

---

# Supplement to 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

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

<b>S1 Additional background</b>	<b>3</b>
S1.1 Asymptotic theory of partial sums . . . . .	3
S1.1.1 Iterated logarithm and Marcinkiewicz-Zygmund strong laws . . . . .	3
S1.1.2 Strong approximations . . . . .	3
S1.2 Confidence sequences . . . . .	3
S1.2.1 Confidence intervals vs. confidence sequences . . . . .	3
S1.2.2 Nonnegative supermartingale and Ville’s inequality . . . . .	4
S1.2.3 Method of mixture . . . . .	5
<b>S2 Secondary results</b>	<b>5</b>
S2.1 Strong coupling with i.i.d. Gaussian . . . . .	5
S2.2 Confidence sequence for i.i.d. Gaussian variables with known variance . . . . .	7
S2.3 Optimal control-variate parameter for PPI++ . . . . .	9
<b>S3 Proofs</b>	<b>9</b>
S3.1 Proofs of Theorem 1 and Theorem 2 . . . . .	9
S3.1.1 Proof of Theorem 2 . . . . .	9
S3.1.2 Proof of Theorem 1 . . . . .	10
S3.2 Proofs of Theorem 3 and Theorem 4 . . . . .	10

---

\*Equal contribution. Order decided by coin toss.

S3.2.1	Proof of Theorem 4 . . . . .	10
S3.2.2	Proof of Theorem 3 . . . . .	11
S3.3	Proof of Theorem 5 . . . . .	11
S3.4	Proof of Proposition 1 . . . . .	13
S3.5	Proof of Proposition 2 . . . . .	13
S3.6	Proof of Proposition 4 . . . . .	14
S3.7	Proof of Proposition 5 . . . . .	14
<b>S4</b>	<b>Multivariate AsympCS</b>	<b>14</b>
S4.1	Definitions . . . . .	14
S4.2	Nonasymptotic Bayes-assisted CS for i.i.d. Gaussian random vectors . . . . .	15
S4.3	Multivariate extension of Theorems 3 and 4 . . . . .	15
<b>S5</b>	<b>Derivations for prediction-powered mean estimation</b>	<b>16</b>
<b>S6</b>	<b>Experimental details</b>	<b>19</b>
S6.1	Implementation . . . . .	19
S6.2	Datasets . . . . .	19
S6.3	Predictor performance . . . . .	20
S6.4	AsympCS hyperparameters . . . . .	20
<b>S7</b>	<b>Additional experimental results</b>	<b>21</b>
S7.1	Synthetic data . . . . .	21
S7.1.1	Noisy predictions . . . . .	21
S7.1.2	Biased predictions . . . . .	21
S7.1.3	Multivariate biased predictions . . . . .	24
S7.2	Real data . . . . .	24
S7.2.1	Mean estimation . . . . .	24
S7.2.2	Other estimation tasks . . . . .	24
<b>S8</b>	<b>Alternative non-assisted AsympCS</b>	<b>26</b>
S8.1	Parameter-free AsympCS via improper prior . . . . .	26
S8.2	Experiments . . . . .	27

## S1 Additional background

### S1.1 Asymptotic theory of partial sums

#### S1.1.1 Iterated logarithm and Marcinkiewicz-Zygmund strong laws

**Theorem S1.** (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,

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

which implies

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

**Theorem S2.** (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,

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

#### 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. We state the version from Csörgö and Hall [4, Theorem 3.2].

**Theorem S3** (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

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

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

**Theorem S4.** (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  $\Sigma$ , 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,

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

### 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 prespecified error probability (so the confidence level is  $1 - \alpha$ ).

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

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

This guarantee is *marginal in  $t$* : it holds for any fixed, 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  $\tau$ ,

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

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 S1), 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

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

Equivalently,

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

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  $\tau$ ,

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

Thus CSs support continual monitoring and sequential decision-making without inflating error rates. In practice, 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.

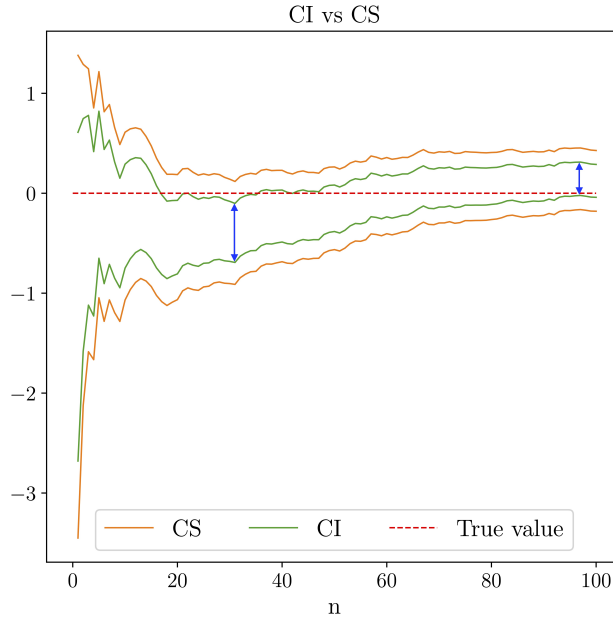


Figure S1: Comparison of fixed-time confidence intervals (CIs) and a confidence sequence (CS) for data from  $\mathcal{N}(0, 1)$ . Two fixed-time CIs at different sample sizes happen to be disjoint (highlighted), illustrating that marginal coverage at each  $t$  does not imply simultaneous coverage over  $t$ . 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 S1** (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 equality holds, then  $M$  is a nonnegative martingale.

**Proposition S1** (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 S2.** *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 S3** (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  $\mu'$  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  $\mathbb{E}[M_t(\mu') \mid \mathcal{F}_{t-1}]$  is  $\mathcal{F}_{t-1} \otimes \mathcal{B}$ -measurable.*

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

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

*Then the mixture  $(\tilde{M}_t)_{t \in \mathbb{N}}$ , 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 supermartingales (say) of the form  $M_t(\mu')$  for  $\mu' \in \mathbb{R}$  which satisfy conditions 1 and 2 above, together with a mixture distribution  $\pi$  satisfying the assumptions of Proposition S3, 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  $\alpha \in (0, 1)$

$$\Pr \left( \int_{\mu' \in \mathbb{R}} M_t(\mu') d\pi(\mu') \leq \frac{1}{\alpha} \text{ for all } t \geq 1 \right) \geq 1 - \alpha. \quad (\text{S1})$$

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 S3). It will be used in the proofs of Theorem 3 and Proposition 2.

**Proposition S4.** *Let  $\xi_1, \xi_2 \dots$  be i.i.d. random variables with mean  $\mu$  and variance  $\sigma^2$  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  $\mu$  and variance  $r\sigma^2$  such that*

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

*Proof.* By Theorem S3, there exists a Brownian motion  $B$  such that, a.s. as  $n \rightarrow \infty$ ,

$$\begin{aligned} \sum_{i=1}^{N_n} \frac{\xi_i - \mu}{\sigma} &= B_{N_n} + o\left(N_n^{1/q}\right) \\ &= B_{N_n} + o\left(n^{1/q}\right) \\ &= \frac{N_n}{n} r B_{n/r} + (B_{N_n} - \frac{N_n}{n} r B_{n/r}) + o\left(n^{1/q}\right). \end{aligned} \quad (\text{S2})$$

We have

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

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

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

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

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

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

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

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  $|B_{N_n} - B_{n/r}| = o(n^{1/q})$  a.s. as  $n \rightarrow \infty$ . 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  $B_{n/r}(1 - \frac{N_n}{n} r) = o(n^{1/q})$  a.s. as  $n \rightarrow \infty$ . So, from Equation (S2), 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,

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

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

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

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

$$|\lambda^* - \hat{\lambda}| = o(n^{-\frac{\delta}{2+\delta}}) \text{ a.s. as } n \rightarrow \infty.$$

*Proof.* By the mean value theorem, we obtain:

$$\begin{aligned} |\lambda^* - \hat{\lambda}| &\leq \frac{1}{\text{var}(U)} \left| \frac{1}{n} \sum_{i=1}^n (U_i V_i - \mathbb{E}(UV)) \right| + \frac{|\mathbb{E}(U)|}{\text{var}(U)} |\bar{V} - \mathbb{E}V| + \frac{|\bar{V}|}{\text{var}(U)} |\bar{U} - \mathbb{E}U| \\ &\quad + \widehat{\text{cov}}((U_i, V_i)_{i=1}^n) K_1 \left| \frac{1}{n} \sum_{i=1}^n (U_i^2 - \mathbb{E}(U^2)) \right| + \widehat{\text{cov}}((U_i, V_i)_{i=1}^n) K_1 K_2 |\bar{U} - \mathbb{E}U| \end{aligned}$$

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  $2 + \delta$  (hence also of order  $1 + \delta/2$ ). Thus, applying Theorem S2 with  $p = 1 + \delta/2$  yields the result.  $\square$

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

An important step in the proofs 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 S5.** Let  $W_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ . For any parameter  $\rho > 0$ , the sequence of intervals defined by

$$\mathcal{C}_{\alpha,t}^{\text{NA}}(\bar{W}_t, \sigma; \rho) := \left[ \bar{W}_t \pm \frac{\sigma}{\sqrt{t}} \sqrt{\left(1 + \frac{1}{t\rho^2}\right) \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)} \right] \quad (\text{S3})$$

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

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

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 S6.** Let  $W_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$  and let  $\pi$  be a continuous and proper prior on  $\mu/\sigma$ , then

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

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

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

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

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

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

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

where  $\bar{w}_t = \frac{1}{t} \sum_{i=1}^t w_i$ ,  $C(\sigma, w_1, \dots, w_t)$  does not depend on  $\mu$  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

$$\tilde{M}_t(\bar{w}_t, \mu) = \int_{\mathbb{R}} \frac{p_{\mu'}(w_{1:t})}{p_{\mu}(w_{1:t})} \pi\left(\frac{\mu'}{\sigma}\right) \frac{d\mu'}{\sigma} \quad (\text{S4})$$

$$= \int_{\mathbb{R}} \frac{\phi_t\left(\frac{\bar{w}_t - \mu'}{\sigma}\right)}{\phi_t\left(\frac{\bar{w}_t - \mu}{\sigma}\right)} \pi\left(\frac{\mu'}{\sigma}\right) \frac{d\mu'}{\sigma} \quad (\text{S5})$$

$$= \frac{\eta_t(\bar{w}_t/\sigma)}{\phi_t((\bar{w}_t - \mu)/\sigma)}. \quad (\text{S6})$$

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

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

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 each  $W_i$  and in  $\mu'$ . Hence, it is  $\mathcal{F}_t \otimes \mathcal{B}(\mathbb{R})$ -measurable. Similarly,  $\mathbb{E}[M_t(\mu') \mid \mathcal{F}_{t-1}] = M_{t-1}(\mu')$  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  $\lambda$  is the Lebesgue measure on  $\mathbb{R}$  (or any other measure dominated by the Lebesgue measure on  $\mathbb{R}$ ). Then, by Proposition S3,  $(\tilde{M}_t(\bar{W}_t, \mu))_{t \geq 1}$  is a nonnegative martingale with respect to the adapted filtration  $(\mathcal{F}_t)_{t \in \mathbb{N}}$ . Hence, by Ville's inequality, we have:

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

It follows that the sequence

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

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

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

□

### 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 S5** (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*

$$\lambda_\theta^* = \frac{\text{cov}(\ell'_\theta(X_1, Y_1), \ell'_\theta(X_1, f(X_1)))}{\text{var}(\ell'_\theta(X_1, f(X_1)))}. \quad (\text{S7})$$

Then, for any  $\theta$ , almost surely as  $n \rightarrow \infty$ ,

$$\widehat{g}_{\theta,n}^{PP+} = \left[ \frac{1}{n} \sum_{i=1}^n \ell'_\theta(X_i, Y_i) \right] - \lambda_\theta^* \left( \left[ \frac{1}{n} \sum_{i=1}^n \ell'_\theta(X_i, f(X_i)) \right] - \widehat{m}_\theta \right) + o\left(\sqrt{\frac{\log \log n}{n}}\right), \quad (\text{S8})$$

$$\begin{aligned} \widehat{\Delta}_{\theta,n}^{PP+} &= \frac{1}{n} \sum_{i=1}^n (\ell'_\theta(X_i, Y_i) - \ell'_\theta(X_i, f(X_i))) - (\lambda_\theta^* - 1) \left( \frac{1}{n} \left[ \sum_{i=1}^n \ell'_\theta(X_i, f(X_i)) \right] - \widehat{m}_\theta \right) \\ &\quad + o\left(\sqrt{\frac{\log \log n}{n}}\right). \end{aligned} \quad (\text{S9})$$

Additionally, in the case of the squared loss:

$$\widehat{\theta}_n^{PP+} = \frac{1}{n} \sum_{i=1}^n Y_i - \lambda_0^* \left( \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) \right) + o\left(\sqrt{\frac{\log \log n}{n}}\right), \quad (\text{S10})$$

with  $\lambda_0^* = \text{cov}(Y_1, f(X_1))/\text{var}(f(X_1))$ .

## S3 Proofs

### S3.1 Proofs of Theorem 1 and Theorem 2

Theorem 1 is a corollary of Theorem 2; we therefore begin 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  $\overline{W}_t = \frac{1}{t} \sum_{i=1}^t W_i$ .

By Theorem S5, the sequence of intervals  $\mathcal{C}_{\alpha,t}^{\text{NA}}(\overline{W}_t, \sigma; \rho) = \mathcal{C}_{\alpha,t}^{\text{NA}}(\widehat{\mu}_t - \varepsilon_t, \sigma; \rho) = [\widehat{\mu}_t - L_t^*, \widehat{\mu}_t + U_t^*]$ , where

$$\begin{aligned} U_t^* &= \frac{\sigma}{\sqrt{t}} \sqrt{\left(1 + \frac{1}{t\rho^2}\right) \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)} - \varepsilon_t, \text{ and} \\ L_t^* &= \frac{\sigma}{\sqrt{t}} \sqrt{\left(1 + \frac{1}{t\rho^2}\right) \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)} + \varepsilon_t, \end{aligned}$$

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

$$\begin{aligned} \frac{1}{a_t} (L_t - L_t^*) &= \frac{1}{a_t} \left[ \frac{\hat{\sigma}_t}{\sqrt{t}} \sqrt{\left(1 + \frac{1}{t\rho^2}\right) \log \left(\frac{t\rho^2 + 1}{\alpha^2}\right)} \right. \\ &\quad \left. - \frac{\sigma}{\sqrt{t}} \sqrt{\left(1 + \frac{1}{t\rho^2}\right) \log \left(\frac{t\rho^2 + 1}{\alpha^2}\right)} \right] + o(1). \end{aligned}$$

We have

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

Hence  $\frac{1}{a_t} (L_t - L_t^*) = o(1)$ . Similarly,  $\frac{1}{a_t} (U_t - U_t^*) = o(1)$ . It follows that  $(\mathcal{C}_{\alpha,t}^{\text{NA}})$  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 S3), there exists a sequence of i.i.d. Gaussian random variables  $(W_i)_{i \geq 1}$  with mean  $\mu$  and variance  $\sigma^2$  such that, a.s.,

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

We also need to satisfy the condition on the variance. Assuming  $\mathbb{E}[|Y|^{2+\delta}] < \infty$ , the Marcinkiewicz-Zygmund strong law of large numbers (Theorem S2) with  $p = 1 + \delta/2 \in (1, 2)$  yields a polynomial a.s. rate for  $\bar{Y}_t$  and  $\bar{Y}_t^2$ . Consequently,  $|\hat{\sigma}_t - \sigma| = o(t^{-\gamma})$  for some  $\gamma > 0$ , which implies

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

The result then 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  $\pi$  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  $Z_t \rightarrow c$  a.s. as  $t \rightarrow \infty$ . Let*

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

Then

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

Theorem 3 is a corollary of Theorem 4; we therefore begin by proving Theorem 4.

#### S3.2.1 Proof of Theorem 4

By assumption, we have, almost surely,

$$\hat{\mu}_t = \bar{W}_t + \varepsilon_t,$$

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

Using Theorem S6, the sequence of intervals  $\mathcal{C}_{\alpha,t}^{\text{BA}}(\bar{W}_t, \sigma; \pi) = \mathcal{C}_{\alpha,t}^{\text{BA}}(\hat{\mu}_t - \varepsilon_t, \sigma; \pi) = [\hat{\mu}_t - L_t^*, \hat{\mu}_t + U_t^*]$ , where

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

$$L_t^* = \frac{\sigma}{\sqrt{t}} \sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\hat{\mu}_t - \varepsilon_t}{\sigma}\right)} + \varepsilon_t,$$

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

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

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

$$\begin{aligned} \frac{1}{a_t} (L_t - L_t^*) &= \frac{1}{a_t} \left[ \frac{\hat{\sigma}_t}{\sqrt{t}} \sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\hat{\mu}_t}{\hat{\sigma}_t}\right)} \right. \\ &\quad \left. - \frac{\sigma}{\sqrt{t}} \sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\hat{\mu}_t - \varepsilon_t}{\sigma}\right)} \right] + o(1) \end{aligned}$$

We have

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

as, by Lemma S2, we have

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

It follows that  $(\mathcal{C}_{\alpha,t}^{\text{BA}})$  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 the same argument used in 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  $(\mathcal{C}_{\alpha,t}^*)$ . For any  $\alpha' > \alpha$ , the narrower exact CS  $(\mathcal{C}_{\alpha',t}^*)$  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  $\mathcal{C}_{\alpha,t} = [\hat{\mu}_t \pm U_{\alpha,t}]$  where

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

for the non-assisted case, and

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

for the Bayes-assisted case.  $\mathcal{C}_{\alpha,t}$  approximates a reference exact CS of the form  $\mathcal{C}_{\alpha,t}^* = [\hat{\mu}_t - L_{\alpha,t}^*, \hat{\mu}_t + U_{\alpha,t}^*]$ , 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

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

$$L_{\alpha,t}^* = \frac{\sigma}{\sqrt{t}} \sqrt{\log\left(\frac{t}{2\pi\alpha^2}\right) - 2\log\eta_t\left(\frac{\hat{\mu}_t - \varepsilon_t}{\sigma}\right) + \varepsilon_t},$$

for the Bayes-assisted case, where  $\varepsilon_t = \hat{\mu}_t - \bar{W}_t = o(a_t)$  a.s. does not depend on  $\alpha$ . In both cases,  $(\mathcal{C}_{\alpha,t}^*)_{t \geq 1}$  is an exact CS (Theorems S5 and S6); hence, for  $E_\alpha^* = \{\mu \in \mathcal{C}_{\alpha,t}^* \text{ for all } t \geq 1\}$ ,

$$\Pr(E_\alpha^*) \geq 1 - \alpha.$$

Additionally,

$$U_{\alpha,t} \sim U_{\alpha,t}^* \sim L_{\alpha,t}^* \sim \sigma \sqrt{\frac{\log t}{t}} \text{ a.s. as } t \rightarrow \infty,$$

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

$$U_{\alpha,t} - U_{\alpha,t}^* = o(a_t) \text{ and } U_{\alpha,t} - L_{\alpha,t}^* = o(a_t). \quad (\text{S11})$$

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

$$U_{\alpha,t}^* - U_{\alpha',t}^* = \frac{(U_{\alpha,t}^*)^2 - (U_{\alpha',t}^*)^2}{U_{\alpha,t}^* + U_{\alpha',t}^*} \sim \frac{\sigma^2/t \log(\frac{\alpha'}{\alpha^2})}{2\sigma\sqrt{\log t/t}} \sim c(\alpha, \alpha')a_t \quad (\text{S12})$$

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

$$L_{\alpha,t}^* - L_{\alpha',t}^* \sim c(\alpha, \alpha')a_t \text{ a.s. as } t \rightarrow \infty. \quad (\text{S13})$$

Combining Equations (S12) and (S13) with Equation (S11), we obtain, a.s.

$$U_{\alpha,t} - U_{\alpha',t}^* \sim c(\alpha, \alpha')a_t \quad (\text{S14})$$

$$L_{\alpha,t} - L_{\alpha',t}^* \sim c(\alpha, \alpha')a_t. \quad (\text{S15})$$

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

$$U_{\alpha,t}(\omega) - U_{\alpha',t}^*(\omega) \geq \frac{1}{2}c(\alpha, \alpha')a_t \quad (\text{S16})$$

$$L_{\alpha,t}(\omega) - L_{\alpha',t}^*(\omega) \geq \frac{1}{2}c(\alpha, \alpha')a_t. \quad (\text{S17})$$

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

$$\Pr(\mu \in \mathcal{C}_{\alpha,t} \text{ for all } t \geq m) \geq \Pr(E_{\alpha'}^* \cap \{T_{\alpha'} \leq m\}) \xrightarrow{m \rightarrow \infty} \Pr(E_{\alpha'}^*) \geq 1 - \alpha'.$$

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

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

### S3.4 Proof of Proposition 1

Let

$$\begin{aligned}\epsilon_n &= \hat{\gamma}^{\text{cv}+} - \hat{\gamma}_{\lambda^*}^{\text{cv}} \\ &= \hat{\gamma}^{\text{cv}+} - (\bar{V} - \lambda^*(\bar{U} - \hat{\mu})) \\ &= (\lambda^* - \hat{\lambda})(\bar{U} - \mu) - (\lambda^* - \hat{\lambda})(\hat{\mu} - \mu).\end{aligned}$$

By the triangle inequality,

$$|\epsilon_n| \leq |\lambda^* - \hat{\lambda}|(|\bar{U} - \mu| + |\hat{\mu} - \mu|).$$

By Lemma S1 we have

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

and, by the law of the iterated logarithm

$$|\bar{U} - \mu| = O\left(\sqrt{\frac{\log \log n}{n}}\right), \quad |\hat{\mu} - \mu| = O\left(\sqrt{\frac{\log \log n}{n}}\right) \text{ a.s. as } n \rightarrow \infty. \quad (\text{S18})$$

It follows that

$$|\epsilon_n| = O\left(\frac{1}{n^{2/(2+\delta)}} \sqrt{\frac{\log \log n}{n}}\right) = o\left(\frac{1}{\sqrt{n \log n}}\right) \text{ a.s. as } n \rightarrow \infty. \quad (\text{S19})$$

### S3.5 Proof of Proposition 2

We have

$$\hat{\gamma}_{\lambda^*}^{\text{cv}} = \bar{V} - \lambda(\bar{U} - \mu) + \lambda(\hat{\mu} - \mu). \quad (\text{S20})$$

The random variables  $\bar{V} - \lambda(\bar{U} - \mu)$  and  $\lambda(\hat{\mu} - \mu)$  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 S3), there exist i.i.d. Gaussian random variables  $(G_i^{(1)})_{i \geq 1}$  with mean  $\gamma$  and variance  $\text{var}(V - \lambda U)$  such that

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

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

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

If  $r > 0$ , then, by Proposition S4, 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  $r\lambda^2 \text{var}(U)$  such that,

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

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

$$\nu^{\text{cv}+} := \nu_{\lambda^*}^{\text{cv}} = \text{var}(V) [1 - (1-r)\rho_{U,V}^2]. \quad (\text{S21})$$

We have

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

and

$$\frac{n\lambda^2}{N_n(N_n - 1)} \sum_{j=1}^{N_n} (\tilde{U}_j - \hat{\mu})^2 \rightarrow r\lambda^2 \text{var}(U) \text{ a.s.}$$

therefore  $\hat{\nu}_\lambda^{\text{cv}} \rightarrow \nu_\lambda^{\text{cv}}$  a.s. Finally, we show that  $\hat{\nu}^{\text{cv}+}$  is a consistent estimator of  $\nu^{\text{cv}+}$ . Set  $\delta_i = V_i - \bar{V} - \hat{\lambda}(U_i - \bar{U})$ . We have  $\delta_i = V_i - \hat{\alpha} - \hat{\beta}U_i$  where  $\hat{\alpha} = \bar{V} - \hat{\lambda}\bar{U}$  and  $\hat{\beta} = \hat{\lambda}$  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

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

we have

$$\frac{1}{n-2} \sum_{i=1}^n \delta_i^2 \rightarrow \mathbb{E}[(V - \alpha^* - \beta^* U)^2] = \text{var}(V)(1 - \rho_{U,V}^2) \text{ a.s. as } n \rightarrow \infty. \quad (\text{S22})$$

Additionally, by the strong law of large numbers,

$$\frac{n/N_n}{n-1} \sum_{i=1}^n (V_i - \bar{V})^2 \rightarrow r\text{var}(V) \text{ a.s. as } n \rightarrow \infty. \quad (\text{S23})$$

Hence,

$$\hat{\nu}^{\text{cv}+} \rightarrow (1-r)\text{var}(V)(1 - \rho_{U,V}^2) + r\text{var}(V) = \text{var}(V)(1 - (1-r)\rho_{U,V}^2)$$

almost surely as  $n \rightarrow \infty$ .

### S3.6 Proof of Proposition 4

We apply Theorem 1 to the i.i.d. sequence  $(\ell'_\theta(\tilde{X}_i, f(\tilde{X}_i)))_{i \geq 1}$ , to obtain an AsympCS  $(\tilde{\mathcal{R}}_{\delta, \theta, i})_{i \geq 1}$  for  $m_\theta$  with approximation rate  $1/\sqrt{i \log i}$ . The subsequence  $(\mathcal{R}_{\delta, \theta, n})_{n \geq 1}$  with  $\mathcal{R}_{\delta, \theta, n} = \tilde{\mathcal{R}}_{\delta, \theta, N_n}$  is also an AsympCS for  $m_\theta$  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  $\theta$  can be extended to obtain asymptotic confidence regions for  $\theta \in \mathbb{R}^d$ .

We first provide definitions of a multivariate confidence sequence and 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 S2.** (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 S3.** (Asymptotic Spherical Confidence Sequence) Let  $\alpha \in (0, 1)$  and  $(a_t)_{t \geq 0}$  be a real sequence such that  $\lim_{t \rightarrow \infty} a_t = 0$ . Let  $(\hat{\mu}_t)_{t \geq 1}$  be a consistent sequence of estimators of  $\mu$ . The sequence of random balls  $(C_{\alpha,t})_{t \geq 1}$ , with  $C_{\alpha,t} = B(\hat{\mu}_t, R_t)$  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  $(C_{\alpha,t}^*)_{t \geq 1}$ , with  $C_{\alpha,t}^* = B(\hat{\mu}_t, R_t^*)$ , such that

$$\Pr(\mu \in C_{\alpha,t}^* \text{ for all } t \geq 1) \geq 1 - \alpha$$

and

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

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

Let  $Y_1, Y_2, \dots$  be i.i.d. Gaussian random vectors with mean  $\mu \in \mathbb{R}^d$  and a known  $d$ -by- $d$  positive definite covariance matrix  $\Sigma$ . Let  $\pi$  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 to Theorem S6), the sequence of ellipsoid regions defined by

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

forms a  $(1 - \alpha)$  confidence sequence for  $\mu$ . One could also consider spherical confidence regions using  $\Lambda_{\max}(\Sigma)$ , the maximum eigenvalue of  $\Sigma$ , 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  $\mu$  is

$$C_{\alpha,t}^{\text{mBA}}(\bar{Y}_t, \Sigma; \pi) = \left\{ \mu \in \mathbb{R}^d \mid \|\mu - \bar{Y}_t\| \leq \frac{\sqrt{\Lambda_{\max}(\Sigma)}}{\sqrt{t}} \sqrt{\log \left( \frac{t^d}{(2\pi)^d \alpha^2 \eta_t(\Sigma^{-1/2} \bar{Y}_t)^2} \right)} \right\}. \quad (\text{S24})$$

#### 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 S7.** (Multivariate version of Theorem 4) Let  $(\hat{\mu}_t)_{t \geq 1}$  be a consistent sequence of estimators of  $\mu$ . Assume that there exists a sequence of i.i.d. Gaussian vectors with mean  $\mu$  and positive-definite covariance matrix  $\Sigma$  such that, a.s. as  $t \rightarrow \infty$ ,

$$\hat{\mu}_t = \frac{1}{t} \sum_{i=1}^t W_i + \varepsilon_t \quad \text{where} \quad \varepsilon_t = o\left(\frac{1}{\sqrt{t \log t}}\right). \quad (\text{S25})$$

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

*Proof.* (Theorem S7) By hypothesis, we have, almost surely,

$$\hat{\mu}_t = \bar{W}_t + \varepsilon_t,$$

where  $\varepsilon_t = o\left(\frac{1}{\sqrt{t \log t}}\right)$  and  $\bar{W}_t = \frac{1}{t} \sum_{i=1}^t W_i$ . As stated in Section S4.2, the sequence of balls  $C_{\alpha,t}^{\text{mBA}}(\bar{W}_t, \Sigma; \pi) = C_{\alpha,t}^{\text{mBA}}(\hat{\mu}_t - \varepsilon_t, \Sigma; \pi)$  forms an exact  $(1 - \alpha)$  CS for  $\mu$ . Let

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

As  $B(\hat{\mu}_t, R_t^*) \supseteq C_{\alpha,t}^{\text{mBA}}(\hat{\mu}_t - \varepsilon_t, \Sigma; \pi)$ , the sequence of random balls  $B(\hat{\mu}_t, R_t^*)$  also forms an exact  $(1 - \alpha)$  CS for  $\mu$ . Define  $C_{\alpha,t}^{\text{mBA}}(\hat{\mu}_t, \hat{\Sigma}_t; \pi) = B(\hat{\mu}_t, R_t)$ , where

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

By the Courant-Fischer theorem, we have,

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

Additionally, by Lemma S2, we have

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

It therefore follows that

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

a.s. Then  $|R_t - R_t^*| = o(1/\sqrt{t \log t})$  and thus  $\mathcal{C}_{\alpha,t}^{\text{mBA}}(\hat{\mu}_t, \hat{\Sigma}_t; \pi)$  is a  $(1 - \alpha)$ -AsympCS with approximation rate  $1/\sqrt{t \log t}$ .  $\square$

**Theorem S8.** (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  $\mu$  and such that  $\mathbb{E}\|Y_1\|^{2+\delta} < \infty$  for some  $\delta > 0$ . Then,  $\mathcal{C}_{\alpha,t}^{\text{mBA}}(\bar{Y}_t, \hat{\Sigma}_t; \pi)$  is a  $(1 - \alpha)$ -AsympCS with approximation rate  $1/\sqrt{t \log t}$ , where  $\bar{Y}_t$  is the sample mean and  $\hat{\Sigma}_t$  the sample covariance.*

*Proof.* (Theorem S8) By the strong law of large numbers,  $\bar{Y}_t$  and  $\hat{\Sigma}_t$  are consistent estimators of  $\mu$  and  $\Sigma$ , respectively. By the multivariate KMT coupling due to Einmahl [5] (see Theorem S4), there exists a sequence of i.i.d. Gaussian random vectors  $(W_i)_{i \geq 1}$  with mean  $\mu$  and covariance matrix  $\Sigma$  such that

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

The result then follows from Theorem S7.  $\square$

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 using

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

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

$$m_\theta = \theta - \mathbb{E}[f(X)], \quad (\text{S26})$$

whereas the rectifier  $\Delta_\theta$  is given by

$$\Delta_\theta = \mathbb{E}[f(X) - Y]. \quad (\text{S27})$$

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

As discussed in Section 4, PPI uses the sample mean as an estimator of  $m_\theta$ , which here is

$$\hat{m}_{\theta,n} = \theta - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j). \quad (\text{S28})$$

For  $\Delta_0$ , either the PPI estimator  $\hat{\Delta}_{0,n}^{\text{PP}}$  (20) or the PPI++ estimator  $\hat{\Delta}_{0,n}^{\text{PP+}}$  (23) may be used. In the case of mean estimation, these are given by

$$\hat{\Delta}_{0,n}^{\text{PP}} = \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i), \quad (\text{S29})$$

$$\hat{\Delta}_{0,n}^{\text{PP+}} = \hat{\Delta}_n^{\text{PP}} - (\hat{\lambda}_{\theta,n} - 1) \left( \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) - \frac{1}{n} \sum_{i=1}^n f(X_i) \right) \quad (\text{S30})$$

$$= \frac{1}{n} \sum_{i=1}^n (\hat{\lambda}_{\theta,n} f(X_i) - Y_i) - (\hat{\lambda}_{\theta,n} - 1) \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j), \quad (\text{S31})$$

where

$$\hat{\lambda}_{\theta,n} = \frac{\widehat{\text{cov}}((\ell'_\theta(X_i, Y_i), \ell'_\theta(X_i, f(X_i)))_{i=1}^n)}{\widehat{\text{var}}((\ell'_\theta(X_i, f(X_i)))_{i=1}^n)} = \frac{\widehat{\text{cov}}((Y_i, f(X_i))_{i=1}^n)}{\widehat{\text{var}}((f(X_i))_{i=1}^n)}. \quad (\text{S32})$$

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

Given  $\hat{m}_{\theta,n}$  and an estimator  $\hat{\Delta}_{0,n}$  of  $\Delta_0$ , the associated prediction-powered estimator of  $\theta^*$  is found by solving for  $\theta$  the equation

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

For the two estimators of  $\Delta_0$  discussed above, this quantity takes the form

$$\hat{g}_{\theta,n}^{\text{PP}} = \theta - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) + \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i) \quad (\text{S33})$$

$$= \theta - \frac{1}{n} \sum_{i=1}^n Y_i + \left( \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) \right), \quad (\text{S34})$$

$$\hat{g}_{\theta,n}^{\text{PP+}} = \theta - \hat{\lambda}_{0,n} \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) + \frac{1}{n} \sum_{i=1}^n (\hat{\lambda}_{0,n} f(X_i) - Y_i) \quad (\text{S35})$$

$$= \theta - \frac{1}{n} \sum_{i=1}^n Y_i + \hat{\lambda}_{0,n} \left( \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) \right), \quad (\text{S36})$$

whose zeroes are given by

$$\hat{\theta}_n^{\text{PP}} = \frac{1}{n} \sum_{i=1}^n Y_i - \left( \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) \right), \quad (\text{S37})$$

$$\hat{\theta}_n^{\text{PP+}} = \frac{1}{n} \sum_{i=1}^n Y_i - \hat{\lambda}_{0,n} \left( \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) \right). \quad (\text{S38})$$

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

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

Section 5.1 defines valid AsympCS for  $g_\theta$  that do not incorporate prior information. In particular,  $\mathcal{C}_{\alpha,\theta,n}^g$  is constructed as

$$\mathcal{C}_{\alpha,\theta,n}^g = \mathcal{C}_{\alpha,t}^{\text{NA}}(\hat{g}_{\theta,n}, \hat{\sigma}_{\theta,n}^g; \rho) \quad (\text{S39})$$

$$= \left[ \hat{g}_{\theta,n} \pm \frac{\hat{\sigma}_{\theta,n}^g}{\sqrt{n}} \sqrt{\left(1 + \frac{1}{n\rho^2}\right) \log\left(\frac{n\rho^2 + 1}{\alpha^2}\right)} \right], \quad (\text{S40})$$

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

$$(\hat{\sigma}_{\theta,n}^g)^2 = \frac{1}{n-2} \sum_{i=1}^n \left( Y_i - f(X_i) - \frac{1}{n} \sum_{k=1}^n (Y_k - f(X_k)) \right)^2 \quad (\text{S41})$$

$$+ \frac{n/N_n}{N_n-1} \sum_{j=1}^{N_n} \left( f(\tilde{X}_j) - \frac{1}{N_n} \sum_{k=1}^{N_n} f(\tilde{X}_k) \right)^2, \quad (\text{S42})$$

whereas, for  $\hat{g}_{\theta,n}^{\text{PP+}}$ , we obtain

$$(\hat{\sigma}_{\theta,n}^g)^2 = \frac{1-n/N_n}{n-2} \sum_{i=1}^n \left( Y_i - \hat{\lambda}_{0,n} f(X_i) - \frac{1}{n} \sum_{k=1}^n (Y_k - \hat{\lambda}_{0,n} f(X_k)) \right)^2 \quad (\text{S43})$$

$$+ \frac{n/N_n}{n-1} \sum_{i=1}^n \left( Y_i - \frac{1}{n} \sum_{k=1}^n Y_k \right)^2. \quad (\text{S44})$$

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

$$\mathcal{C}_{\alpha,n}^{\text{avpp}} = \left\{ \theta \mid 0 \in \mathcal{C}_{\alpha,\theta,n}^g \right\} \quad (\text{S45})$$

$$= \left[ \hat{\theta}_n \pm \frac{\hat{\sigma}_{0,n}^g}{\sqrt{n}} \sqrt{\left(1 + \frac{1}{n\rho^2}\right) \log\left(\frac{n\rho^2 + 1}{\alpha^2}\right)} \right], \quad (\text{S46})$$

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

Similarly, Section 5.2 defines valid AsympCS for  $g_\theta$  that incorporate prior information via a zero-mean prior  $\pi$  on  $\Delta_\theta$ . In particular, for  $\delta \in (0, \alpha)$ , Proposition 4 first constructs a standard  $(1 - \delta)$  AsympCS  $\mathcal{R}_{\delta,\theta,n}$  for  $m_\theta$ , which, in the case of mean estimation, takes the form

$$\mathcal{R}_{\delta,\theta,n} = \mathcal{C}_{\delta,n}^{\text{NA}}(\hat{m}_{\theta,n}, \hat{\sigma}_{\theta,n}^f; \rho) \quad (\text{S47})$$

$$= \left[ \theta - \frac{1}{N_n} \sum_{j=1}^{N_n} f(\tilde{X}_j) \pm \frac{\hat{\sigma}_{\theta,n}^f}{\sqrt{N_n}} \sqrt{\left(1 + \frac{1}{N_n\rho^2}\right) \log\left(\frac{N_n\rho^2 + 1}{\delta^2}\right)} \right], \quad (\text{S48})$$

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

$$(\hat{\sigma}_{\theta,n}^f)^2 = \widehat{\text{var}}((\ell'_\theta(\tilde{X}_i, f(\tilde{X}_i)))_{i=1}^{N_n}) \quad (\text{S49})$$

$$= \frac{1}{N_n-1} \sum_{j=1}^{N_n} \left( f(\tilde{X}_j) - \frac{1}{N_n} \sum_{k=1}^{N_n} f(\tilde{X}_k) \right)^2. \quad (\text{S50})$$

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

$$\mathcal{T}_{\alpha-\delta,\theta,n} = \mathcal{C}_{\alpha-\delta,n}^{\text{BA}}(\hat{\Delta}_{0,n}, \hat{\sigma}_{\theta,n}^\Delta; \pi) \quad (\text{S51})$$

$$= \left[ \hat{\Delta}_{0,n} \pm \frac{\hat{\sigma}_{\theta,n}^\Delta}{\sqrt{n}} \sqrt{\log\left(\frac{n(2\pi(\alpha - \delta)^2)^{-1}}{\eta_n(\hat{\Delta}_{0,n}/\hat{\sigma}_{\theta,n}^\Delta)^2}\right)} \right], \quad (\text{S52})$$

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

$$(\hat{\sigma}_{\theta,n}^\Delta)^2 = \frac{1}{n-1} \sum_{i=1}^n \left( Y_i - f(X_i) - \frac{1}{n} \sum_{k=1}^n (Y_k - f(X_k)) \right)^2, \quad (\text{S53})$$

for PPI, and is given by

$$(\hat{\sigma}_{\theta,n}^\Delta)^2 = \frac{1-n/N_n}{n-2} \sum_{i=1}^n \left( Y_i - \hat{\lambda}_{0,n} f(X_i) - \frac{1}{n} \sum_{k=1}^n (Y_k - \hat{\lambda}_{0,n} f(X_k)) \right)^2 \quad (\text{S54})$$

$$+ \frac{n/N_n}{n-1} \sum_{i=1}^n \left( Y_i - f(X_i) - \frac{1}{n} \sum_{k=1}^n (Y_k - f(X_k)) \right)^2 \quad (\text{S55})$$

for PPI++. Finally,  $\mathcal{T}_{\alpha-\delta,\theta,n}$  and  $\mathcal{R}_{\delta,\theta,n}$  are combined by Minkowski summation to obtain a valid  $(1-\alpha)$ -AsympCS for  $g_\theta$  as

$$\mathcal{C}_{\alpha,\theta,n}^g = \mathcal{T}_{\alpha-\delta,\theta,n} + \mathcal{R}_{\delta,\theta,n} \quad (\text{S56})$$

$$= \left[ \hat{g}_{\theta,n} \pm \left\{ \frac{\hat{\sigma}_{\theta,n}^\Delta}{\sqrt{n}} \sqrt{\log \left( \frac{n(2\pi(\alpha-\delta)^2)^{-1}}{\eta_n(\hat{\Delta}_{0,n}/\hat{\sigma}_{\theta,n}^\Delta)^2} \right)} + \frac{\hat{\sigma}_{\theta,n}^f}{\sqrt{N_n}} \sqrt{\left( 1 + \frac{1}{N_n \rho^2} \right) \log \left( \frac{N_n \rho^2 + 1}{\delta^2} \right)} \right\} \right], \quad (\text{S57})$$

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

$$\mathcal{C}_{\alpha,n}^{\text{avPP}} = \left\{ \theta \mid 0 \in \mathcal{C}_{\alpha,\theta,n}^g \right\} \quad (\text{S58})$$

$$= \left[ \hat{\theta}_n \pm \left\{ \frac{\hat{\sigma}_{\theta,n}^\Delta}{\sqrt{n}} \sqrt{\log \left( \frac{n(2\pi(\alpha-\delta)^2)^{-1}}{\eta_n(\hat{\Delta}_{0,n}/\hat{\sigma}_{\theta,n}^\Delta)^2} \right)} + \frac{\hat{\sigma}_{\theta,n}^f}{\sqrt{N_n}} \sqrt{\left( 1 + \frac{1}{N_n \rho^2} \right) \log \left( \frac{N_n \rho^2 + 1}{\delta^2} \right)} \right\} \right], \quad (\text{S59})$$

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

## 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<sup>2</sup>, while all the others are available as part of the ppi-python package<sup>3</sup>.

**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.  $\theta^* = \mathbb{E}[Y] \in \mathbb{R}$ .

<sup>2</sup><https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

<sup>3</sup><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.  $\theta^* = \mathbb{E}[Y] \in [0, 1]$ .

**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.  $\theta^* = \mathbb{E}[Y] \in [0, 1]$ .

**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  $\Delta_\theta$ , 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  $\rho$  for Equation (6) in Theorem 1. As mentioned at the end of Section 3.1,  $\rho$  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

$$\rho = \sqrt{\frac{-W_{-1}(-\alpha^2 \exp(-1)) - 1}{t^*}}, \quad (\text{S60})$$

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

The Bayes-assisted prediction-powered AsympCS discussed in Section 5.2 requires the specification of both the parameter  $\rho$  used for the non-assisted AsympCS for the measure of fit  $m_\theta$  (Proposition 4) and the scale parameter  $\tau$  of the prior  $\pi$  for the Bayes-assisted AsympCS for the rectifier  $\Delta_\theta$  (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  $\bar{Y}_t/\hat{\sigma}_t$  in Equation (9) of Theorem 3. Instead, we propose the following heuristic for choosing  $\tau$ . If  $Z_1, Z_2, \dots | \mu \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$ , then the posterior mean  $\mathbb{E}[\mu | Z_1, \dots, Z_t]$  after  $t$  observations under a Gaussian prior  $\mu \sim \mathcal{N}(\mu_0, \tau^2)$  is given by

$$\mathbb{E}[\mu | Z_1, \dots, Z_t] = \frac{1}{1 + t\tau^2} \mu_0 + \frac{t\tau^2}{1 + t\tau^2} \bar{Z}_t, \quad (\text{S61})$$

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 (S61) is of the same form as the one implicitly used for the construction of Bayes-assisted AsympCS (see e.g. Equation (8)), we choose  $\tau$  so that the prior and the data have the same influence on the posterior mean (S61) at time  $t^*$ , i.e.

$$\tau = \frac{1}{\sqrt{t^*}}. \quad (\text{S62})$$

Section S7 reports the hyperparameter value used for each experiment in terms of  $t^*$ . 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  $m_\theta$ . As a result of this, the only hyperparameter that needs to be chosen for the Bayes-assisted procedure is the prior scale  $\tau$ .

## 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  $t^* = 500$  (see Section S6.4) for all methods.

Figure S2 reports the average cumulative miscoverage rate for the results in Figure 1. As desired, the cumulative miscoverage rate remains below the threshold  $\alpha$  for all  $n$ .

Figure S3 shows the performance of the Bayes-assisted prediction-powered AsympCS procedures under a Gaussian prior on the noisy predictions experiment in Section 6.1. These 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  $\sigma_Y$ . 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  $t^* = 500$  (see Section S6.4) for all methods shown in Figures 2 and S4. The values of  $t^*$  used for Figure S5 are reported in the figure legend.

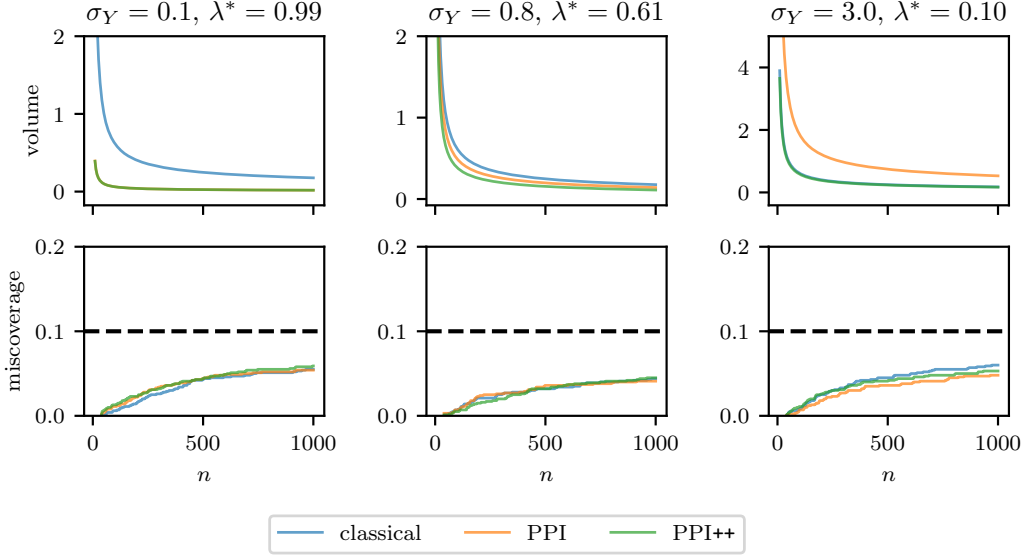


Figure S2: Noisy predictions study. The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for noise levels  $\sigma_Y = 0.1, 0.8, 3.0$ .

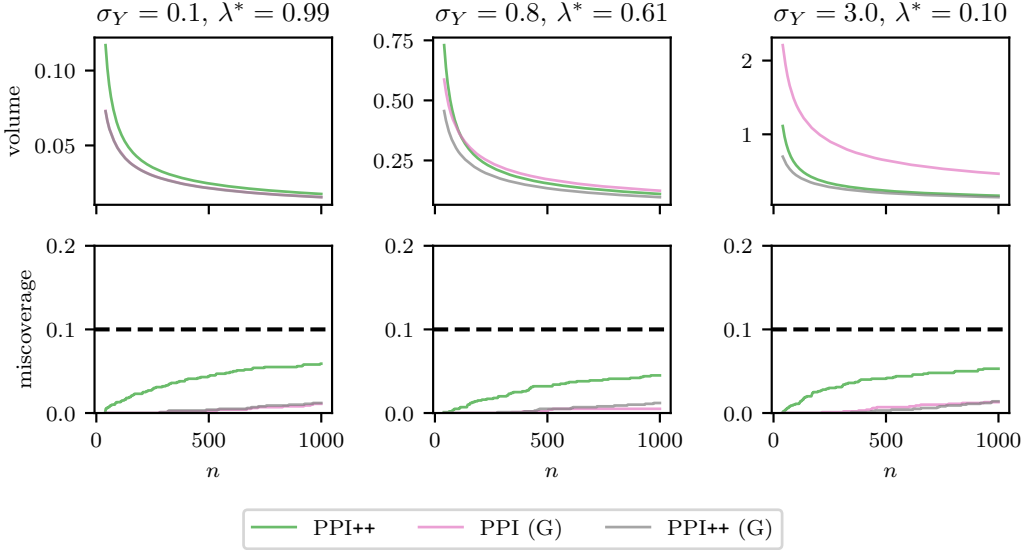


Figure S3: Noisy predictions study with Gaussian prior. The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for noise levels  $\sigma_Y = 0.1, 0.8, 3.0$ . Results for non-assisted PPI++ are included for reference.

Figure S4 reports the average cumulative miscoverage rate for the results shown in Figure 2 at  $v = 0$ . As discussed in Section 6.1, the cumulative miscoverage rate increases slightly as we decrease df, but remains below the threshold  $\alpha$  for all  $n$ , as desired.

Figure S5 repeats the simulation of Figure S4 for different values of  $t^*$ , which affects the procedure-specific hyperparameters as discussed in Section S6.4. For non-assisted PPI,  $t^*$  represents the time at which the procedure's interval width is minimised. Therefore, as expected, increasing  $t^*$  above 100 leads to larger intervals at  $n = 100$ . However, the qualitative behaviour of the non-assisted methods as  $v$  varies is the same across all values of  $t^*$ : their volumes remain constant across bias

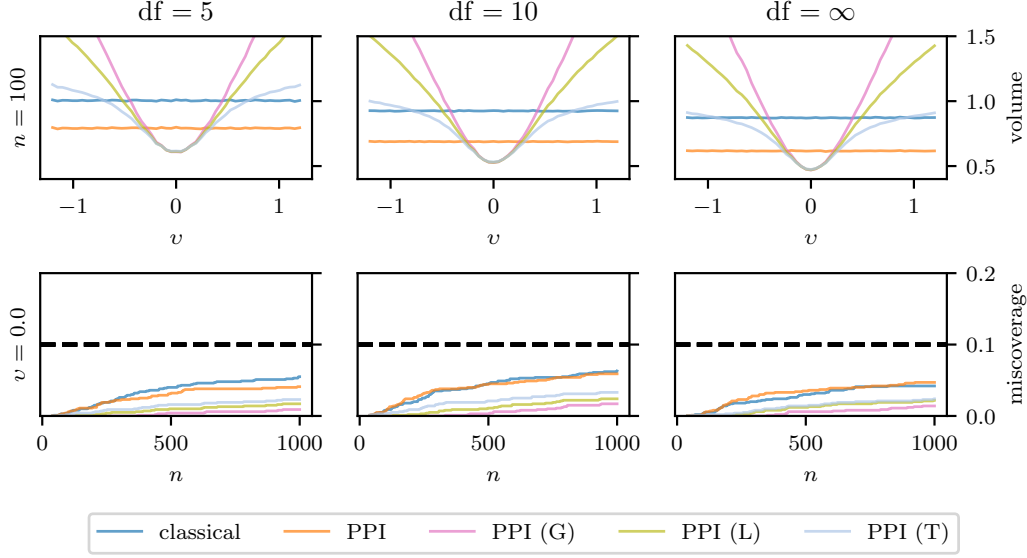


Figure S4: Biased predictions study. The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for  $df = 5, 10, \infty$ .

levels, reflecting the lack of prior information. On the other hand, for Bayes-assisted methods, a larger  $t^*$  implies a smaller prior scale  $\tau$ . As a result of this, increasing  $t^*$  above 100 leads to stronger prior influence at  $n = 100$ . In particular, a large  $t^*$  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.

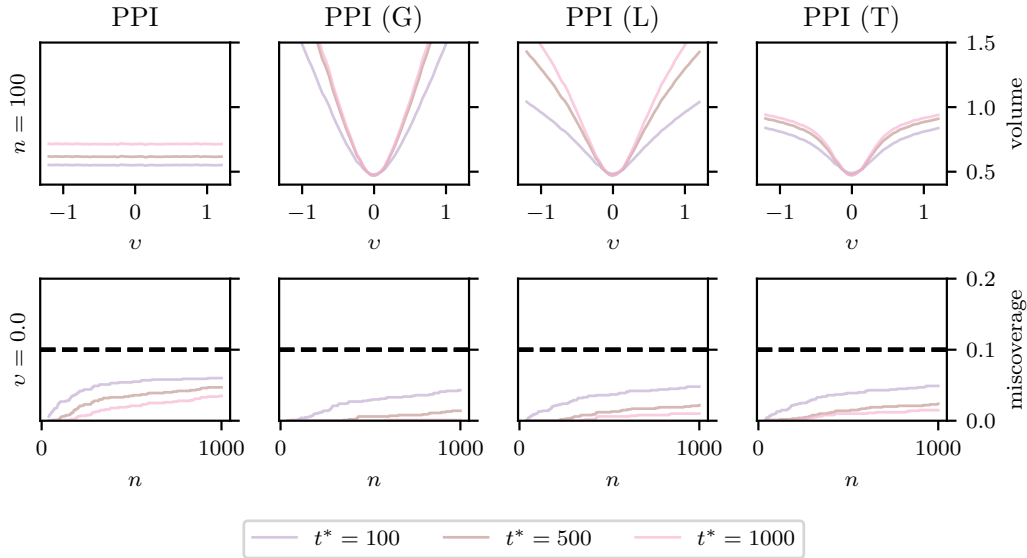


Figure S5: Biased predictions study with different hyperparameters. Each column correspond to one of the prediction-powered methods in Figure 2 for  $t^* = 100, 500, 1000$ . The top and bottom columns show average interval volume and cumulative miscoverage rate over 1000 repetitions with  $df = \infty$ .

### S7.1.3 Multivariate biased predictions

Here, we illustrate the multivariate AsympCS procedure described in Section 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  $Y_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$ , 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  $\epsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_d)$ , so that  $\theta^* = \mathbb{E}[Y] = \mathbf{0}$ . Then, we proceed as in Section 6.1 and define biased predictions  $f(X_i) = X_i + v$ , where  $v \in \mathbb{R}$  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 Section S4. The AsympCS hyperparameters are set using natural extensions of the rules in Section S6.4 to the multivariate case, with  $t^* = 1000$  for all methods. Figure 2 shows the average spherical interval volumes of each method as a function of  $v$ , which we vary between  $-6$  and  $6$ , at  $n \in \{100, 250, 500\}$ . These results are consistent with those

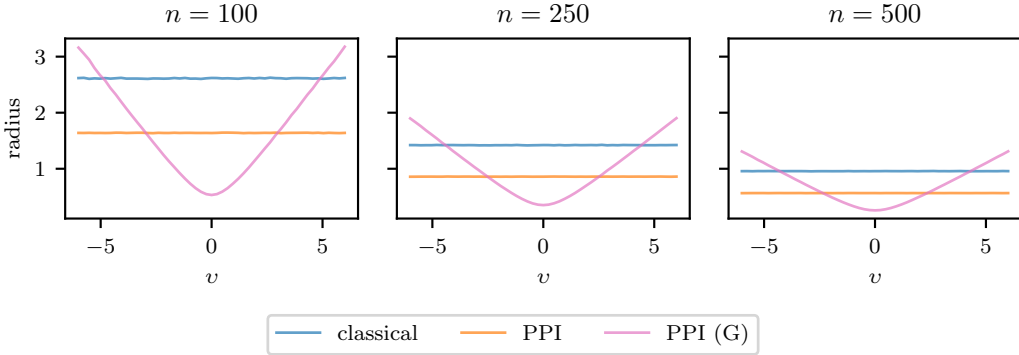


Figure S6: Multivariate biased predictions study. The left, middle and right panels show average spherical region radius over 1000 repetitions for  $n = 100, 250, 500$ .

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  $v$ , but its radius grows quickly with  $|v|$  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  $v$  considered, with cumulative miscoverage rates close to zero, likely due to the conservative spherical construction mentioned in Section S4.

## S7.2 Real data

### S7.2.1 Mean estimation

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

Figure S7 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  $t^* = 2000$  (see Section S6.4), as that is the largest  $n$  considered in all experiments.

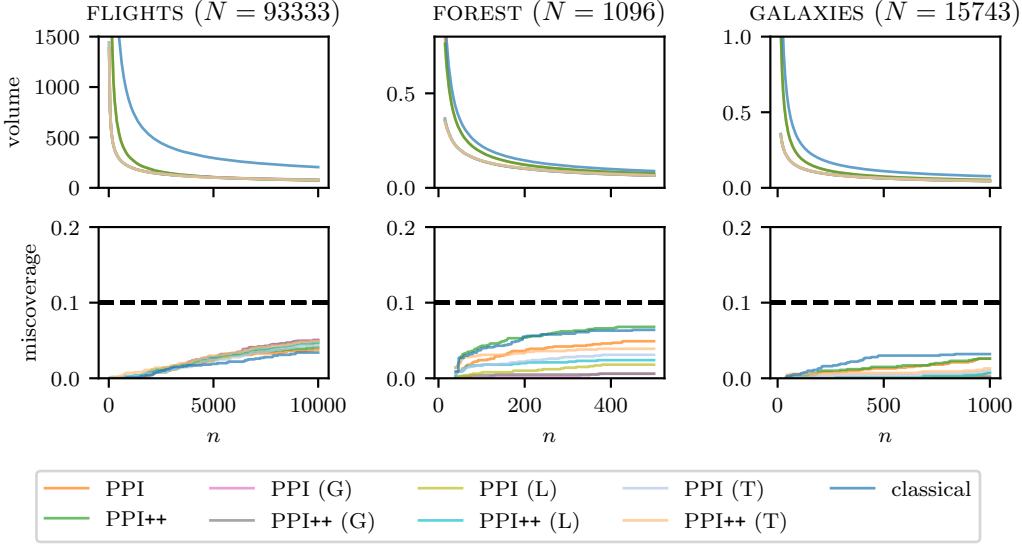


Figure S7: Mean estimation. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 1000 repetitions for the FLIGHTS, FOREST, and GALAXIES datasets.

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  $\theta$  through Equation (4). To initialise the grid, we use the first  $n_0$  labelled data points to compute a preliminary estimate of  $\theta^*$ , 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  $\eta_t$  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  $\theta$  values simultaneously. Because of this, we only report results for the Gaussian and Laplace priors, which admit closed-form expressions for  $\eta_t$ .

Figure S8 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 yields 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 Section 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 S9, the results are strikingly different: all methods yield significantly larger cumulative miscoverage rates compared to the preprocessed case in Figure S8, 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

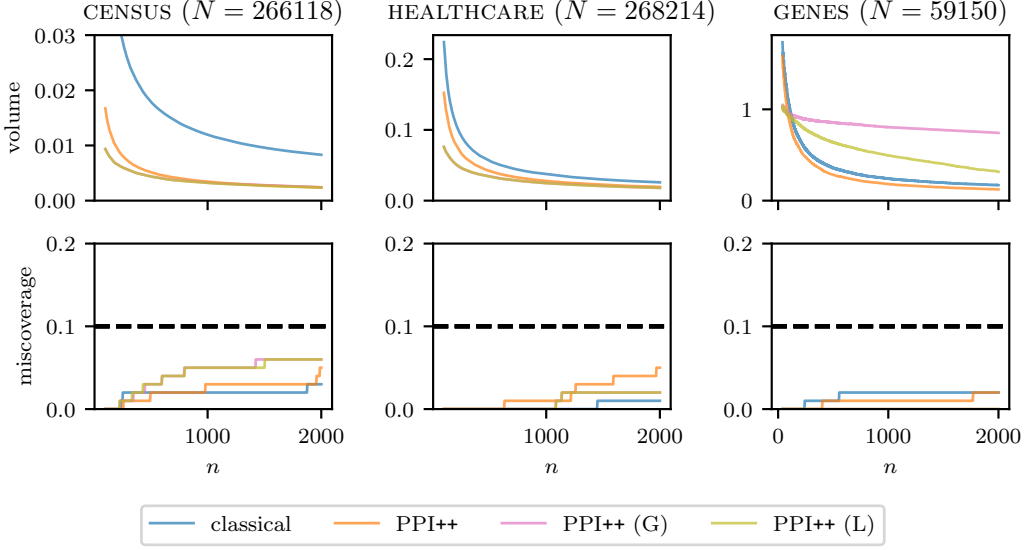


Figure S8: Other estimation tasks. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 100 repetitions for the CENSUS, HEALTHCARE, and GENES datasets.

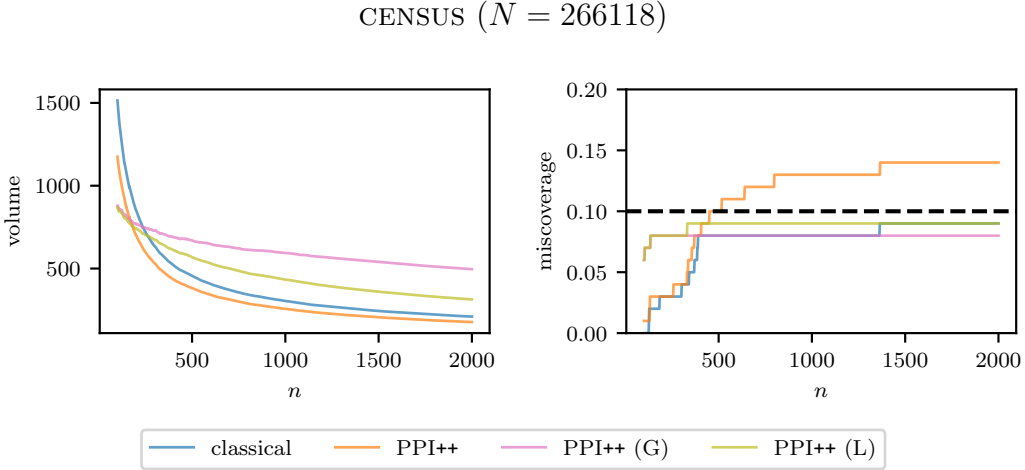


Figure S9: Linear regression on the CENSUS dataset without preprocessing. The top and bottom rows show the average interval volume and cumulative miscoverage rate over 100 repetitions.

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  $\mathcal{C}_{\alpha,t}^{\text{NA}}(\bar{Y}_t, \hat{\sigma}_t; \rho)$  in Equation (6) approximates the exact CS in Equation (S3) 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  $g: [1, \infty) \rightarrow (0, 1]$  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  $\Phi$  and  $\phi$  the standard normal CDF and PDF, and let  $z_\alpha := g^{-1}(\alpha)$ , which is well-defined. The corresponding AsympCS is given by

$$\mathcal{C}_{\alpha,t}^{\text{NA}'}(\bar{Y}_t, \hat{\sigma}_t) := \left[ \bar{Y}_t \pm \frac{\hat{\sigma}_t}{\sqrt{t}} \sqrt{\log(tz_\alpha^2)} \right]. \quad (\text{S63})$$

Notably, for i.i.d. Gaussian observations with known variance, the exact (nonasymptotic) counterpart of (S63) 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  $\mathcal{C}_{\alpha,t}^{\text{NA}'}$  from (S63) as a parameter-free drop-in replacement for  $\mathcal{C}_{\alpha,t}^{\text{NA}}$  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

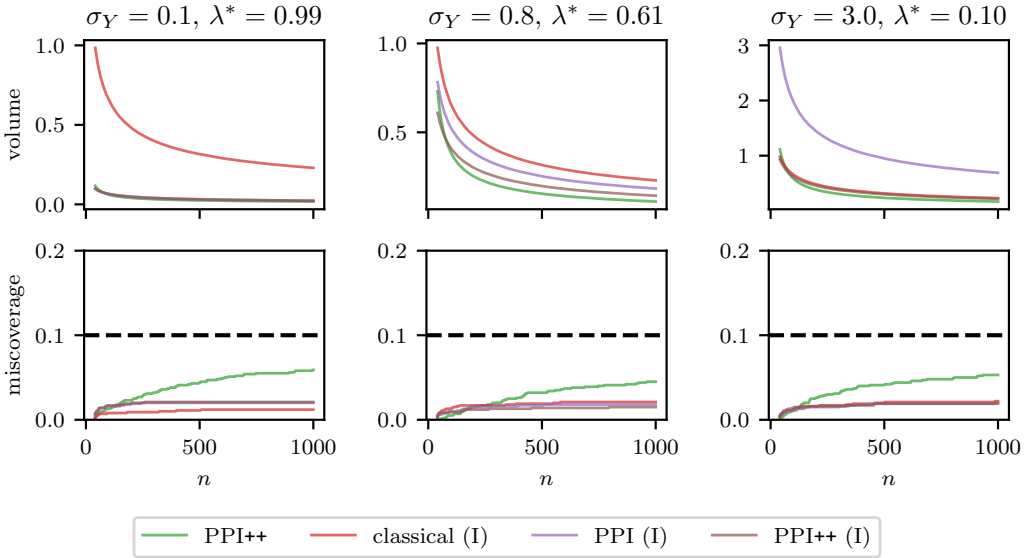


Figure S10: Noisy predictions study with alternative AsympCS . The left, middle and right panels show average interval volume and cumulative miscoverage rate over 1000 repetitions for noise levels  $\sigma_Y = 0.1, 0.8, 3.0$ . Results for non-assisted PPI++ based on Equation (6) are shown for reference.

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

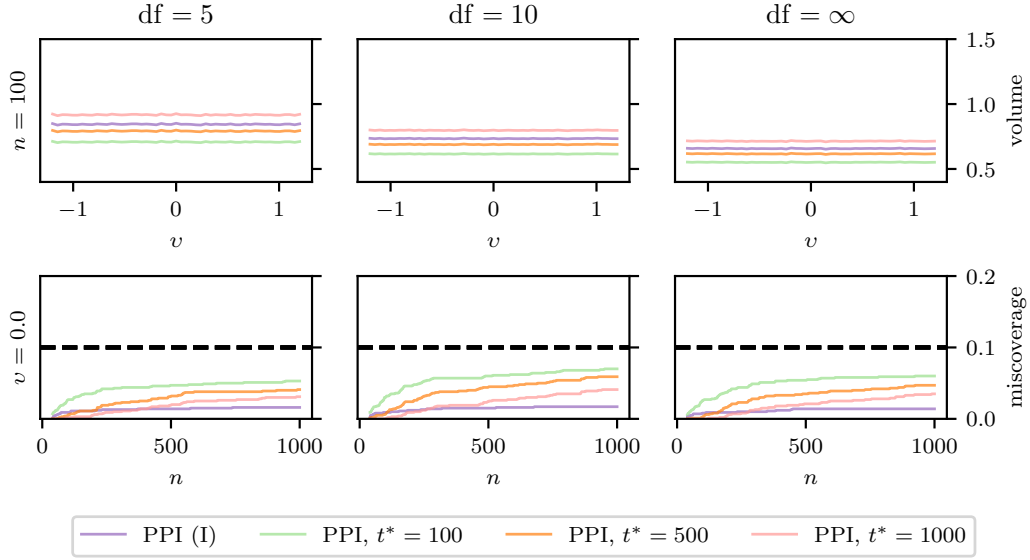


Figure S11: Biased predictions study with alternative AsympCS . The left, middle and right panels show average interval volume and cumulative miscoverage rate over 100 repetitions for  $df = 5, 10, \infty$ . Results for non-assisted PPI based on Equation (6) are shown for reference.

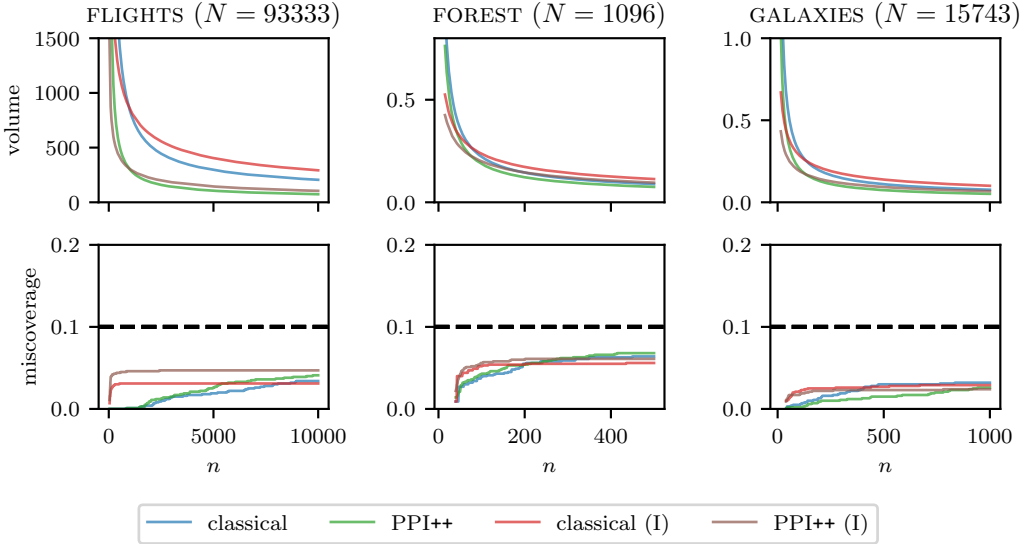


Figure S12: Real data study with alternative AsympCS . The top and bottom rows show the average interval volume and cumulative miscoverage rate over 1000 repetitions for the FLIGHTS, FOREST, and GALAXIES datasets. Results for classical inference and non-assisted PPI++ based on Equation (6) are shown 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  $\mathbb{R}^d$  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.