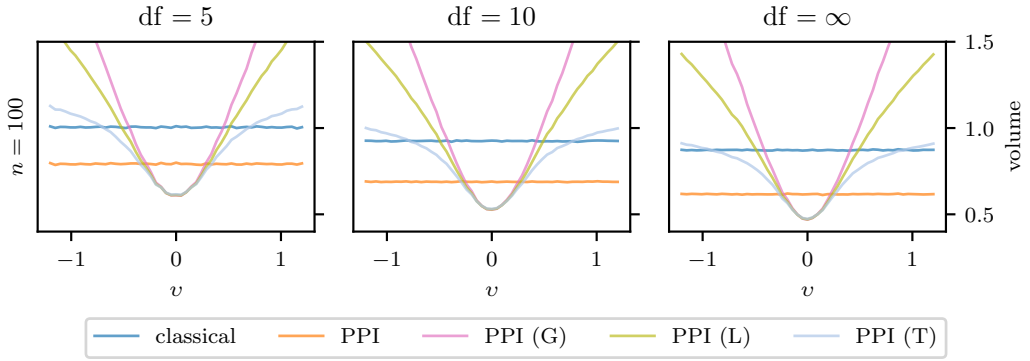

Figure 2: Biased predictions study. The left, middle and right panels show average interval volume over 100 repetitions as a function of the bias level υ for $\mathrm{df} = 5, 10, \infty$.

by leveraging imputed predictions. On the other hand, the volume of the Bayes-assisted procedures varies widely with the bias level υ: it 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 Section 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 Section 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, serving 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 S8 reports results for three additional estimation tasks: linear regression (CENSUS), logistic regression (HEALTHCARE), and quantile estimation (GENES). For the first two tasks, the same conclusions as for mean estimation hold: PPI methods consistently outperform classical inference,

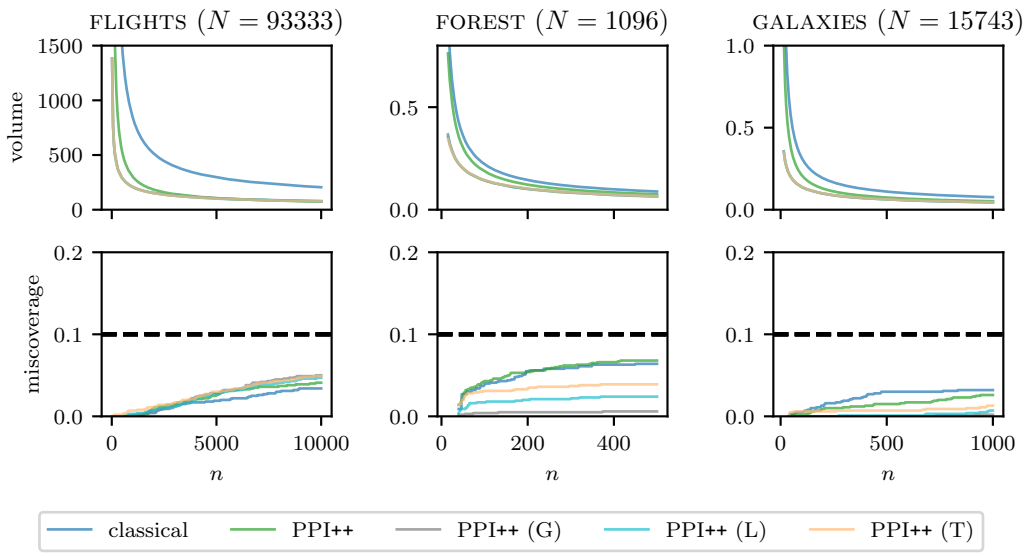

Figure 3: Mean estimation. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 1000 repetitions for the FLIGHTS, FOREST, and GALAXIES datasets.

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 allow for the 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 Section S4, building on earlier work by Waudby-Smith et al. [2, §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 Section S8, the parameter-free CS proposed by Wang and Ramdas [29], 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. [2, §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 S8), 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 requires 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.

[2] 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.

[3] 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.

[5] 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.

[14] 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.

[16] 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.

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

[22] 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.

[23] 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.

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

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.