A. Proofs of main results

We first introduce a lemma which will aid in the proofs to follow.

Lemma 1 (Predictable plug-in Chernoff supermartingales). *Suppose that $X_1, X_2, \cdots \sim P$, and for some μ, v_t and $\psi(\lambda)$, we have that for any $\lambda \in \Lambda \subseteq \mathbb{R}$,*

$$\mathbb{E}_P \left[\exp(\lambda(X_t - \mu) - v_t\psi(\lambda)) \mid \mathcal{F}_{t-1}\right] \leq 1 \quad \text{for each } t \geq 1 \ . \tag{35}$$

Then, for any Λ-valued sequence $(\lambda_t)_{t=1}^\infty$ that is predictable with respect to $\mathcal{F}$,

$$M_t^\psi(\mu) := \prod_{i=1}^t \exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right)$$

forms a test supermartingale with respect to $\mathcal{F}$.

Proof. Writing out the conditional expectation of M_t^ψ for any $t \geq 2$,

$$\begin{aligned}\mathbb{E}\left(M_t^\psi(\mu) \mid \mathcal{F}_{t-1}\right) &= \mathbb{E}\left(\prod_{i=1}^t \exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right) \,\middle|\, \mathcal{F}_{t-1}\right)\\ &\overset{(i)}{=} \prod_{i=1}^{t-1} \exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right) \underbrace{\mathbb{E}\left[\exp\left(\lambda_t(X_t - \mu) - v_t\psi(\lambda_t)\right) \mid \mathcal{F}_{t-1}\right]}_{\leq 1 \text{ by assumption}}\\ &= M_{t-1}^\psi(\mu),\end{aligned}$$

where (i) follows from the fact that $\exp\left(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)\right)$ is $\mathcal{F}_{t-1}$-measurable for $i \leq t-1$. Since $\mathcal{F}_0$ was assumed to be trivial, for M_1 we have that

$$\mathbb{E}[M_1^\psi(\mu) | \mathcal{F}_0] = \underbrace{\mathbb{E}\left[\exp\left(\lambda_1(X_1 - \mu) - v_1\psi(\lambda_1)\right)\right]}_{\leq 1 \text{ by assumption}},$$

which completes the proof. □

A.1. Proof of Proposition 1

The proof proceeds in three steps. First, apply a standard MGF bound by Hoeffding [1963]. Second, we apply Lemma 1. Finally, we apply Theorem 1 to obtain a CS and take a union bound.

Step 1. By Hoeffding [1963], we have that $\mathbb{E}\left[\exp(\lambda_t(X_t - \mu) - \psi_H(\lambda_t)) \mid \mathcal{F}_{t-1}\right] \leq 1$ since $X_t \in [0, 1]$ almost surely and since λ_t is $\mathcal{F}_{t-1}$-measurable.

Step 2. By Step 1 and Lemma 1, we have that

$$M_t^{\text{PrPl-H}}(\mu) := \prod_{i=1}^t \exp\left(\lambda_i(X_i - \mu) - \psi_H(\lambda_i)\right)$$

forms a test supermartingale.

Step 3. By Step 2 combined with Theorem 1, we have that

$$P\left(\exists t \geq 1 : \mu \leq \frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right)$$
$$= P\left(\exists t \geq 1 : M_t^{\text{PrPl-H}}(\mu) \geq 1/\alpha\right) \leq \alpha.$$

Applying the same bound to $(-X_t)_{t=1}^\infty$ with mean $-\mu$ and taking a union bound, we have the desired result,

$$P\left(\exists t \geq 1 : \mu \notin \left(\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right)\right) \leq \alpha,$$

which completes the proof. □

A.2. Proof of Theorem 2

By Lemma 1 combined with Theorem 1, it suffices to prove that

$$\mathbb{E}_P\left[\exp\left\{\lambda_t(X_t - \mu) - v_t\psi_E(\lambda_t)\right\} \mid \mathcal{F}_{t-1}\right] \leq 1.$$

For succinctness, denote

$$Y_t := X_t - \mu \quad \text{and} \quad \delta_t := \widehat{\mu}_t - \mu.$$

Note that $\mathbb{E}_P(Y_t \mid \mathcal{F}_{t-1}) = 0$. It then suffices to prove that for any $[0,1)$-bounded, $\mathcal{F}_{t-1}$- measurable $\lambda_t \equiv \lambda_t(X_1^{t-1})$,

$$\mathbb{E}\left[\exp\left\{\lambda_t Y_t - 4(Y_t - \delta_{t-1})^2\psi_E(\lambda_t)\right\} \,\middle|\, \mathcal{F}_{t-1}\right] \leq 1.$$

Indeed, in the proof of Proposition 4.1 in Fan et al. [2015], $\exp\{\xi\lambda - 4\xi^2\psi_E(\lambda)\} \leq 1 + \xi\lambda$ for any $\lambda \in [0,1)$ and $\xi \geq -1$. Setting $\xi := Y_t - \delta_{t-1} = X_t - \widehat{\mu}_{t-1}$,

$$\mathbb{E}\left[\exp\left\{\lambda_t Y_t - 4(Y_t - \delta_{t-1})^2\psi_E(\lambda_t)\right\} \,\middle|\, \mathcal{F}_{t-1}\right]$$
$$= \mathbb{E}\Big[\exp\Big\{\lambda_t(Y_t - \delta_{t-1}) - 4(Y_t - \delta_{t-1})^2\psi_E(\lambda_t)\Big\} \mid \mathcal{F}_{t-1}\Big]\exp(\lambda_t\delta_{t-1})$$
$$\leq \mathbb{E}\Big[1 + (Y_t - \delta_{t-1})\lambda_t \mid \mathcal{F}_{t-1}\Big]\exp(\lambda_t\delta_{t-1}) \overset{(i)}{=} \mathbb{E}\big[1 - \delta_{t-1}\lambda_t \mid \mathcal{F}_{t-1}\big]\exp(\lambda_t\delta_{t-1}) \overset{(ii)}{\leq} 1,$$

where equality (i) follows from the fact that Y_t is conditionally mean zero, and inequality (ii) follows from the inequality $1 - x \leq \exp(-x)$ for all $x \in \mathbb{R}$. This completes the proof. □

A.3. Proof of Proposition 2

We proceed by proving $(d) \implies (c) \implies (b) \implies (a) \implies (d)$.

Proof of $(d) \implies (c)$. This claim follows from the fact that for $\lambda \in (-1/(1-\mu), 1/\mu)$, we have that $(\lambda, \lambda, \dots)$ is a $(-1/(1-\mu), 1/\mu)$-valued predictable sequence.

Proof of $(c) \implies (b)$. By the assumption of (c), we have that for $\lambda = 0.5$, $\mathcal{K}_t(\mu)$ forms a test martingale. Furthermore, since $X_i, \mu \in [0,1]$ for each $i \in \{1, 2, \dots\}$, we have that $1 + 0.5(X_i - \mu) > 0$ almost surely for each i. Therefore, $(\mathcal{K}_t(\mu))_{t=1}^{\infty}$ is a strictly positive test martingale.

Proof of $(b) \implies (a)$. Suppose that there exists $\lambda \in \mathbb{R} \setminus \{0\}$ such that $\mathcal{K}_t(\mu)$ forms a strictly positive martingale. Then we must have

$$\begin{aligned} \mathcal{K}_{t-1}(\mu) &= \mathbb{E}\left(\mathcal{K}_t(\mu) \mid \mathcal{F}_{t-1}\right) \\ &= \mathcal{K}_{t-1}(\mu) \cdot \mathbb{E}\left(1 + \lambda(X_i - \mu) \mid \mathcal{F}_{t-1}\right) \\ &= \mathcal{K}_{t-1}(\mu) \cdot \left[1 + \lambda(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu)\right]. \end{aligned}$$

Now since $\mathcal{K}_{t-1}(\mu) > 0$, we have that

$$1 + \lambda(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu) = 1.$$

Since $\lambda \neq 0$ by assumption, we have that $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = \mu$ as required.

Proof of $(a) \implies (d)$. Let $(\lambda_t(\mu))_{t=1}^{\infty}$ be a $(-1/(1-\mu), 1/\mu)$-valued predictable sequence. Then $\mathcal{K}_t(\mu)$ is clearly nonnegative and $\mathcal{K}_0(\mu) = 1$ by definition. Writing out the conditional mean of the capital process for any $t \geq 1$,

$$\begin{aligned} \mathbb{E}\left(\mathcal{K}_t(\mu) \mid \mathcal{F}_{t-1}\right) &= \mathcal{K}_{t-1}(\mu) \cdot \mathbb{E}\left(1 + \lambda_t(\mu)(X_i - m) \mid \mathcal{F}_{t-1}\right) \\ &= \mathcal{K}_{t-1}(\mu) \cdot \left[1 + \lambda_t(\mu)(\mathbb{E}(X_i \mid \mathcal{F}_{t-1}) - \mu)\right] \\ &= \mathcal{K}_{t-1}(\mu), \end{aligned}$$

and thus $\mathcal{K}_t(\mu)$ forms a test martingale.

The proof of the final part of the proposition is simple. Let (M_t) be a test martingale for $\mathcal{P}^\mu$. Define $Y_t := M_t/M_{t-1}$ if $M_{t-1} > 0$, and as $Y_t := 0$ otherwise. Now note that $M_t = \prod_{i=1}^t Y_t$ and $\mathbb{E}_P[Y_t | \mathcal{F}_{t-1}] = 1$ for any $P \in \mathcal{P}^\mu$. In other words, every test martingale is a product of nonnegative random variables with conditional mean one. Now rewrite Y_t as $(1 + f_t(X_t))$ for some predictable function f_t. Since Y_t is nonnegative, we must have $f_t(X_t) \geq -1$, and since Y_t is conditional mean one, we must have $f_t(X_t)$ is conditional mean zero. Such a representation in fact holds true for any test martingale, and we have not yet used the fact that we are working with test martingales for $\mathcal{P}^\mu$. Now, the proof ends by noting that the only predictable functions f_t with the latter property under every $P \in \mathcal{P}^\mu$ has the form $\lambda_t(X_t - \mu)$ for some predictable λ_t; any nonlinear function of X_t would not have mean zero under *every distribution* with mean μ.

This completes the proof of Proposition 2 altogether. □

A.4. Proof of Proposition 3

We only prove the martingale part of the proposition, since the supermartingale aspect follows analogously, and as mentioned early in the paper, inequalities and equalities are meant in an almost sure sense.

First, it is easy to check that if (M_t) is a test martingale for $\mathcal{S}$, then M_t is the product of nonnegative conditionally unit mean terms, that is $M_t = \prod_{i=1}^t Y_i$ such that for all $S \in \mathcal{S}$, we have $\mathbb{E}_S[Y_i|\mathcal{F}_{i-1}] = 1$ and $Y_i \geq 0$. (Indeed, one can identify $Y_i := \frac{M_i}{M_{i-1}}\mathbf{1}_{M_{i-1}>0}$.) Now, define $Z_i' := Y_i - 1$, and note that $Z_i' \geq -1$, and $\mathbb{E}_S[Z_i'|\mathcal{F}_{t-1}] = 0$. Thus, M_t has been represented as $\prod_{i=1}^t(1+Z_i')$. Now, the proof is completed by noting that any such Z_i' can be written as $\lambda_i Z_i$ for a predictable λ_i (this step is purely cosmetic). □

A.5. Proof of Theorem 3

First, we present Lemma 2 which establishes that the hedged capital process is a quasiconvex function of m (and thus has convex sublevel sets). We then invoke this lemma to prove the main result.

LEMMA 2. *Let $\theta \in [0,1]$ and*

$$\begin{aligned}\mathcal{K}_t^{\pm}(m) &:= \max\left\{\theta\mathcal{K}_t^{+}(m), (1-\theta)\mathcal{K}_t^{-}(m)\right\}\\ &\equiv \max\left\{\theta\prod_{i=1}^t(1+\lambda_i^+(m)\cdot(X_i-m)), (1-\theta)\prod_{i=1}^t(1-\lambda_i^-(m)\cdot(X_i-m))\right\}\end{aligned}$$

be the hedged capital process as in Section 4. Consider the $(1-\alpha)$ confidence set of the same theorem,

$$\mathfrak{B}_t^{\pm} \equiv \mathfrak{B}^{\pm}(X_1,\dots,X_t) := \left\{m \in [0,1] : \mathcal{K}_t^{\pm}(m) < \frac{1}{\alpha}\right\}.$$

Then $\mathfrak{B}_t^{\pm}$ is an interval on $[0,1]$.

PROOF. Since sublevel sets of quasiconvex functions are convex, it suffices to prove that $\mathcal{K}_t^{\pm}(m)$ is a quasiconvex function of $m \in [0,1]$. The crux of the argument is: the product of nonnegative nonincreasing functions is quasiconvex, the product of nonnegative nondecreasing functions is also quasiconvex, and the maximum of quasiconvex functions is quasiconvex.

To elaborate, we will proceed in two steps. First, we use an induction argument to show that $\mathcal{K}_t^{+}(m)$ and $\mathcal{K}_t^{-}(m)$ are nonincreasing and nondecreasing, respectively, and hence quasiconvex. Finally, we note that $\mathcal{K}_t^{\pm}(m) := \max\left\{\theta\mathcal{K}_t^{+}(m), (1-\theta)\mathcal{K}_t^{-}(m)\right\}$ is a maximum of quasiconvex functions and is thus itself quasiconvex.

Step 1. First, since $\dot{\lambda}_t^+$ does not depend on m, we have that

$$1+\lambda_t^+(m)(X_t-m) := 1 + \left(|\dot{\lambda}_t^+| \wedge \frac{c}{m}\right)(X_t-m)$$

is nonnegative and nonincreasing in m for each $t \in \{1, 2, \dots\}$. (To see this, consider the terms with and without truncation separately.) Suppose for the sake of induction that

$$\prod_{i=1}^{t-1} \left(1 + \lambda_i^+(m)(X_i - m)\right)$$

is nonnegative and nonincreasing in m. Then,

$$\begin{aligned}\mathcal{K}_t^+(m) &:= \prod_{i=1}^{t} \left(1 + \lambda_i^+(m)(X_i - m)\right) \\ &= \left(1 + \lambda_t^+(m)(X_t - m)\right) \cdot \prod_{i=1}^{t-1} \left(1 + \lambda_i^+(m)(X_i - m)\right)\end{aligned}$$

is a product of nonnegative and nonincreasing functions, and is thus itself nonnegative and nonincreasing. By a similar argument, $\mathcal{K}_t^-(m)$ is nonnegative and *nondecreasing*. $\mathcal{K}_t^+(m)$ and $\mathcal{K}_t^-(m)$ are thus both quasiconvex.

Step 2. Since the maximum of quasiconvex functions is quasiconvex, we infer that

$$\mathcal{K}_t^{\pm}(m) := \max\left\{\theta \mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\right\}$$

is quasiconvex. In particular, the sublevel sets of quasiconvex functions is convex, and thus

$$\mathfrak{B}_t^{\pm} := \left\{ m \in [0,1] : \mathcal{K}_t^{\pm}(m) < \frac{1}{\alpha} \right\}$$

is an interval, which completes the proof of Lemma 2. □

Proof (Theorem 3). The proof proceeds in three steps. First we show that $\mathcal{K}_t^{\pm}(\mu)$ is upper-bounded by test martingale. Second, we apply the 4-step procedure in Theorem 1 to get a CS for μ. Third and finally, we invoke Lemma 2 to conclude that the CS is indeed convex at each time t.

Step 1. We first upper bound $\mathcal{K}_t^{\pm}(m)$ as follows:

$$\begin{aligned}\mathcal{K}_t^{\pm}(m) &:= \max\left\{\theta \mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\right\} \\ &\leq \theta \mathcal{K}_t^+(m) + (1-\theta)\mathcal{K}_t^-(m) =: \mathcal{M}_t^{\pm}(m).\end{aligned}$$

By Proposition 2, we have that $\mathcal{K}_t^+(\mu)$ and $\mathcal{K}_t^-(\mu)$ are test martingales for $\mathcal{P}$. For each $P \in \mathcal{P}$, writing out the conditional expectation of $\mathcal{M}_t^{\pm}(\mu)$ for any $t \geq 1$,

$$\begin{aligned}\mathbb{E}_P\left[\mathcal{M}_t^{\pm}(\mu) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}_P\left[\theta \mathcal{K}_t^+(\mu) + (1-\theta)\mathcal{K}_t^-(\mu) \;\middle|\; \mathcal{F}_{t-1}\right] \\ &= \theta \mathbb{E}_P(\mathcal{K}_t^+(\mu) \mid \mathcal{F}_{t-1}) + (1-\theta)\mathbb{E}_P(\mathcal{K}_t^-(\mu) \mid \mathcal{F}_{t-1}) \\ &= \theta \mathcal{K}_{t-1}^+(\mu) + (1-\theta)\mathcal{K}_{t-1}^-(\mu) \\ &= \mathcal{M}_{t-1}^{\pm}(\mu),\end{aligned}$$

and $\mathcal{M}_0^{\pm}(\mu) = \theta \mathcal{K}_0^+(\mu) + (1-\theta)\mathcal{K}_0^-(\mu) = 1$. Therefore, $(\mathcal{M}_t^{\pm}(\mu))_{t=0}^{\infty}$ is a test martingale for $\mathcal{P}$.

Step 2. By Step 1 combined with Theorem 1 we have that

$$\mathfrak{B}_t^{\pm} := \left\{ m \in [0,1] : \mathcal{K}_t^{\pm}(m) < \frac{1}{\alpha} \right\}$$

forms a $(1-\alpha)$-CS for μ.

Step 3. Finally, by Lemma 2, we have that $\mathfrak{B}_t^{\pm}$ is an interval for each $t \in \{1, 2, \dots\}$, which completes the proof of Theorem 3. □

A.6. Proof of Lemma 3

Following the proof of Lemma 4.1 in Fan et al. [2015], we have that the function

$$f(x) := \begin{cases} \dfrac{\log(1+x) - x}{x^2/2} & x \in (-1, \infty) \setminus \{0\} \\ -1 & x = 0 \end{cases} \tag{36}$$

is an increasing and continuous function in x (note that $f(0)$ is defined as -1 because it is a removable singularity). For any $y \geq -m$ and $\lambda \in [0, 1/m)$ we have

$$\lambda y \geq -m\lambda > -1. \tag{37}$$

Combining (36) and (37), we have

$$\frac{\log(1+\lambda y) - \lambda y}{\lambda^2 y^2/2} \geq \frac{\log(1-m\lambda) + m\lambda}{\lambda^2 m^2/2},$$
$$\text{and thus,} \quad \log(1+\lambda y) - \lambda y \overset{(i)}{\geq} \frac{y^2}{m^2}\left(\log(1-m\lambda) + m\lambda\right).$$

Above, (i) can be quickly verified for the case when $\lambda y = 0$, and follows from (36) and (37) otherwise. Rearranging terms, we obtain the first half of the desired result,

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{m^2}(\log(1-m\lambda) + m\lambda). \tag{38}$$

Now, for any $y \leq 1-m$ and $\lambda \in (-1/(1-m), 0]$, we have

$$\lambda y \geq (1-m)\lambda > -1,$$

and proceed similarly to before to obtain

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{(1-m)^2}(\log(1+(1-m)\lambda) - (1-m)\lambda),$$

which completes the proof. □

A.7. Proof of Proposition 5

Since sublevel sets of convex functions are convex, it suffices to prove that with probability one, $\mathcal{K}_n^{\text{hgKelly}}(m)$ is a convex function in m on the interval $[0, 1]$.

We proceed in three steps. First, we show that if two functions are (a) both nonincreasing (or both nondecreasing), (b) nonnegative, and (c) convex, then their product is convex. Second, we use Step 1 and an induction argument to prove that $\prod_{i=1}^t (1 + \gamma(X_i/m - 1))$ is convex for any fixed $\gamma \in [0, 1]$. Third and finally, we show that $\mathcal{K}_n^{\text{hgKelly}}(m)$ is a convex combination of convex functions and is thus itself convex.

Step 1. The claim is that if two functions f and g are (a) both nonincreasing (or both nondecreasing), (b) nonnegative, and (c) convex on a set $\mathcal{S} \subseteq \mathbb{R}$, then their product is also convex on $\mathcal{S}$. Let $x_1, x_2 \in \mathcal{S}$, and let $t \in [0, 1]$. Furthermore, abbreviate $f(x_1)$ by f_1, $g(x_1)$ by g_1, and similarly for f_2 and g_2. Writing out the product fg evaluated at $tx_1 + (1-t)x_2$,

$$\begin{aligned}(fg)(tx_1 + (1-t)x_2) &= f(tx_1 + (1-t)x_2)g(tx_1 + (1-t)x_2)\\ &= |f(tx_1 + (1-t)x_2)||g(tx_1 + (1-t)x_2)|\\ &\le |tf_2 + (1-t)f_2||tg_1 + (1-t)g_2|\\ &= t^2 f_1 g_1 + t(1-t)\left(f_1 g_2 + f_2 g_1\right) + (1-t)^2 f_2 g_2,\end{aligned}$$

where the second equality follows from assumption that f and g are nonnegative, and the inequality follows from the assumption that they are both convex. To show convexity of (fg), it then suffices to show that,

$$\Big(tf_1 g_1 + (1-t)f_2 g_2\Big) - \Big(t^2 f_1 g_1 + t(1-t)\left[f_1 g_2 + f_2 g_1\right] + (1-t)^2 f_2 g_2\Big) \ge 0. \quad (39)$$

To this end, write out the above expression and group terms,

$$\begin{aligned}&\Big(tf_1 g_1 + (1-t)f_2 g_2\Big) - \Big(t^2 f_1 g_1 + t(1-t)\left[f_1 g_2 + f_2 g_1\right] + (1-t)^2 f_2 g_2\Big)\\ &= (1-t)tf_1 g_1 + t(1-t)f_2 g_2 - t(1-t)[f_1 g_2 + f_2 g_1]\\ &= t(1-t)\Big(f_1 g_1 + f_2 g_2 - f_1 g_2 - f_2 g_1\Big)\\ &= t(1-t)(f_1 - f_2)(g_1 - g_2).\end{aligned}$$

Now, notice that $t(1-t) \ge 0$ since $t \in [0, 1]$ and that $(f_1 - f_2)(g_1 - g_2) \ge 0$ by the assumption that f and g are both nonincreasing or nondecreasing. Therefore, we have satisfied the inequality in (39), and thus fg is convex on $\mathcal{S}$.

Step 2. Now, we prove convexity of $\prod_{i=1}^t (1 + \gamma(X_i/m - 1))$ for a fixed $\gamma \in [0, 1]$. First note that for any $\gamma \in [0, 1]$, $1 + \gamma(X_i/m - 1)$ is a nonincreasing, nonnegative, and convex function in $m \in [0, 1]$. Suppose for the sake of induction that conditions (a), (b), and (c) hold for $\prod_{i=1}^{n-1} (1 + \gamma(X_i/m - 1))$. By the inductive hypothesis, we have that

$$\prod_{i=1}^{n}(1+\gamma(X_i/m-1)) = (1+\gamma(X_n/m-1)) \cdot \prod_{i=1}^{n-1}(1+\gamma(X_i/m-1))$$

is a product of functions satisfying (a) through (c). By Step 1, $\prod_{i=1}^{n}(1+\gamma(X_i/m-1))$ is convex in $m \in [0,1]$. A similar argument can be made for $\mathcal{K}_n^-(m)$, but instead of the multiplicands being nonincreasing, they are now nondecreasing.

Step 3. Now, notice that for the evenly-spaced points $(\lambda^{1+}, \dots, \lambda^{G+})$ on $[0, 1/m]$, we have that $(\gamma^{1+}, \dots, \gamma^{G+}) = (m\lambda^{1+}, \dots, m\lambda^{G+})$ are G evenly-spaced points on $[0,1]$. It then follows that for any m and any $g \in \{0, 1, \dots, G\}$,

$$m \mapsto \prod_{i=1}^{n}(1+\lambda^{g+}(X_i - m))$$

is a nonincreasing, nonnegative, and convex function in $m \in [0,1]$. It follows that

$$\frac{1}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g+}(X_i - m))$$

is convex in $m \in [0,1]$. A similar argument goes through for $\frac{1}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g+}(X_i - m)$. Finally, since $\theta \in [0,1]$, we have that

$$\frac{\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g+}(X_i - m)) + \frac{1-\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{n}(1+\lambda^{g-}(X_i - m))$$

is a convex combination of convex functions in $m \in [0,1]$. It then follows that

$$\{m \in [0,1] : \mathcal{K}_t^{\text{hgKelly}}(m) < 1/\alpha\}$$

is an interval, which completes the proof. □

A.8. Proof of Proposition 4

Proof of (1) $\implies$ (2). By definition of $\mathcal{K}_t^{\text{WoR}}(\mu)$, we have

$$\begin{aligned}\mathbb{E}\left(\mathcal{K}_t^{\text{WoR}}(\mu) \mid \mathcal{F}_{t-1}\right) &= \prod_{i=1}^{t-1}\left(1+\lambda_i(\mu)\cdot(X_i - \mu_t^{\text{WoR}})\right)\cdot \mathbb{E}\left(1+\lambda_t(\mu)\cdot(X_t - \mu_t^{\text{WoR}}) \mid \mathcal{F}_{t-1}\right) \\ &= \mathcal{K}_{t-1}^{\text{WoR}}(\mu)\cdot\left(1+\lambda_t(\mu)\cdot(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu_t^{\text{WoR}})\right) \\ &= \mathcal{K}_{t-1}^{\text{WoR}}(\mu).\end{aligned}$$

Since $\mathcal{K}_0^{\text{WoR}}(\mu) \equiv 1$ by convention, we have that $\mathcal{K}_t^{\text{WoR}}(\mu)$ is a martingale.

Now, note that since $X_t \in [0,1]$ and $\lambda_t^{\mathrm{WoR}}(\mu) \in [-1/(1-\mu_t^{\mathrm{WoR}}), 1/\mu_t^{\mathrm{WoR}}]$ for each t by assumption, we have that $1 + \lambda_t(\mu) \cdot \left(X_t - \mu_t^{\mathrm{WoR}}\right) \geq 0$ and thus $\mathcal{K}_t^{\mathrm{WoR}}(\mu) \geq 0$. Therefore, $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$ is a test martingale.

Proof of (2) $\implies$ (1). Suppose that $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$ is a test martingale for any $(\lambda_t(\mu))_{t=1}^N$ with $\lambda_t(\mu) \in [-1/(1-\mu_t^{\mathrm{WoR}}, 1/\mu_t^{\mathrm{WoR}}]$, but suppose for the sake of contradiction that $\mathbb{E}(X_{t^\star} \mid \mathcal{F}_{t^\star-1}) \neq \mu_{t^\star}^{\mathrm{WoR}}$ for some $t^\star \in \{1,2,\dots\}$. Set $\lambda_1 = \lambda_2 = \cdots = \lambda_{t^\star-1} = 0$ and $\lambda_{t^\star} = 1$. Then,

$$\mathcal{K}_{t^\star}^{\mathrm{WoR}}(\mu) \equiv \mathcal{K}_{t^\star-1}^{\mathrm{WoR}}(\mu) \cdot (1 + \lambda_{t^\star}(X_{t^\star} - \mu_{t^\star}^{\mathrm{WoR}})) = 1 + X_{t^\star} - \mu_{t^\star}^{\mathrm{WoR}}.$$

By assumption of $\mathcal{K}_t^{\mathrm{WoR}}(\mu)$ forming a martingale, we have that $\mathbb{E}\left(\mathcal{K}_{t^\star}^{\mathrm{WoR}}(\mu) \mid \mathcal{F}_{t^\star-1}\right) = \mathcal{K}_{t^\star-1}^{\mathrm{WoR}}(\mu) = 1$. On the other hand, since $\mathbb{E}\left(X_{t^\star} \mid \mathcal{F}_{t^\star-1}\right) \neq \mu_{t^\star}^{\mathrm{WoR}}$, we have

$$\mathbb{E}\left(\mathcal{K}_{t^\star}^{\mathrm{WoR}}(\mu) \mid \mathcal{F}_{t^\star-1}\right) = \mathbb{E}\left(1 + X_{t^\star} - \mu_{t^\star}^{\mathrm{WoR}} \mid \mathcal{F}_{t^\star-1}\right) \neq 1,$$

a contradiction. Therefore, we must have that $\mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right) = \mu_t^{\mathrm{WoR}}$ for each t, which completes the proof of (2) $\implies$ (1) and Proposition 4. □

A.9. Proof of Theorem 4

The proof that $\mathfrak{B}_t^{\pm,\mathrm{WoR}}$ forms a $(1-\alpha)$-CS for μ proceeds in exactly the same manner as Theorem 3, noting that $\mathbb{E}\left(X_t \mid \mathcal{F}_{t-1}\right) = \mu_t^{\mathrm{WoR}}$ instead of μ.

To show that $\mathfrak{B}_t^{\pm,\mathrm{WoR}}$ is indeed an interval for each $t \geq 1$, we note that the proof of Theorem 3 applies since m_t^{WoR} is increasing or decreasing if and only if m is increasing or decreasing, respectively. □

B. How to bet: deriving adaptive betting strategies

In Section 4.4, we presented CSs and CIs via the hedged capital process. We suggested a specific betting scheme which has strong empirical performance but did not discuss where it came from. In this section, we derive various betting strategies and discuss their statistical and computational properties.

B.1. Predictable plug-ins yield good betting strategies

First and foremost, we will examine why any predictable plug-in for empirical Bernstein-type CSs and CIs (i.e. those recommended in Theorem 2 and Remark 1) yield effective betting strategies. Consider the hedged capital process

$$\begin{aligned}\mathcal{K}_t^{\pm}(m) &:= \max\left\{\theta \prod_{i=1}^t (1 + \lambda_i^+(X_i - m)), (1-\theta)\prod_{i=1}^t (1 - \lambda_i^-(X_i - m))\right\} \\ &\equiv \max\left\{\theta \mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\right\} ,\end{aligned}$$

where $(\lambda_t^+(m))_{t=1}^\infty$ and $(\lambda_t^-(m))_{t=1}^\infty$ are $[0, 1/m]$-valued and $[0, 1/(1-m)]$-valued predictable sequences as in Theorem 3. First, consider the "positive" capital process, $\mathcal{K}_t^+(\mu)$ evaluated at $m = \mu$. An inequality that has been repeatedly used to derive empirical Bernstein inequalities [Howard et al., 2020, 2021, Waudby-Smith and Ramdas, 2020], including the current paper is the following due to Fan et al. [2015, equation 4.12]: for any $y \geq -1$ and $\lambda \in [0, 1)$, we have

$$\log(1 + \lambda y) \geq \lambda y - 4\psi_E(\lambda) y^2. \tag{40}$$

where $\psi_E(\lambda)$ is as defined in (14). If the predictable sequence $(\lambda_t^+(m))_{t=1}^\infty$ is further restricted to $[0, 1)$, then by (40) we have

$$\begin{aligned}\mathcal{K}_t^+(\mu) := \prod_{i=1}^t (1 + \lambda_i^+(X_i - \mu)) &\geq \exp\left(\sum_{i=1}^t \lambda_i^+(X_i - \mu) - \sum_{i=1}^t 4(X_i - \mu)^2 \psi_E(\lambda_i^+) \right) \\ &\overset{(i)}{\approx} \exp\left(\sum_{i=1}^t \lambda_i^+(X_i - \mu) - \sum_{i=1}^t 4(X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i^+) \right) \\ &= M_t^{\text{PrPl-EB}}(\mu),\end{aligned}$$

where (i) follows from the approximations $\widehat{\mu}_{t-1} \approx \mu$ for large t. Not only does the approximate inequality $\mathcal{K}_t^+(\mu) \gtrsim M_t^{\text{PrPl-EB}}(\mu)$ shed light on why a sensible empirical Bernstein predictable plug-in translates to a sensible betting strategy, but also why we might expect $\mathcal{K}_t^+(m)$ to be more powerful than $M_t^{\text{PrPl-EB}}(m)$ for the same $[0, 1)$-valued predictable sequence $(\lambda_t^+(m))_{t=1}^\infty$. Moreover, $\mathcal{K}_t^+(m)$ has the added flexibility of allowing $(\lambda_t(m))_{t=1}^\infty$ to take values in $[0, 1/m] \supset [0, 1)$ which we find — through simulations — tends to improves empirical performance (see Figure 19 in Section E.2.2). Finally, a similar story holds for $\mathcal{K}_t^-(\mu)$ with the added caveat that $(\lambda_t^-)_{t=1}^\infty$ can instead take values in $[0, 1/(1 - m)] \supset [0, 1)$ which as before, seems to improve empirical performance.

Despite the success of predictable plug-ins as betting strategies, it is natural to wonder whether it is preferable to focus on directly maximizing capital over time. As will be seen in the following section, these capital-maximizing approaches tend to have improved empirical performance, but are not always guaranteed to produce convex confidence sets (i.e. intervals). Nevertheless, it is worth examining some of these strategies both for their intuitive appeal and excellent empirical performance.

B.2. Growth rate adaptive to the particular alternative (GRAPA)

As alluded to in Section 6, Kelly Jr [1956] dealt with capital processes, betting strategies, etc. in the fields of information and communication theory in the pursuit of maximizing the information rate over a channel. Kelly suggested that an effective betting strategy is one that maximizes a gambler's expected *log-capital* — i.e. the growth rate of the gambler's capital — under a particular alternative.[§] However, Kelly's setup was a simplified special case of ours: Kelly's observations were binary, and the exact alternative was assumed known, while ours are merely bounded in

[§] This objective has also been arrived at indirectly as the dual in optimization programs for deriving regret bounds for Kullback-Leibler-based UCB algorithms in multi-armed bandit problems [Honda and Takemura, 2010, Cappé et al., 2013].

$[0,1]$ with an unknown alternative. Nevertheless, the principle of maximizing the log-capital can be adapted to our setting under bounded observations and an unknown alternative. We summarize this adaptation here and refer to it as maximizing the "growth rate adaptive to the particular alternative" or "GRAPA" for short.

Write the log-capital process at time t as

$$\ell_t(\lambda_1^t, m) := \log(\mathcal{K}_t(m)) = \sum_{i=1}^{t} \log(1 + \lambda_i(m)(X_i - m)), \tag{41}$$

for a general $[-1/(1-m), 1/m]$-valued sequence $(\lambda_t(m))_{t=1}^{\infty}$. If we were to choose a single value of $\lambda^{\mathrm{HS}} := \lambda_1 = \cdots = \lambda_t$ which maximizes the log-capital ℓ_t "in hindsight" (i.e. based on *all* of the previous data), then this value is given by

$$\frac{\partial \ell_t(\lambda^{\mathrm{HS}}, m)}{\partial \lambda^{\mathrm{HS}}} = \sum_{i=1}^{t} \frac{X_i - m}{1 + \lambda^{\mathrm{HS}}(X_i - m)} \overset{\text{set}}{=} 0.$$

However, λ^{HS} is clearly not predictable. Following Kumon et al. [2011] (who referred to this as the "sequential optimization strategy"), we set $(\lambda_t^{\mathrm{GRAPA}}(m))_{t=1}^{\infty}$ such that

$$\frac{1}{t-1} \sum_{i=1}^{t-1} \frac{X_i - m}{1 + \lambda_t^{\mathrm{GRAPA}}(m)(X_i - m)} \overset{\text{set}}{=} 0, \tag{42}$$

truncated to lie between $(-c/(1-m), c/m)$ using some $c \leq 1$. Importantly, $\lambda_t^{\mathrm{GRAPA}}(m)$ only depends on $X_1, \ldots, X_{t-1}$, and is thus predictable.

This rule is a sequentially adaptive version of the worst-case "GROW" criterion of Grünwald et al. [2019]. To see the connection, one can derive (42) from a slightly different motivation. At the t-th step, we want to choose $\lambda_t(m)$ so that the wealth multiplier $(1 + \lambda_t(m)(X_t - m))$ is as large as possible. The ideal choice would be

$$\lambda_t^*(m) := \operatorname*{argmax}_{\lambda \in [-1/(1-m), 1/m]} \mathbb{E}_{P^\mu}[\log(1 + \lambda(X_t - m)) \mid \mathcal{F}_{t-1}], \tag{43}$$

where P^μ is the unknown true distribution. Writing down the stationary condition for this optimization problem by differentiating through the expectation, we get

$$\mathbb{E}_{P^\mu}\left[\frac{X_t - m}{1 + \lambda_t^*(m)(X_t - m)} \mid \mathcal{F}_{t-1}\right] = 0. \tag{44}$$

Since P^μ is unknown, using a simple empirical plug-in estimator yields (42).

CSs constructed from $(\lambda_t^{\mathrm{GRAPA}}(m))_{t=1}^{\infty}$ tend to have excellent empirical performance, but can be prohibitively slow due to the required root-finding in (42) for each time t and $m \in [0,1]$ (or a sufficiently fine grid of $[0,1]$). A similar but computationally inexpensive alternative to GRAPA is "approximate GRAPA" (aGRAPA), which we derive now.

B.3. Approximate GRAPA (aGRAPA)

Rather than solve (42), we take the Taylor approximation of $(1+y)^{-1}$ by $(1-y)$ for $y \approx 0$ to obtain

$$\begin{aligned}\frac{1}{t-1}\sum_{i=1}^{t-1}\frac{X_i - m}{1+\lambda_t^{\mathrm{aGRAPA}}(m)(X_i - m)} &\approx \frac{1}{t-1}\sum_{i=1}^{t-1}\left(1-\lambda_t^{\mathrm{aGRAPA}}(m)(X_i - m)\right)(X_i - m)\\ &= \frac{1}{t-1}\sum_{i=1}^{t-1}(X_i - m) - \frac{\lambda_t^{\mathrm{aGRAPA}}(m)}{t-1}\sum_{i=1}^{t-1}(X_i - m)^2\\ &\overset{\text{set}}{=} 0,\end{aligned}$$

which, after appropriate truncation leads what we call the "approximate GRAPA" (aGRAPA) betting strategy,

$$\lambda_t^{\mathrm{aGRAPA}}(m) := -\frac{c}{1-m} \vee \frac{\widehat{\mu}_{t-1} - m}{\widehat{\sigma}_{t-1}^2 + (\widehat{\mu}_{t-1} - m)^2} \wedge \frac{c}{m},$$

for some truncation level $c \leq 1$. This expression is quite natural: we bet more aggressively if our empirical mean is far away from m, and are further emboldened if the empirical variance is small.

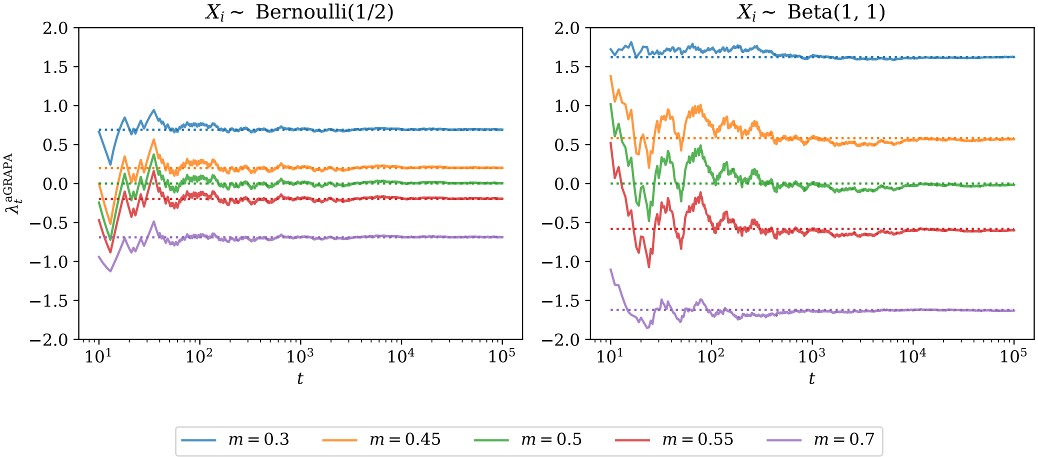

Figure 10. $\lambda_t^{\mathrm{aGRAPA}}$ for various values of m under two distributions: Bernoulli(1/2) and Beta(1, 1). The dotted lines show the "oracle" bets, meaning $\lambda_t^{\mathrm{aGRAPA}}$ with estimates of the mean and variance replaced by their true values. As time passes, bets stabilize and approach their oracle quantities.

As alluded to at the end of Section B.1, CSs derived using the capital process $\mathcal{K}_t(m)$ with arbitrary betting schemes are not always guaranteed to produce a convex set (interval). In fact, it is possible to construct scenarios where the sublevel sets of $\mathcal{K}_t^{\mathrm{aGRAPA}}(m)$ are nonconvex in m (see Section E.4 for an example). In our experience, this type of situation is not common, and one must actively search for such pathological examples.

B.4. Lower-bound on the wealth (LBOW)

Instead of maximizing $\log(\mathcal{K}_t(m))$, we may aim to do so for a tight lower-bound on the wealth (LBOW). This technique has proven useful in the game-theoretic probability literature [Shafer and Vovk, 2001, Proof of Lemma 3.3] and [Cutkosky and Orabona, 2018, Proof of Theorem 1]. Our lower bound will rely on an extension of Fan's inequality (40) to $\lambda \in (-1/(1-m), 1/m)$, summarized in the following lemma.

LEMMA 3. *If $y \geq -m$, then for any $\lambda \in [0, 1/m)$, we have*

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{m^2}(\log(1-m\lambda) + m\lambda).$$

On the other hand, if $y \leq 1-m$, then for any $\lambda \in (-1/(1-m), 0]$, we have

$$\log(1+\lambda y) \geq \lambda y + \frac{y^2}{(1-m)^2}(\log(1+(1-m)\lambda) - (1-m)\lambda).$$

Thus, for $y \in [-m, 1-m]$, both of the above inequalities hold.

The proof is an easy generalization of inequality (40) by Fan et al. [2015], and also follows from similar observations about the subexponential function ψ_E in Howard et al. [2020, 2021], but we prove it from first principles in Section A.6 for completeness. Using Lemma 3, we have for $\lambda^{\mathrm{L}+} \in [0, 1/m)$, the following lower-bound on $\ell(\lambda^{\mathrm{L}+}, m)$,

$$\begin{aligned}\ell(\lambda^{\mathrm{L}+}, m) &:= \log\left(\prod_{i=1}^t (1+\lambda^{\mathrm{L}+}(X_i - m))\right) \\ &\geq \lambda^{\mathrm{L}+}\sum_{i=1}^t (X_i - m) + \frac{\log(1-m\lambda^{\mathrm{L}+}) + m\lambda^{\mathrm{L}+}}{m^2}\sum_{i=1}^t (X_i-m)^2, \qquad (45)\end{aligned}$$

and for $\lambda^{\mathrm{L}-} \in (-1/(1-m), 0]$, we have

$$\begin{aligned}\ell(\lambda^{\mathrm{L}-}, m) &:= \log\left(\prod_{i=1}^t (1+\lambda^{\mathrm{L}-}(X_i - m))\right) \\ &\geq \lambda^{\mathrm{L}-}\sum_{i=1}^t (X_i - m) + \frac{\log(1+(1-m)\lambda^{\mathrm{L}-}) - (1-m)\lambda^{\mathrm{L}-}}{(1-m)^2}\sum_{i=1}^t (X_i-m)^2. \qquad (46)\end{aligned}$$

Importantly, if $\sum_{i=1}^t (X_i - m)$ is positive, then (45) is concave, while if negative, (46) is concave. Maximizing (45) or (46) depending on the sign of $\sum_{i=1}^t (X_i - m)$ we obtain the following "hindsight" choice for λ^{L},

$$\lambda^{\mathrm{L}} = \begin{cases} \dfrac{\sum_{i=1}^t (X_i - m)}{m\sum_{i=1}^t (X_i - m) + \sum_{i=1}^t (X_i - m)^2} & \text{if } \displaystyle\sum_{i=1}^t (X_i - m) \geq 0, \\ \dfrac{\sum_{i=1}^t (X_i - m)}{-(1-m)\sum_{i=1}^t (X_i - m) + \sum_{i=1}^t (X_i - m)^2} & \text{if } \displaystyle\sum_{i=1}^t (X_i - m) \leq 0. \end{cases}$$

Of course, this choice of λ^{L} is not predictable and thus is not a valid betting strategy in the framework of the current paper. This motivates the following strategy, $(\lambda_t^{\mathrm{L}}(m))_{t=1}^{\infty}$ given by

$$\lambda_t^{\mathrm{L}}(m) := \frac{-c}{1-m} \vee \frac{\widehat{\mu}_{t-1} - m}{\omega_{t-1}|\widehat{\mu}_{t-1} - m| + \widehat{\sigma}_{t-1}^2 + (\widehat{\mu}_{t-1} - m)^2} \wedge \frac{c}{m}, \tag{47}$$

$$\text{where} \quad \omega_t := \begin{cases} m & \text{if } \widehat{\mu}_t - m \geq 0 \ , \\ 1-m & \text{if } \widehat{\mu}_t - m < 0 \ . \end{cases}$$

Similarly to the aGRAPA betting procedure, LBOW is computationally-inexpensive but is not guaranteed to produce an interval. The expression also carries similar intuition to the GRAPA case.

B.5. Online Newton Step (ONS-m)

Betting algorithms play an essential role in online learning as several optimization problems can be framed in terms of coin-betting games [Cutkosky and Orabona, 2018, Orabona and Tommasi, 2017, Jun et al., 2017, Jun and Orabona, 2019]. While the downstream application is different, the game-theoretic techniques of maximizing wealth are almost immediately applicable to the problem at hand. Here, we consider a slight modification to the Online Newton Step (ONS) algorithm due to Cutkosky and Orabona [2018].

Algorithm 1: ONS-m.

```
Result: $(\lambda_t^{\mathrm{O}}(m))_{t=1}^{T}$
$\lambda_1^{\mathrm{O}}(m) \leftarrow 1$;
for $t \in \{1, \dots, T-1\}$ do
    $y_t \leftarrow X_t - m$ ;
    Set $z_t \leftarrow y_t/(1 - y_t\lambda_t^{\mathrm{O}}(m))$ ;
    $A_t \leftarrow 1 + \sum_{i=1}^{t} z_i^2$ ;
    $\lambda_{t+1}^{\mathrm{O}}(m) \leftarrow \frac{-c}{1-m} \vee \left(\lambda_t^{\mathrm{O}}(m) - \frac{2}{2-\log(3)}\frac{z_t}{A_t}\right) \wedge \frac{c}{m}$ ;
end
```

Through simulations, we find that ONS-m performs competitively. However, its lack of closed-form expression makes it a slightly more computationally-expensive alternative to aGRAPA and LBOW, but not nearly as expensive as GRAPA (see Table 2).

B.6. Diversified Kelly betting (dKelly)

Instead of committing to one betting strategy such as aGRAPA or LBOW, we can simply take the average capital among D separate strategies. This follows from the fact that an average of test martingales is itself a test martingale. That is, if

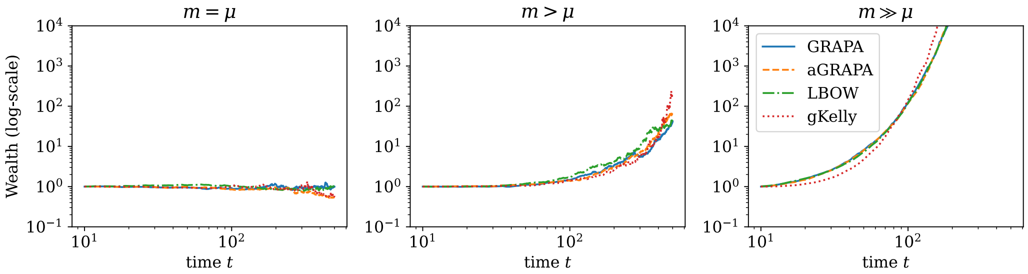

Figure 11. Comparison of the wealth process under various game-theoretic betting strategies with 100 repeats. In this example, the 1000 observations are drawn from a Beta(10, 10) distribution, and the candidate means m being tested are 0.5, 0.51, and 0.55 (from left to right). Notice that these strategies perform similarly, but have varying computational costs (see Table 2).

$(\lambda_t^1)_{t=1}^\infty, (\lambda_t^2)_{t=1}^\infty, \dots, (\lambda_t^D)_{t=1}^\infty$ are D separate betting strategies, then

$$\mathcal{K}_t^{\text{dKelly}}(\mu) := \frac{1}{D}\sum_{d=1}^{D}\prod_{i=1}^{t}\Big(1 + \lambda_i^d(\mu)(X_i - \mu)\Big)$$

forms a test martingale. Following Kelly's original motivation to maximize (expected) log-capital, notice that by Jensen's inequality,

$$\log\left(\mathcal{K}_t^{\text{dKelly}}\right) > \frac{1}{D}\sum_{d=1}^{D}\log\left(\prod_{i=1}^{t}\Big(1 + \lambda_i^d(\mu)(X_i - \mu)\Big)\right).$$

In other words, the log-capital of the diversified bets is strictly larger than the average log-capital among the diverse candidate bets.

Grid Kelly betting (gKelly). While it is possible to use any finite collection of strategies, we focus our attention on a particularly simple (and useful) example where the bets are constant values on a grid. Specifically, divide the interval $[-1/(1-m), 1/m]$ up into G evenly-spaced points $\lambda^1, \dots, \lambda^G$. Then define the gKelly capital process $\mathcal{K}_t^{\text{gKelly}}$ by

$$\mathcal{K}_t^{\text{gKelly}}(m) := \frac{1}{G}\sum_{g=1}^{G}\prod_{i=1}^{t}(1 + \lambda^g(X_i - m)).$$

When used to construct confidence sequences for μ, $\mathcal{K}_t^{\text{gKelly}}$ demonstrates excellent empirical performance. Moreover, this procedure can be slightly modified into "Hedged gKelly" (hgKelly) so that confidence sequences constructed using gKelly are intervals almost surely.

In order to mimic the unknown optimal λ^*, D or G should not be kept constant, but itself grow slowly (say logarithmically) with t. In game-theoretic terms, one should slowly add more strategies to the portfolio, in order to asymptotically match the performance of the optimal one over time. (When adding a new λ^g to an existing mixture, it obviously only begins to contribute to the wealth from the following step onwards; formally G would be replaced by G_t, and $\prod_{i=1}^{t}(1+\lambda^g(X_i-m)$ would be replaced by $\prod_{i=t_g}^{t}(1+\lambda^g(X_i-m)$ if λ^g was first introduced after t_g-1 steps.)

Hedged gKelly. First, divide the interval $[-1/(1-m),0]$ and $[0,1/m]$ into G evenly-spaced points: $(\lambda^{1-},\ldots,\lambda^{G-})$ and $(\lambda^{1+},\ldots,\lambda^{G+})$, respectively. Then define the "Hedged grid Kelly capital process" $\mathcal{K}_t^{\mathrm{hgKelly}}$ given by

$$\mathcal{K}_t^{\mathrm{hgKelly}}(m) := \frac{\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{t}\left(1+\lambda^{g+}(X_i-m)\right) + \frac{1-\theta}{G}\sum_{g=1}^{G}\prod_{i=1}^{t}\left(1+\lambda^{g-}(X_i-m)\right),$$

where $\theta \in [0,1]$ (a reasonable default being $\theta = 1/2$).

Proposition 5. *If $(X_t)_{t=1}^{\infty} \sim P$ for some $P \in \mathcal{P}^\mu$, then $\mathcal{K}_t^{\mathrm{hgKelly}}(\mu)$ forms a test martingale and $\mathfrak{B}_t^{\mathrm{hgKelly}} := \left\{m \in [0,1] : \mathcal{K}_t^{\mathrm{hgKelly}}(m) < 1/\alpha\right\}$ is a CS for μ that forms an interval for each $t \geq 1$.*

The proof in Section A.7 proceeds by showing that $\mathcal{K}_t^{\mathrm{hgKelly}}$ is a convex function of m and hence its sublevel sets are intervals.

B.7. Confidence Boundary (ConBo)

The aforementioned strategies benefit from targeting bets against a particular null hypothesis, H_0^m for each $m \in [0,1]$, but this has the drawback of $\mathcal{K}_t(m)$ potentially not being quasiconvex in m. One of the advantages of the hedged capital process as described in Theorem 3 is that $\mathcal{K}_t^{\pm}(m)$ is always quasiconvex, and thus its sublevel sets (and hence the confidence sets $\mathfrak{B}_t^{\pm}$) are intervals.

In an effort to develop game-theoretic betting strategies which generate confidence sets which are intervals, we present the Confidence Boundary (ConBo) bets. Rather than bet against the null hypotheses H_0^m for each $m \in [0,1]$, consider two sequences of nulls, $(H_0^{u_t})_{t=1}^{\infty}$ and $(H_0^{l_t})_{t=1}^{\infty}$ corresponding to upper and lower confidence boundaries, respectively. The ConBo bet λ_t^{CB} is then targeted against u_{t-1} and l_{t-1} using *any* game-theoretic betting strategy (e.g. ∗GRAPA, ∗Kelly, LBOW, or ONS-m). Letting $\lambda_t^{\mathrm{G}}(m)$ be any such strategy, we summarize the ConBo betting scheme in Algorithm 2.

Corollary 1 (Confidence boundary CS **[ConBo]**). *In Algorithm 2,*

$$\mathfrak{B}_t^{\mathrm{CB}} \quad \textit{forms a } (1-\alpha)\textit{-CS for } \mu,$$

as does $\bigcap_{i\leq t}\mathfrak{B}_i^{\mathrm{CB}}$. Further, $\mathfrak{B}_t^{\mathrm{CB}}$ is an interval for any $t \geq 1$.

We can also adapt the ConBo betting scheme outlined in Algorithm 2 to the without-replacement setting by replacing m by m_t^{WoR} for each time t.

Algorithm 2: ConBo

```
Result: $(\mathcal{K}_t^{\mathrm{CB}}(m))_{t=1}^T$
$l_0 \leftarrow 0$ ; $u_0 \leftarrow 1$;
$\mathcal{K}_0^{\mathrm{CB}+}(m) \leftarrow \mathcal{K}_0^{\mathrm{CB}-}(m) \leftarrow 1$;
for $t \in \{1, \dots, T\}$ do
    $\lambda_t^{\mathrm{CB}+} \leftarrow \max\left\{\lambda_t^{\mathrm{G}}(l_{t-1}), 0\right\} \wedge c/m$;        // Compute ConBo bets
    $\lambda_t^{\mathrm{CB}-} \leftarrow \left|\min\left\{\lambda_t^{\mathrm{G}}(u_{t-1}), 0\right\}\right| \wedge c/(1-m)$;
    $\mathcal{K}_t^{\mathrm{CB}+}(m) \leftarrow \left[1 + \lambda_t^{\mathrm{CB}+}(X_t - m)\right] \cdot \mathcal{K}_{t-1}^{\mathrm{CB}+}(m)$;        // Update capital
    $\mathcal{K}_t^{\mathrm{CB}-}(m) \leftarrow \left[1 - \lambda_t^{\mathrm{CB}-}(X_t - m)\right] \cdot \mathcal{K}_{t-1}^{\mathrm{CB}-}(m)$;
    $\mathcal{K}_t^{\mathrm{CB}}(m) \leftarrow \max\left\{\theta\mathcal{K}_t^{\mathrm{CB}+}(m), (1-\theta)\mathcal{K}_t^{\mathrm{CB}-}(m)\right\}$;        // Hedging
    $\mathfrak{B}_t^{\mathrm{CB}} \leftarrow \{m \in [0,1] : \mathcal{K}_t(m) < 1/\alpha\}$ ;
    $l_t \leftarrow \inf \mathfrak{B}_t^{\mathrm{CB}}$;        // Update confidence boundaries to bet against
    $u_t \leftarrow \sup \mathfrak{B}_t^{\mathrm{CB}}$;
end
```

Corollary 2 (WoR confidence boundary CS **[ConBo-WoR]**). *Under the same conditions as Theorem 4, define $\lambda_t^{\text{CB-WoR}+}$ and $\lambda_t^{\text{CB-WoR}-}$ as in Algorithm 2 but with m replaced by m_t^{WoR}. Then,*

$$\mathfrak{B}_t^{\text{CB-WoR}} := \left\{m \in [0,1] : \mathcal{K}_t^{\text{CB-WoR}} < 1/\alpha\right\} \quad \textit{forms a } (1-\alpha)\textit{-CS for } \mu,$$

as does $\bigcap_{i \le t} \mathfrak{B}_i^{\text{CB-WoR}}$. Further, $\mathfrak{B}_t^{\text{CB-WoR}}$ is an interval for each $t \geq 1$.

B.8. Sequentially Rebalanced Portfolio (SRP)

Implicitly, none of the aforementioned strategies take advantage of "rebalancing", meaning the ability to take ones capital $\mathcal{K}_t$ at time t, diversify it in any manner at time $t+1$, and repeat. This has had the mathematical advantage of being able to write the resulting capital process $(\mathcal{K}_t(m))_{t=1}^\infty$ in the following general, but closed-form expression:

$$\mathcal{K}_t(m) := \sum_{d=1}^{D} \theta_d \prod_{i=1}^{t} (1 + \lambda_i^d(m) \cdot (X_i - m)),$$

where $D \geq 1$ is as in Section B.6, $(\lambda_t^1(m))_{t=1}^\infty, \dots, (\lambda_t^D(m))_{t=1}^\infty$ are $[-1/(1-m), 1/m]$-valued predictable sequences as usual, and $(\theta_d)_{d=1}^D$ are convex weights such that $\sum_{d=1}^D \theta_d = 1$. However, a more general capital process martingale can be written but instead of having a closed-form product expression, it can be written recursively as

$$\mathcal{K}_t^{\mathrm{SRP}}(m) := \sum_{d=1}^{D_t} (1 + \lambda_t^d(m) \cdot (X_t - m)) \cdot \theta_t^d \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m), \tag{48}$$

where $(\lambda_t^d)_{d=1}^{D_t}$ are $[1/(1-m), 1/m]$-valued predictable bets, $(\theta_t^d)_{d=1}^{D_t}$ are predictable convex weights that sum to 1 (conditional on X_1^{t-1}), and we have set the initial capital $\mathcal{K}_0^{\mathrm{SRP}}(m)$ to 1 as usual.

Adopting the betting interpretation, (48) is a rather intuitive procedure. At each time step t, the gambler divides their previous capital $\mathcal{K}_{t-1}^{\mathrm{SRP}}(m)$ up into $D_t \geq 1$ portions given by $\theta_t^1 \cdot K_{t-1}^{\mathrm{SRP}}(m), \dots, \theta_t^{D_t} \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m)$, then invests these wealths with bets $\lambda_t^1(m), \dots, \lambda_t^{D_t}(m)$, respectively. The gambler's wealths are then updated to

$$(1 + \lambda_t^1(m) \cdot (X_t - m)) \cdot \theta_t^1 \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m), \dots, (1 + \lambda_t^{D_t}(m) \cdot (X_t - m)) \cdot \theta_t^{D_t} \cdot \mathcal{K}_{t-1}^{\mathrm{SRP}}(m),$$

which are then combined via summation to yield a final capital of (48).

It is now routine to check that the process given by (48) is a nonnegative martingale when evaluated at μ since

$$\begin{aligned}\mathbb{E}\left(\mathcal{K}_t^{\mathrm{SRP}}(\mu) \mid X_1^{t-1}\right) &= \sum_{d=1}^{D_t} \mathcal{K}_{t-1}^{\mathrm{SRP}}(\mu) \cdot \theta_t^{D_t} \cdot \left(1 + \lambda_t(\mu) \left(\underbrace{\mathbb{E}(X_t \mid X_1^{t-1}) - \mu}_{=0}\right)\right) \\ &= \mathcal{K}_{t-1}^{\mathrm{SRP}}(\mu) \underbrace{\sum_{d=1}^{D_t} \theta_t^{D_t}}_{=1} = \mathcal{K}_{t-1}^{\mathrm{SRP}}(\mu).\end{aligned}$$

Note that SRP is the most general and customizable betting strategy presented in this paper, since it can be composed of any of the previously discussed strategies, and includes each of them as a special case.

C. Simulations

This section contains a comprehensive set of simulations comparing our new confidence sets presented against previous works. We present simulations for building both time-uniform CSs and fixed-time CIs with or without replacement. Each of these are presented under four distributional "themes": (1) discrete, high-variance; (2) discrete, low-variance; (3) real-valued, evenly spread; and (4) real-valued, concentrated.

C.1. Time-uniform confidence sequences (with replacement)

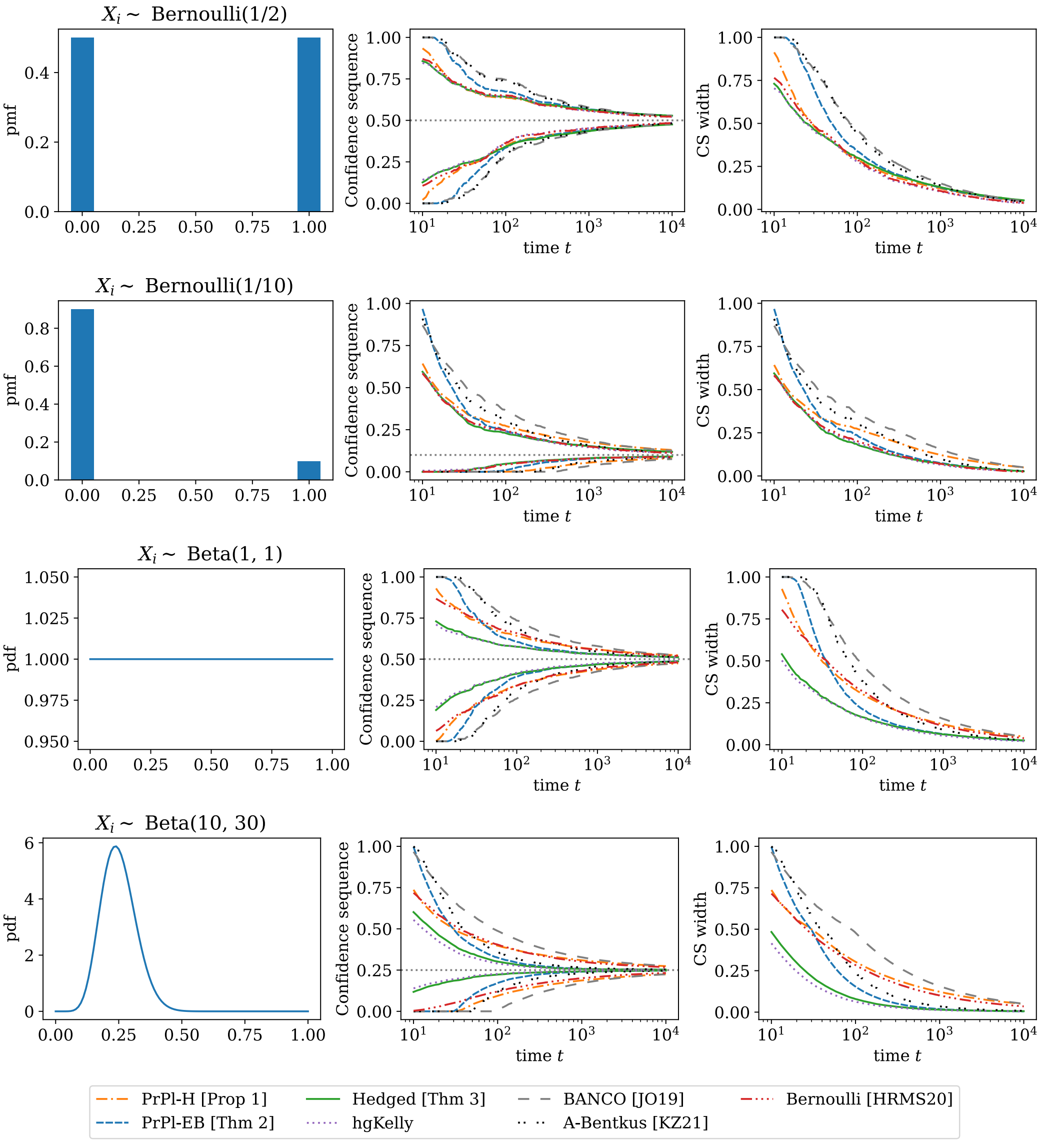

Figure 12. Comparing Hedged, hgKelly, PrPl-EB, and PrPl-H CSs alongside other time-uniform confidence sequences in the literature; further details in Section D.1. Clearly, the betting approach is dominant in all settings.

C.2. Fixed-time confidence intervals (with replacement)

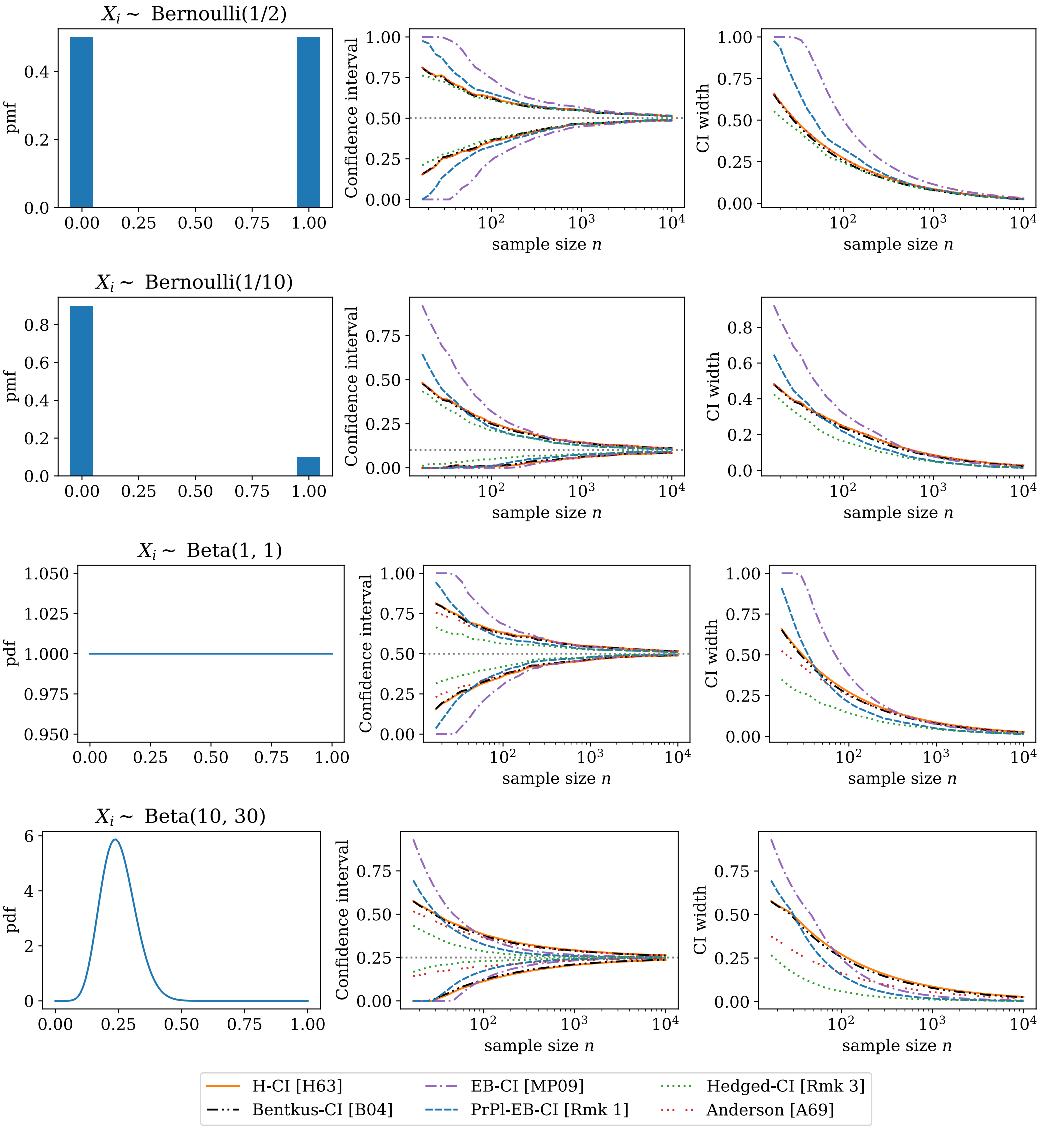

Figure 13. Hedged capital, Anderson, Bentkus, Maurer-Pontil empirical Bernstein, and predictable plug-in empirical Bernstein CIs under four distributional scenarios. Further details can be found in Section D.2. Clearly, the betting approach is dominant in all settings.

C.3. Time-uniform confidence sequences (without replacement)

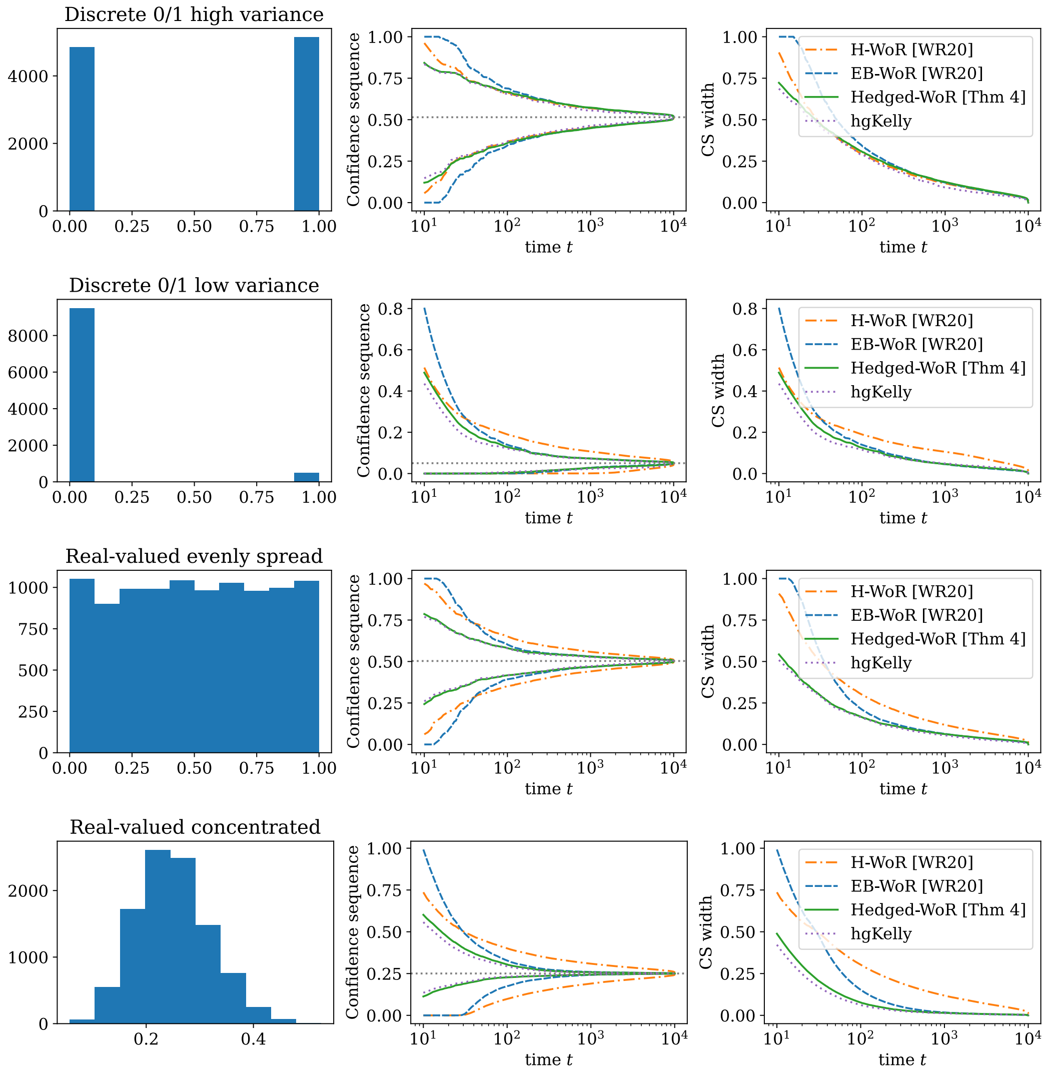

Figure 14. Hedged capital, Hoeffding, and empirical Bernstein CSs for the mean of a finite set of bounded numbers when sampling WoR. Further details can be found in Section D.3. Clearly, the betting approach is dominant in all settings.

C.4. Fixed-time confidence intervals (without replacement)

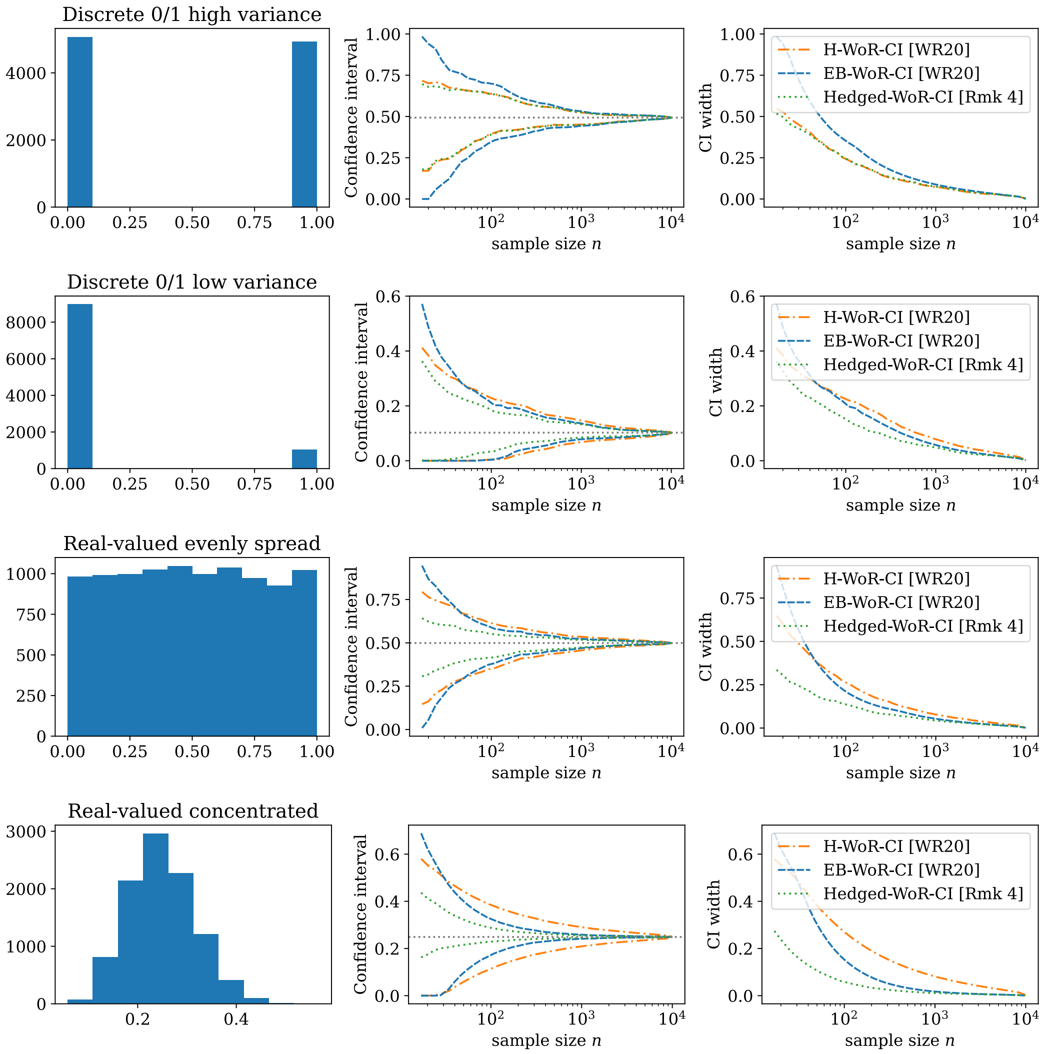

Figure 15. Fixed-time hedged capital, Hoeffding-type, and empirical Bernstein-type CIs for the mean of a finite set of bounded numbers when sampling WoR. Further details can be found in Section D.4. Clearly, the two betting approaches (Hedged and ConBo) are dominant in all settings.

Table 2. Typical computation time for constructing a CS from time 1 to 10^3 for the mean of Bernoulli$(1/2)$-distributed random variables. The three betting CSs were computed for 1000 evenly-spaced values of m in $[0, 1]$, while a coarser grid would have sped up computation. All CSs were calculated on a laptop powered by a quad-core 2GHz 10th generation Intel Core i5. Parallelization was carried out using the Python library, `multiprocess` [McKerns et al., 2011].

Betting scheme	Interval (a.s.)	Computation time (seconds)
ConBo+LBOW	✓	0.08
Hedged+$(\lambda_t^{\mathrm{PrPl}\pm})_{t=1}^{\infty}$	✓	0.25
hgKelly $(G = 20)$	✓	1.38
aGRAPA		0.35
LBOW		0.25
ONS-m		12.45
Kelly		197.38

D. Simulation details

In each simulation containing confidence sequences or intervals and their widths, we took an average over 5 random draws from the relevant distribution. For example, in the "Time-uniform confidence sequences" plot of Figure 1, the CSs (PrPl-H, PrPl-EB, and Hedged) were averaged over 5 random draws from a Beta$(10, 30)$ distribution. Computation times for various strategies are given in Table 2.

D.1. Time-uniform confidence sequences (with replacement)

Each of the CSs considered in the time-uniform (with replacement) case are presented as explicit theorems and propositions throughout the paper. Specifically,

- **PrPl-H**: Predictable plug-in Hoeffding (Proposition 1);
- **PrPl-EB**: Predictable plug-in empirical Bernstein (Theorem 2);
- **Hedged**: Hedged capital process (Theorem 3); and
- **hgKelly**: Hedged grid-Kelly (Proposition 5).

Bernoulli [HRMS20] Section C compared these against the conjugate mixture sub-Bernoulli confidence sequence by Howard et al. [2021], recalled below.

Hoeffding [1963, Equation (3.4)], presented the sub-Bernoulli upper-bound on the moment generating function of bounded random variables for any $\lambda > 0$:

$$\mathbb{E}_P \left(\exp \left\{ \lambda(X_i - \mu) \right\} \right) \leq 1 - \mu + \mu \exp\{\lambda\},$$

which can be used to construct an e-value by noting that

$$\mathbb{E}_P \left(\exp \left\{ \lambda(X_i - \mu) - \log(1 - \mu + \mu e^{\lambda}) \right\} \mid \mathcal{F}_{i-1} \right) \leq 1.$$

Then, Howard et al. [2021] showed that the cumulative product process

$$\prod_{i=1}^{t} \left(\exp\left\{ \lambda(X_i - \mu) - \log(1 - \mu + \mu e^{\lambda}) \right\} \right) \tag{49}$$

forms a test supermartingale, as does a mixture of (49) for any probability distribution $F(\lambda)$ on $\mathbb{R}^+$:

$$\int_{\lambda \in \mathbb{R}^+} \prod_{i=1}^{t} \left(\exp\left\{ \lambda X_i - \log(1 - \mu + \mu e^{\lambda}) \right\} \right) dF(\lambda). \tag{50}$$

In particular, Howard et al. [2021] take $F(\lambda)$ to be a beta distribution so that the integral (50) can be computed in closed-form. Using (50) in Step (b) in Theorem 1 yields the "Bernoulli [HRMS20]" confidence sequence.

There are yet other improvements of Hoeffding's inequality, for example one that goes by the name of Kearns-Saul [Kearns and Saul, 1998] but was incidentally noted in Hoeffding's original paper itself. This inequality, and other variants, are looser than the sub-Bernoulli bound and so we exclude them here; see Howard et al. [2020] for more details. Most importantly, none of these adapt to the true underlying variance of the random variables, unlike most of our new techniques.

A-Bentkus [KZ21] We also compared our bounds against the "adaptive Bentkus confidence sequence" (A-Bentkus) due to Kuchibhotla and Zheng [2021, Section 3.5]. These combine a maximal version of Bentkus et al.'s concentration inequality [Kuchibhotla and Zheng, 2021, Theorem 1] with the "stitching" technique Zhao et al. [2016], Mnih et al. [2008], Howard et al. [2021] — a method to obtain infinite-horizon concentration inequalities by taking a union bound over exponentially-spaced *finite* time horizons.

D.2. Fixed-time confidence intervals (with replacement)

For the fixed-time CIs included from this paper, we have

- **PrPl-EB-CI**: Predictable plug-in empirical Bernstein CI (Remark 1); and
- **Hedged-CI**: Hedged capital process CI (Remark 3).

These were compared against CIs due to Hoeffding [1963], Maurer and Pontil [2009], Anderson [1969], and Bentkus [2004] which we now recall.

H-CI [H63] These intervals refer to the CIs based on Hoeffding's classical concentration inequalities [Hoeffding, 1963]. Specifically, for a sample size $n \geq 1$, "H-CI [H63]" refers to the CI,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \pm \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

Anderson [A69] These intervals refer to the confidence intervals due to Anderson [1969] which take a unique approach by considering the entire sample cumulative distribution function, rather than just the mean and variance. Consequently, however, Anderson's CIs require iid observations, rather than the more general setup we consider. We nevertheless find that even in the iid setting, our approach outperforms Anderson's.

Suppose $X_1, \dots, X_n \overset{iid}{\sim} P$ are $[0,1]$-bounded with mean $\mathbb{E}_P(X_1) = \mu$. Let $X_{(1)}, \dots, X_{(n)}$ denote the order statistics of X_1^n with the convention that $X_{(0)} := 0$ and $X_{(n+1)} := 1$. Following the notation of Learned-Miller and Thomas [2019], Anderson's CI is given by

$$\left[\sum_{i=1}^n u_i^{\mathrm{DKW}} \left(-X_{(n-(i+1))} + X_{(n-i)} \right),\ 1 - \sum_{i=1}^n u_i^{\mathrm{DKW}} \left(X_{(i+1)} - X_{(i)} \right) \right],$$

where $u_i^{\mathrm{DKW}} = \left(i/n - \sqrt{\log(2/\alpha)/2n} \right) \vee 0$. Learned-Miller and Thomas [2019, Theorem 2] show that Anderson's CI is always tighter than Hoeffding's. The authors also introduce a bound which is strictly tighter than Anderson's which they conjecture has valid $(1-\alpha)$-coverage, but we do not compare to this bound here.

EB-CI [MP09] The empirical Bernstein CI of Maurer and Pontil [2009] is given by

$$\frac{1}{n} \sum_{i=1}^n X_i \pm \sqrt{\frac{2\widehat{\sigma}^2 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)}{3(n-1)},$$

and $\widehat{\sigma}^2$ is the sample variance.

Bentkus-CI [B04] Bentkus' confidence interval requires an a-priori upper bound on $\mathrm{Var}(X_i)$ for each i. As alluded to in the introduction, we do not consider concentration bounds which require knowledge of the variance. However, since we assume $X_i \in [0,1]$, we have the trivial upper bound, $\mathrm{Var}(X_i) \leq \frac{1}{4}$, which we implicitly use throughout our computation of Bentkus' confidence interval.

Define the independent, mean-zero random variables $(G_i)_{i=1}^n$ as

$$G_i := \begin{cases} -\frac{1}{4} & \text{w.p. } \frac{4}{5} \\ 1 & \text{w.p. } \frac{1}{5} \end{cases},$$

an important technical device which has appeared in seminal works by Hoeffding [1963, Equation (2.14)] and Bennett [1962, Equation (10)]. Then the "Bentkus-CI" is

$$\frac{1}{n} \sum_{i=1}^n X_i \pm \frac{W_\alpha^\star}{n},$$

where $W_\alpha^\star \in [0, n]$ is given by the value of W_α such that

$$\inf_{y \in [0,n]\ :\ y \leq W_\alpha} \frac{\mathbb{E}\left[\sum_{i=1}^n (G_i - y)_+^2 \right]}{(W_\alpha - y)_+^2} = \alpha.$$

Efficient algorithms have been developed to solve the above [Bentkus et al., 2006, Section 9], [Kuchibhotla and Zheng, 2021].

PTL-ℓ_2 [PTL21] The work by Phan et al. [2021] proposes an interesting but computationally intensive approach to constructing confidence intervals for means of iid bounded random variables. Specifically, we will focus on their tightest bound (according to [Phan et al., 2021, Figure 4]) which makes use of the ℓ_2 norm in its derivation (and which we thus refer to as PTL-ℓ_2).

For example, computing PTL-ℓ_2 confidence intervals[¶] from a sample $X_1, \dots, X_{300} \sim \mathrm{Unif}[0,1]$ of $n = 300$ uniformly distributed random variables took upwards of 11 minutes while our betting confidence interval (Remark 3) took less than 0.5 seconds. For this reason, we conduct a small-scale simulation of sample sizes 5-200 (see Figure 16). We find that PTL-ℓ_2 performs extremely well for the low-variance continuous distribution Beta(10, 30) but poorly for sample sizes closer to 200 for Bernoulli data. Nevertheless, PTL-ℓ_2 requires i.i.d. data (while we only require boundedness and conditional mean μ) and PTL-ℓ_2 does not have time-uniform or without-replacement analogues.

D.3. Time-uniform confidence sequences (without replacement)

The WoR CSs which were introduced in this paper include

- **Hedged-WoR**: Without replacement hedged capital process (Theorem 4); and
- **hgKelly-WoR**: Without replacement analogue of hgKelly (Proposition 5).

The CSs labeled "H-WoR [WR20]" and "EB-WoR [WR20]" are the without-replacement Hoeffding- and empirical Bernstein-type CSs due to Waudby-Smith and Ramdas [2020] which we recall now.

H-WoR [WR20] Define the weighted WoR mean estimator and the Hoeffding-type λ-sequence,

$$\widehat{\mu}_t^{\mathrm{WoR}}(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i (X_i + \frac{1}{N-i+1}\sum_{j=1}^{i-1} X_j)}{\sum_{i=1}^t \lambda_i (1 + \frac{i-1}{N-i+1})}, \quad \text{and} \quad \lambda_t := \sqrt{\frac{8\log(2/\alpha)}{t\log(t+1)}} \wedge 1,$$

respectively. Then "H-CS [WR20]" refers to the WoR Hoeffding-type CS,

$$\widehat{\mu}_t^{\mathrm{WoR}}(\lambda_1^t) \pm \frac{\sum_{i=1}^t \psi_H(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}.$$

[¶] We used code by Phan et al. [2021] with their default tuning parameters, available at github.com/myphan9/small_sample_mean_bounds.

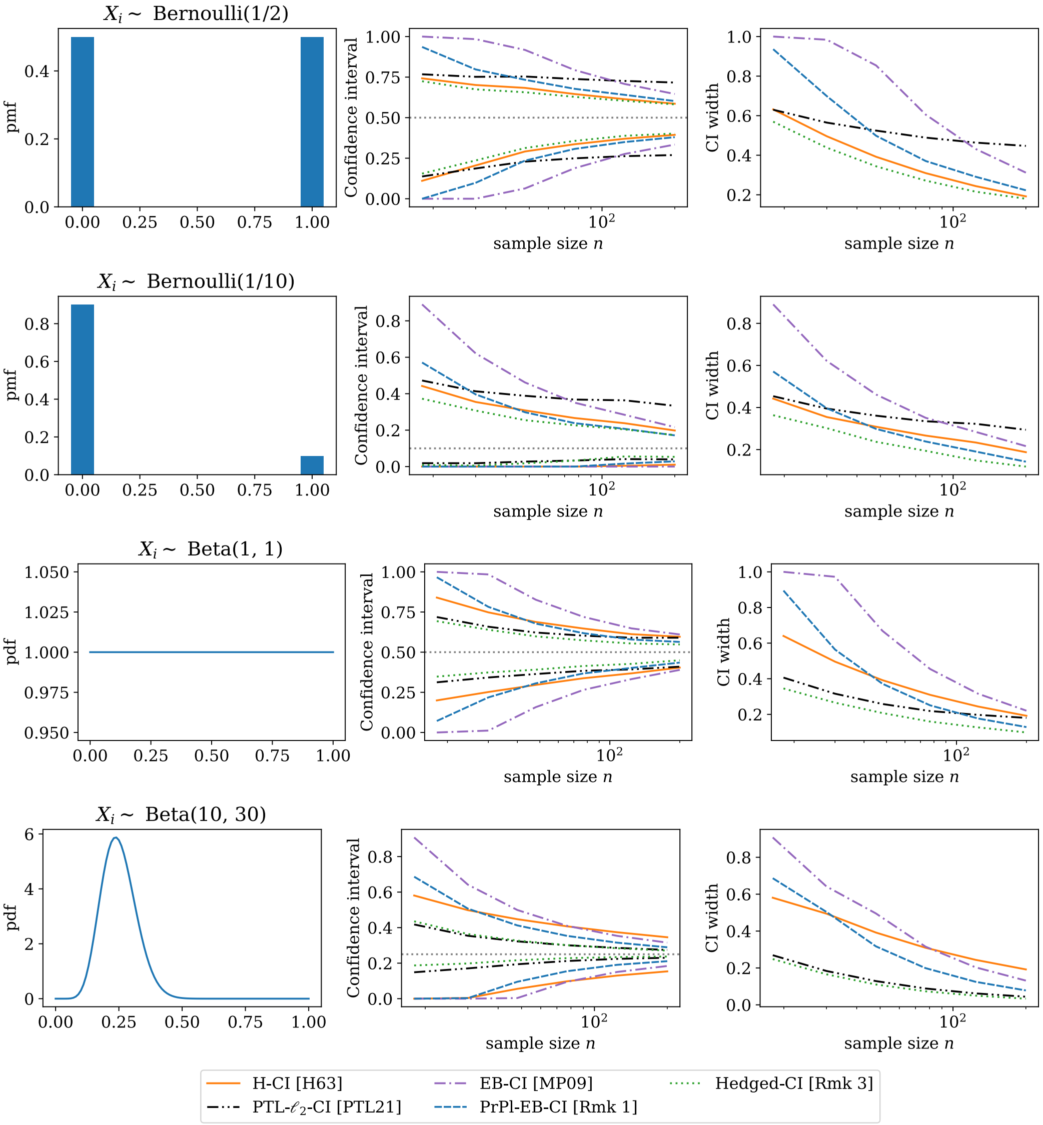

Figure 16. Various with-replacement fixed-time confidence intervals, including that of Phan et al. [2021] (PTL-ℓ_2-CI). While PTL-ℓ_2-CI performs very well in the Beta(10, 30) regime, it appears to suffer for Bernoulli(1/2) with larger n. In any case, PTL-ℓ_2-CI relies on iid data, while the other four methods do not.

EB-WoR [WR20] Analogously to the Hoeffding-type CSs, "EB-CS [WR20]" corresponds to the empirical Bernstein-type CSs for sampling WoR due to Waudby-Smith and Ramdas [2020]. These CSs take the form

$$\widehat{\mu}_t^{\mathrm{WoR}}(\lambda_1^t) \pm \frac{\sum_{i=1}^t 4(X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)},$$

where in this case, we have

$$\lambda_t := \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_{t-1}^2 t \log(t+1)}} \wedge \frac{1}{2}, \quad \widehat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \widehat{\mu}_t := \frac{1}{t}\sum_{i=1}^t X_i. \tag{51}$$

D.4. Fixed-time confidence intervals (without replacement)

The only fixed-time CI introduced in this paper is **Hedged-WoR-CI**: the without-replacement hedged capital process CI described in Section 5. The other two are both due to Waudby-Smith and Ramdas [2020] which we describe now.

H-WoR-CI [WR20] This corresponds to the CI described in Corollary 3.1 of Waudby-Smith and Ramdas [2020]. This has the form

$$\widehat{\mu}_n^{\mathrm{WoR}} \pm \frac{\sqrt{\frac{1}{2}\log(2/\alpha)}}{\sqrt{n} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{i-1}{N-i+1}}.$$

EB-WoR-CI [WR20] Similarly, this CI corresponds to that described in Corollary 3.2 of Waudby-Smith and Ramdas [2020]. Specifically, "EB-WoR-CI [WR20]" is defined as

$$\widehat{\mu}_n^{\mathrm{WoR}}(\lambda_1^n) \pm \frac{\sum_{i=1}^n 4(X_i - \widehat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^n \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)},$$

where

$$\lambda_t := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}} \wedge \frac{1}{2}, \quad \widehat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \text{ and } \widehat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1}, \tag{52}$$

and $\widehat{\mu}_n^{\mathrm{WoR}}$ is defined as

$$\widehat{\mu}_t^{\mathrm{WoR}}(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i (X_i + \frac{1}{N-i+1}\sum_{j=1}^{i-1} X_j)}{\sum_{i=1}^t \lambda_i (1 + \frac{i-1}{N-i+1})}.$$

D.5. Betting "confidence distributions": confidence sets at several resolutions

Figures 17 and 18 demonstrate two tools to visualize CSs at various α and t.

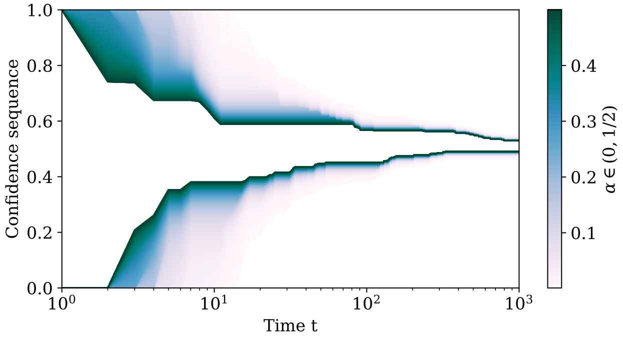

Figure 17. This plot shows the aGRAPA CS for all $\alpha \in [0, 1/2]$ under Unif$[0, 1]$ data.

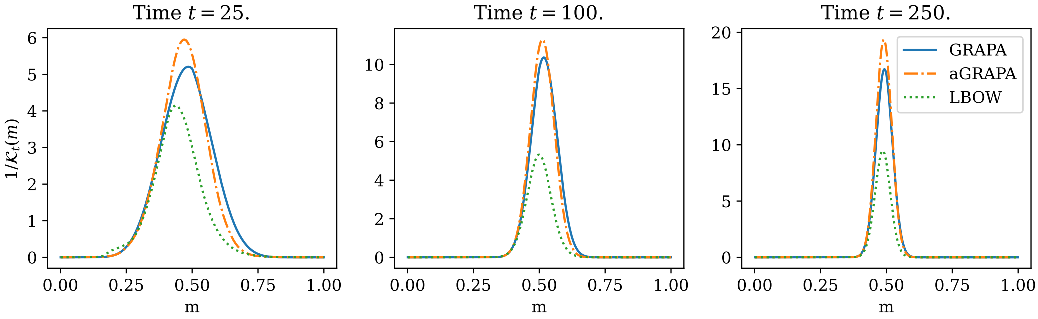

Figure 18. Here we plot the inverse wealth $1/\mathcal{K}_t(m)$ in game m against $m \in [0, 1]$, at $t = 25, 100, 250$ for three different betting strategies. Note the different y-axis scales. Despite not being normalized to yield a "confidence distribution", this is a useful visual tool. For example, the mode in each plot signifies the m against which we have minimum wealth, which is a reasonable point estimator for μ. Further, the superlevel set for any $\alpha \in [0, 1]$ yields exactly the $(1 - \alpha)$-CS for μ (for that corresponding time and strategy) since it yields all m with wealth less than $1/\alpha$. Last, for any $m \in [0, 1]$, the height (truncated at one) is anytime-valid p-value for the null hypothesis that the mean equals m.

E. Additional theoretical results

E.1. Betting confidence sets are tighter than Hoeffding

In this section, we demonstrate that the betting approach can dominate Hoeffding for sufficiently large sample sizes. First, we show that for any $x, m \in (0, 1)$ and any $\lambda \in \mathbb{R}$, then $\gamma \equiv \gamma^m(\lambda)$ can be set as

$$\gamma^m(\lambda) := \exp\left\{-m\lambda - \lambda^2/8\right\}(\exp(\lambda) - 1),$$

so that

$$H^m(x) := \underbrace{\exp\left\{\lambda(x - m) - \lambda^2/8\right\}}_{\text{Hoeffding term}} \leq \underbrace{1 + \gamma(x - m)}_{\text{Capital process term}} =: \mathcal{K}^m(x)$$

for any $x, m \in [0,1]$. In particular, the Hoeffding-type and capital process supermartingales are built from precisely the above terms, respectively, and so if $H^m(x) \le \mathcal{K}^m(x)$ for any $x \in [0,1]$, then their respective supermartingales will satisfy the same inequality almost surely.

Proposition 6 (Capital process dominates Hoeffding process). *Suppose $x, m \in [0,1]$ and $\lambda \in \mathbb{R}$. Then there exists $\gamma^m(\lambda) \in \mathbb{R}$ such that*

$$H^m(x) := \exp\left(\lambda(x-m) - \lambda^2/8\right) \le 1 + \gamma^m(\lambda)(x-m) =: \mathcal{K}^m(x).$$

Note that Proposition 6 alone does not confirm that the Hoeffding-based CIs will be dominated by capital process-based CIs since γ must be within $[-1/(1-m), 1/m]$ for $\mathcal{K}^m(x)$ to be nonnegative. However, it is easy to verify that for all $\lambda \in [-0.45, 0.45]$, we have that $\gamma \in [-1,1]$ and thus $\mathcal{K}^m(x) \ge 0$. When constructing a Hoeffding-type $(1-\alpha)$-confidence interval, for example, one would set $\lambda_n^H := \sqrt{8\log(2/\alpha)/n}$, making $\lambda_n^H \in [-0.45, 0.45]$ whenever $n \ge 40\log(2/\alpha)$, in which case a capital process-based CI will dominate a Hoeffding-based CI almost surely.

Proof (Proposition 6). We prove the result for $\lambda \ge 0$ and remark that this implies the result for the case when $\lambda \le 0$ by considering $(1-x)$ and $(1-m)$ instead of x and m, respectively.

The proof proceeds in 3 steps. First, we consider the line segment $L^m(x)$ connecting $H^m(0)$ and $H^m(1)$ and note that by convexity of $H^m(x)$, we have that $H^m(x) \le L^m(x)$ for all $x \in [0,1]$. We then find the slope of this line segment and set γ to this value so that the line $\mathcal{K}^m(x) := 1 + \gamma(x-m)$ has the same slope as $L^m(x)$. Finally, we demonstrate that $L^m(0) \le \mathcal{K}^m(0)$, and conclude that $H^m(x) \le L^m(x) \le \mathcal{K}^m(x)$ for all $x \in [0,1]$.

Step 1. Note that $H^m(x)$ is a convex function in $x \in [0,1]$, and thus

$$\forall x \in [0,1],\ H^m(x) \le H^m(0) + [H^m(1) - H^m(0)]\, x =: L^m(x).$$

Step 2. Observe that the slope of $L^m(x)$ is $H^m(1) - H^m(0)$. Setting $\gamma := H^m(1) - H^m(0)$ we have that $\mathcal{K}^m(x)$ and $L^m(x)$ are parallel.

Step 3. It remains to show that $\mathcal{K}^m(0) \ge L^m(0) \equiv H^m(0)$ for every $m \in [0,1]$. Consider the following equivalent statements:

$$\begin{aligned}
&\mathcal{K}^m(0) \ge H^m(0)\\
\iff &1 - m\,[H^m(1) - H^m(0)] \ge H^m(0)\\
\iff &1 - m\exp\left(\lambda - \lambda m - \lambda^2/8\right) \ge (1-m)\exp\left(-\lambda m - \lambda^2/8\right)\\
\iff &1 \ge \exp\left(-\lambda m - \lambda^2/8\right)[1 - m + m\exp(\lambda)]\\
\iff &\exp\left(\lambda m + \lambda^2/8\right) \ge [1 - m + m\exp(\lambda)]\\
\iff &a(\lambda) := \exp\left(\lambda m + \lambda^2/8\right) - [1 - m + m\exp(\lambda)] \ge 0.
\end{aligned}$$

Now, note that a is smooth and $a(0) = 0$ and so it suffices to show that its derivative $a'(\lambda) \geq 0$ for all $\lambda \geq 0$. To this end, consider the following equivalent statements.

$$\begin{aligned}
&a'(\lambda) \equiv \left(m + \frac{\lambda}{4}\right) \exp\left(\lambda m + \lambda^2/8\right) - m \exp(\lambda) \geq 0 \\
\iff &\left(m + \frac{\lambda}{4}\right) \exp\left(\lambda m + \lambda^2/8\right) \geq m \exp(\lambda) \\
\iff &\ln\left(1 + \frac{\lambda}{4m}\right) + \lambda m + \lambda^2/8 \geq \lambda \\
\iff &b(\lambda) := \ln\left(1 + \frac{\lambda}{4m}\right) + \lambda m + \lambda^2/8 - \lambda \geq 0,
\end{aligned}$$

and hence it suffices to show that $b(\lambda) \geq 0$. Similar to $a(\lambda)$, we have that $b(0) = 0$ and so it suffices to show that its derivative, $b'(\lambda) \geq 0$ for all $\lambda \geq 0$. Indeed,

$$\begin{aligned}
&b'(\lambda) \equiv \frac{1}{4m + \lambda} + m + \frac{\lambda}{4} - 1 \geq 0 \\
\iff &c(\lambda) := 1 + m(4m + \lambda) + \frac{\lambda}{4}(4m + \lambda) - 4m - \lambda \geq 0
\end{aligned}$$

Since $c(\lambda)$ is a convex quadratic, it is straightforward to check that

$$\operatorname*{argmin}_{\lambda \in \mathbb{R}} c(\lambda) = 2 - 4m,$$

and that $c(2 - 4m) = 0$. In conclusion, if we set $\gamma \equiv \gamma^m(\lambda)$ as

$$\gamma^m(\lambda) := H^m(1) - H^m(0) = \exp\left\{-m\lambda - \lambda^2/8\right\}(\exp(\lambda) - 1),$$

then $H^m(x) \leq \mathcal{K}^m(x) := 1 + \gamma^m(\lambda)(x - m)$ for every $m \in [0, 1]$. This completes the proof. □

E.2. Optimal convergence of betting confidence sets

In Section B, it was mentioned that for nonnegative martingales, Ville's inequality is nearly an equality and hence martingale-based CSs are nearly tight in a time-uniform sense. However, it is natural to wonder what other theoretical guarantees betting CSs/CIs can have in addition to their empirical performance. In the time-uniform setting, CSs for the mean cannot attain widths which scale faster than $\asymp \sqrt{\log\log t/t}$, due to the law of the iterated logarithm. Similarly, fixed-time CIs cannot scale faster than $\asymp 1/\sqrt{n}$. In this section, we show that it is possible to choose betting strategies such that the resulting CSs and CIs scale at the optimal rates of $O(\sqrt{\log\log t/t})$ and $O(1/\sqrt{n})$, respectively.

E.2.1. An iterated logarithm betting confidence sequence

We will establish the law of the iterated logarithm (LIL) convergence rate by carefully constructing a capital process martingale whose resulting CS is — for sufficiently large t — tighter than a larger CS which itself attains the required LIL rate.

Before stating the result in Proposition 7, let $\zeta(s) := \sum_{k=1}^{\infty} \frac{1}{k^s}$ be the Riemann zeta function and for each $k \in \{1, 2, \dots\}$, define

$$\lambda_k := \sqrt{\frac{8 \log\left(k^s \zeta(s)\right)}{\eta^{k+1/2}}}, \quad \text{and}$$
$$\gamma_k(m) = \exp\left\{-m\lambda_k - \lambda_k^2/8\right\}(\exp(\lambda_k) - 1) \wedge 1,$$

where $\eta > 1$ is some user-chosen constant. Let k_t denote the (unique) integer such that $\log_\eta t \le k_t \le \log_\eta t + 1$. Define the process

$$\mathcal{K}_t^{\mathcal{L}} := \frac{1}{2}\mathcal{K}_t^{\mathcal{L}+}(m) + \frac{1}{2}\mathcal{K}_t^{\mathcal{L}-}(m)$$
$$\text{where} \quad \mathcal{K}_t^{\mathcal{L}+}(m) := \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^{t} (1 + \gamma_{k_t}(X_i - m)) \quad \text{and}$$
$$\mathcal{K}_t^{\mathcal{L}-}(m) := \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^{t} (1 - \gamma_{k_t}(X_i - m)).$$

Note that $\mathcal{K}_t^{\mathcal{L}+}(m)$ and $\mathcal{K}_t^{\mathcal{L}-}(m)$ are both upper-bounded by the infinite mixtures

$$\mathcal{K}_t^{\mathcal{L}+}(m) \le \sum_{k=1}^{\infty} \frac{1}{k^s \zeta(s)} \prod_{i=1}^{t} (1 + \gamma_k(X_i - m)) \quad \text{and} \tag{53}$$
$$\mathcal{K}_t^{\mathcal{L}-}(m) \le \sum_{k=1}^{\infty} \frac{1}{k^s \zeta(s)} \prod_{i=1}^{t} (1 - \gamma_k(X_i - m)), \tag{54}$$

which themselves form nonnegative martingales when $m = \mu$ by Fubini's theorem. Consequently,

$$C_t^{\mathcal{L}} := \left\{ m \in [0,1] : \mathcal{K}_t^{\mathcal{L}}(m) < \frac{1}{\alpha} \right\}$$

forms a $(1-\alpha)$-CS for μ. The following proposition establishes the LIL rate of $C_t^{\mathcal{L}}$.

Proposition 7. *The CS $(C_t^{\mathcal{L}})_{t=1}^{\infty}$ has a width of $O(\sqrt{\log\log t/t})$, meaning*

$$\nu(C_t^{\mathcal{L}}) = O\left(\sqrt{\frac{\log\log t}{t}}\right),$$

where ν is the Lebesgue measure.

Proof. The proof proceeds in three steps. In Step 1, we construct a distinct but related CS (which we will denote by $(C_t^{\times})_{t=1}^{\infty}$) via the stitching technique [Howard et al., 2021]. In Step 2, we demonstrate that this stitched CS achieves the desired rate by deriving an analytically tractible superset whose width scales as $O(\sqrt{\log\log t/t})$. Finally, in Step 3, we will show that the stitched CS $C_t^{\times}$ is a superset of $C_t^{\mathcal{L}}$ for all t sufficiently large, thus implying the final result.

Step 1. Constructing the stitched CS $C_t^\times$: In the language of betting, the idea behind stitching is to first divide one's capital up into infinitely many portions $w_1, w_2, \dots$ such that $\sum_{k=1}^\infty w_k = 1$, and then place a constant bet λ_k using a capital of w_k on a designated epoch of time, which will be chosen to be geometrically spaced. In what follows, the portions w_k will be given by $w_k = \frac{1}{\zeta(s)k^s}$, and we will divide time $\{1, 2, 3, \dots\}$ up into epochs demarcated by the endpoints η^{k-1} and η^k for each $k \in \{1, 2, 3, \dots\}$ and for some user-specified $\eta > 1$ (e.g. $\eta = 1.1$). The constant bets λ_k will be chosen so that they are effective between η^{k-1} and η^k and lead to $O(\sqrt{\log\log t/t})$ widths after being combined across epochs.

The construction of the stitched boundary essentially follows (a simplified version of) the proof of Theorem 1 in Howard et al. [2021, Section A.1], but we present the derivation here for completeness. Consider the Hoeffding-type process for a fixed $\lambda \in \mathbb{R}$:

$$M_t^\lambda(m) := \exp\left\{\lambda S_t(m) - t\lambda^2/8\right\}, \tag{55}$$

where $S_t(m) := \sum_{i=1}^t (X_i - m)$. As discussed in Section 3, $M_t(\mu)$ forms a test supermartingale, and hence by Ville's inequality we have

$$P\left(\exists t \geq 1 : S_t(\mu) \geq \underbrace{\frac{r + t\lambda^2/8}{\lambda}}_{g_{\lambda,r}(t)}\right) \leq e^{-r}.$$

We have typically used $r = \log(1/\alpha)$ throughout the paper, but the above alternative notation will help in the following discussion. Using the notation of Howard et al. [2021, Section A.1], define the boundary above as $g_{\lambda,r}(t) := (r + t\lambda^2/8)/\lambda$, and let

$$\lambda_k := \sqrt{\frac{8r_k}{\eta^{k-1/2}}},$$

$$\text{where} \quad r_k := \log\left(\frac{k^s\zeta(s)}{\alpha/2}\right).$$

Some algebra will reveal that plugging the above choices of λ_k and r_k into $g_{\lambda,r}(t)$ yields

$$g_{\lambda_k,r_k}(t) := \sqrt{\frac{r_k t}{8}}\left(\sqrt{\frac{\eta^{k-1/2}}{t}} + \sqrt{\frac{t}{\eta^{k-1/2}}}\right),$$

resulting in the following concentration inequality for each k:

$$P\left(\exists t \geq 1 : S_t(\mu) \geq g_{\lambda_k,r_k}(t)\right) \leq \exp\{-r_k\}.$$

Let k_t denote the (unique) epoch number such that $\eta^{k_t-1} \leq t \leq \eta^{k_t}$ (i.e. such that $\log_\eta t \leq k_t \leq \log_\eta t + 1$). Now, we take a union bound over $k = 1, 2, 3, \dots$ resulting in the following boundary,

$$P\left(\exists t \geq 1 : S_t(\mu) \geq g_{\lambda_{k_t},r_{k_t}}(t)\right) \leq \sum_{k=1}^\infty \exp\{-r_k\} = \frac{\alpha/2}{\zeta(s)}\underbrace{\sum_{k=1}^\infty \frac{1}{k^s}}_{\zeta(s)} = \alpha/2.$$

Repeating all of the previous steps for $-S(\mu)$ and taking a union bound, we arrive at the $(1-\alpha)$ stitched CS $(C_t^\times)_{t=1}^\infty$ given by

$$C_t^\times := \left(\frac{1}{t} \sum_{i=1}^t X_i \pm \frac{g_{\lambda_{k_t}, r_{k_t}}(t)}{t} \right),$$

with the guarantee that $P(\exists t \geq 1 : \mu \notin C_t^\times) \leq \alpha$.

Step 2. Demonstrating that $C_t^\times$ achieves the desired LIL width: Now, we will simply upper-bound $g_{\lambda_{k_t}, r_{k_t}}(t)$ by an analytical boundary depending explicitly on t (rather than implicitly through k_t) to see that it achieves the desired LIL width. First, notice that $\sqrt{\eta^{k_t - 1/2}/t} + \sqrt{t/\eta^{k_t - 1/2}}$ is uniquely minimized when $t = \eta^{k_t - 1/2}$ and hence its maximum on the interval $(\eta^{k_t - 1}, \eta^{k_t})$ must be at the endpoints. Therefore, $\sqrt{\eta^{k_t - 1/2}/t} + \sqrt{t/\eta^{k_t - 1/2}} \leq \eta^{1/4} + \eta^{-1/4}$ and thus for each k, we have

$$g_{\lambda_{k_t}, r_{k_t}}(t) \leq \sqrt{\frac{r_{k_t} t}{8}} \left(\eta^{1/4} + \eta^{-1/4} \right) \quad \text{for all } \eta^{k_t - 1} \leq t \leq \eta^{k_t}.$$

Furthermore, for all $\eta^{k_t - 1} \leq t \leq \eta^{k_t}$, we have that $k_t \leq \log_\eta t + 1$. Applying this inequality to the above, we obtain the final bound which does not depend on k,

$$g_{\lambda_{k_t}, r_{k_t}}(t) \leq \sqrt{\frac{t \log \left(2 \left(\log_\eta t + 1 \right)^s \zeta(s)/\alpha \right)}{8}} \left(\eta^{1/4} + \eta^{-1/4} \right) \quad \text{for all } k.$$

In conclusion, we have that

$$C_t^\times \subseteq \left(\frac{1}{t} \sum_{i=1}^t X_i \pm \sqrt{\frac{\log \left(2 \left(\log_\eta t + 1 \right)^s \zeta(s)/\alpha \right)}{8t}} \left(\eta^{1/4} + \eta^{-1/4} \right) \right),$$

and thus $C_t^\times = O\left(\sqrt{\log \log t / t} \right)$, as desired.

Step 3. Showing that $C_t^{\mathcal{L}} \subseteq C_t^\times$ for all t large enough: This step in the proof essentially follows immediately from the discussion in Section E.1. We justified that for $\lambda \geq 0$, setting γ as

$$\gamma = \exp\left\{ -m\lambda - \lambda^2/8 \right\} (\exp(\lambda) - 1) \wedge 1,$$

yields $1 + \gamma(x - m) \geq \exp\left\{ \lambda(x - m) - \lambda^2/8 \right\}$ for all $x, m \in [0, 1]$ if λ is sufficiently small (i.e. so that γ is not relying on truncation at 1). Since λ_k is decreasing in t, it follows that for t sufficiently large,

$$\prod_{i=1}^t (1 + \gamma_{k_t}(X_i - m)) \geq \exp\left\{ \lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8 \right\} \quad \text{almost surely.}$$

Therefore, for t sufficiently large,

$$\begin{aligned}\mathcal{K}_t^{\mathcal{L}+}(m) &:= \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^t (1 + \gamma_{k_t}(X_i - m)) \\ &\geq \frac{1}{k_t^s \zeta(s)} \exp\left\{\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\right\} =: H_t^{\infty+}(m)\end{aligned}$$

and similarly for $K_t^{\mathcal{L}-}(m)$,

$$\mathcal{K}_t^{\mathcal{L}-}(m) \geq \frac{1}{k_t^s \zeta(s)} \exp\left\{-\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\right\} =: H_t^{\infty-}(m).$$

Therefore, for sufficiently large t, we have

$$\begin{aligned}C_t^{\mathcal{L}} &:= \left\{ m \in [0,1] : \mathcal{K}_t^{\mathcal{L}}(m) < \frac{1}{\alpha} \right\} \\ &\subseteq \underbrace{\left\{ m \in \mathbb{R} : \max\left\{ \frac{1}{2} H_t^{\infty+}(m),\ \frac{1}{2} H_t^{\infty-}(m) \right\} < \frac{1}{\alpha} \right\}}_{(\star)}\end{aligned}$$

and it is straightforward to verify that $(\star)$ is precisely $C_t^{\times}$.

In summary, we constructed a CS $C_t^{\times}$ using the stitching technique in Step 1, and then showed that $\nu(C_t^{\times}) = O(\sqrt{\log\log t/t})$ in Step 2. Finally in Step 3, we showed that our discrete mixture betting CS $C_t^{\mathcal{L}}$ is a subset of $C_t^{\times}$ for t sufficiently large, and hence by subadditivity of measures,

$$\nu(C_t^{\mathcal{L}}) = O\left(\sqrt{\frac{\log\log t}{t}} \right),$$

which completes the proof. □

Remark 5. *Notice that $\mathcal{K}_t^{\mathcal{L}+}$ and $\mathcal{K}_t^{\mathcal{L}-}$ can be made strictly more powerful if they are replaced by adding additional terms, as long as the final sums are upper-bounded by (53) and (54), respectively. In particular, any finite sum analogue of (53) and (54) would have sufficed, as long as $\mathcal{K}_t^{\mathcal{L}+}$ and $\mathcal{K}_t^{\mathcal{L}-}$ form a term in each sum, respectively. We presented $\mathcal{K}_t^{\mathcal{L}+}$ and $\mathcal{K}_t^{\mathcal{L}-}$ in their current forms for the sake of notational (and computational) simplicity.*

E.2.2. The $\sqrt{n}$-convergence of betting CIs

Proposition 8. *Suppose $X_1^n \sim P$ are independent observations from a distribution $P \in \mathcal{P}^\mu$ with mean $\mu \in [0,1]$. Let $\lambda_n \in (0,1)$ such that $\lambda_n \asymp 1/\sqrt{n}$. Then the confidence interval,*

$$C_n := \left\{ m \in [0,1] : \mathcal{K}_n^{\pm} < \frac{1}{\alpha} \right\} \quad \textit{has an asymptotic width of } O(1/\sqrt{n}).$$

PROOF. Writing out the capital process with positive bets, we have by Lemma 3 that for any $m \in [0,1]$,

$$\begin{aligned}
\mathcal{K}_n^+(m) &:= \prod_{i=1}^n (1 + \lambda_n (X_i - m)) \\
&\geq \exp\left(\lambda_n \sum_{i=1}^n (X_i - m) - \psi_E(\lambda_n) \sum_{i=1}^n 4(X_i - m)^2 \right) \\
&\geq \exp\left(\lambda_n \sum_{i=1}^n (X_i - m) - 4n\psi_E(\lambda_n) \right) =: B_t^+(m),
\end{aligned}$$

and similarly for negative bets,

$$\begin{aligned}
\mathcal{K}_n^-(m) &:= \prod_{i=1}^n (1 - \lambda_n (X_i - m)) \\
&\geq \exp\left(-\lambda_n \sum_{i=1}^t (X_i - m) - 4n\psi_E(\lambda_n) \right) =: B_t^-(m).
\end{aligned}$$

For any $\theta \in (0,1)$, consider the set,

$$\mathcal{S}_n := \left\{ m : B_t^+(m) < \frac{1}{\theta\alpha} \right\} \bigcap \left\{ m : B_t^-(m) < \frac{1}{(1-\theta)\alpha} \right\}$$

Now notice that the $1/\alpha$-level set of $\mathcal{K}_n^\pm(m) := \max\{\theta\mathcal{K}_n^+(m), (1-\theta)\mathcal{K}_n^-(m)\}$ is a subset of $\mathcal{S}_n$:

$$C_n = \left\{ m : \mathcal{K}_n^+(m) < \frac{1}{\theta\alpha} \right\} \bigcap \left\{ m : \mathcal{K}_n^-(m) < \frac{1}{(1-\theta)\alpha} \right\} \subseteq \mathcal{S}_n.$$

On the other hand, it is straightforward to derive a closed-form expression for $\mathcal{S}_n$:

$$\left(\frac{\sum_{i=1}^n X_i}{n} - \frac{\log\left(\frac{1}{\theta\alpha}\right) + 4n\psi_E(\lambda_n)}{n\lambda_n}, \frac{\sum_{i=1}^n X_i}{n} + \frac{\log\left(\frac{1}{(1-\theta)\alpha}\right) + 4n\psi_E(\lambda_n)}{n\lambda_n} \right),$$

which in the typical case of $\theta = 1/2$ has the cleaner expression,

$$\frac{\sum_{i=1}^n X_i}{n} \pm \frac{\log(2/\alpha) + 4n\psi_E(\lambda_n)}{n\lambda_n}.$$

As discussed in Section B, we have by two applications of L'Hôpital's rule that $\frac{\psi_E(\lambda_n)}{\psi_H(\lambda_n)} \xrightarrow{n\to\infty} 1$, where $\psi_H(\lambda_n) := \lambda_n^2/8 \asymp 1/n$ and thus the width W_n of $\mathcal{S}_n$ scales as

$$W_n := 2 \cdot \frac{\log(1/\alpha) + 4n\psi_E(\lambda_n)}{n\lambda_n} \asymp \frac{\log(1/\alpha)}{\sqrt{n}} + \frac{4n/n}{\sqrt{n}} \asymp \frac{1}{\sqrt{n}}.$$

Since $C_n \subseteq \mathcal{S}_n$, we have that C_n has a width of $O(1/\sqrt{n})$, which completes the proof. □

Despite these results, the hedged capital CI presented and recommended in Section 4.4 does not satisfy the assumptions of the above proof. In particular, we recommended using the variance-adaptive predictable plug-in,

$$\lambda_t^{\text{PrPl-EB}(n)} := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}}, \quad \widehat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \widehat{\mu}_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1}, \tag{56}$$

using a truncation which depends on m,

$$\lambda_t^+(m) := \lambda_t^\pm \wedge \frac{c}{m}, \quad \lambda_t^-(m) := -\left(\lambda_t^\pm \wedge \frac{c}{1-m}\right), \tag{57}$$

and finally defining the hedged capital process for each $t \in \{1, \dots, n\}$:

$$\mathcal{K}_t^\pm(m) := \max\left\{\theta \prod_{i=1}^t (1 + \lambda_i^+(m) \cdot (X_i - m)), (1-\theta) \prod_{i=1}^t (1 - \lambda_i^-(m) \cdot (X_i - m))\right\}.$$

Furthermore, the resulting CI is defined as an intersection,

$$\mathfrak{B}_n := \bigcap_{t=1}^n \left\{m \in [0,1] : \mathcal{K}_t^\pm(m) < \frac{1}{\alpha}\right\}. \tag{58}$$

All of these tweaks (i.e. making bets predictable, truncating beyond $(0,1)$, and taking an intersection) do not in any way invalidate the type-I error, but we find (through simulations) that they tighten the CIs, especially in low-variance, asymmetric settings (see Figure 19).

E.3. On the width of empirical Bernstein confidence intervals

Recall the predictable plug-in empirical Bernstein confidence interval:

$$C_n^{\text{PrPl-EB}(n)} := \left(\frac{\sum_{i=1}^n \lambda_i X_i}{\sum_{i=1}^n \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}\right),$$

where

$$\lambda_t := \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{t-1}^2}}, \; \widehat{\sigma}_t^2 := \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}, \text{and } \widehat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1}.$$

Below, we analyze the asymptotic behavior of the width of $C_n^{\text{PrPl-EB}(n)}$ in the i.i.d. setting. In Proposition 9, we will show that if the data are drawn i.i.d. from a distribution $Q \in \mathcal{Q}^\mu$ having variance σ^2, then the half-width W_n of $C_n^{\text{PrPl-EB}(n)}$ scales as

$$\sqrt{n}W_n \equiv \sqrt{n}\left(\frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}\right) \xrightarrow{a.s.} \sigma\sqrt{2\log(2/\alpha)}, \tag{59}$$

and hence the width is asymptotically proportional to the standard deviation.

First, let us prove a few lemmas about *nonrandom* sequences of numbers, which will be helpful in what follows. These are simple facts for which we could not find a proof to reference, so we prove them below for completeness.

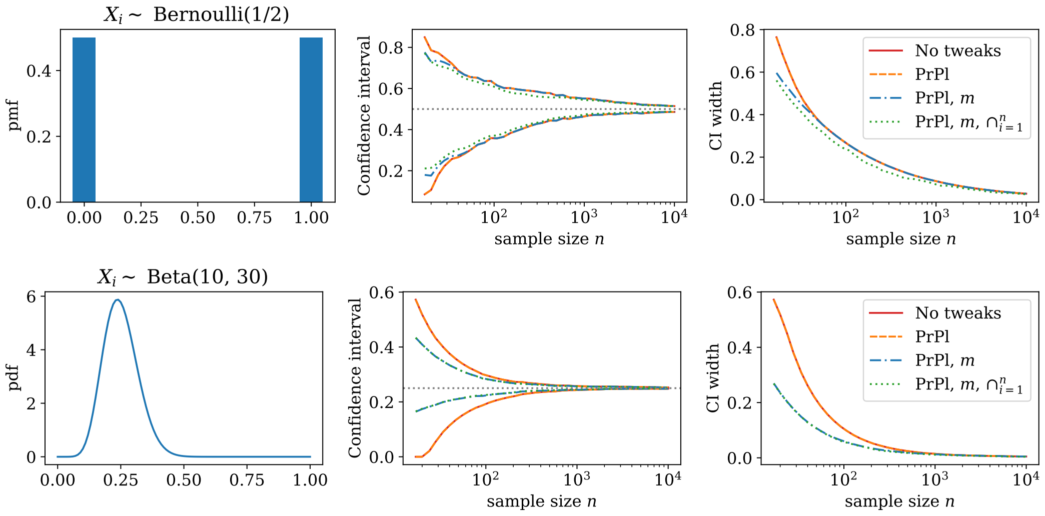

Figure 19. Hedged capital CIs with various added tweaks. The CIs labeled “No tweaks” refer to those which satisfy the conditions of Proposition 8. The other three plots differ in which “tweaks” have been added. Those with “PrPl” in the legend use the predictable plug-in approach defined in (56); those with m in the legend have been truncated using m as outlined in (57); finally, the plots with $\cap_{i=1}^n$ in their legends had their running intersections taken as in (58).

LEMMA 4. *Suppose $(a_n)_{n=1}^\infty$ is a sequence of real numbers such that $a_n \to a$. Then their cumulative average also converges to a, meaning that $\frac{1}{n}\sum_{i=1}^n a_i \to a$.*

PROOF. Let $\epsilon > 0$ and choose $N \equiv N_\epsilon \in \mathbb{N}$ such that whenever $n \geq N$, we have

$$|a_n - a| < \epsilon. \tag{60}$$

Moreover, choose

$$M \equiv M_N > \frac{\sum_{i=1}^N |a_i - a|}{\epsilon} \tag{61}$$

and note that

$$\frac{n - N - 1}{n} < 1. \tag{62}$$

Let $n \geq \max\{N, M\}$. Then we have by the triangle inequality,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n (a_i - a)\right| &\leq \frac{1}{n}\sum_{i=1}^N |a_i - a| + \frac{1}{n}\sum_{i=N+1}^n |a_i - a| \\
&\leq \frac{1}{n}\sum_{i=1}^N |a_i - a| + \frac{1}{n}(n - N - 1)\epsilon && \text{by (60)} \\
&\leq 2\epsilon && \text{by (61) and (62)},
\end{aligned}$$

which can be made arbitrarily small. This completes the proof of Lemma 4. □

LEMMA 5. *Let $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ be sequences of numbers such that*

$$a_n \to 0 \quad \textit{and} \tag{63}$$
$$|b_n| \leq C \quad \textit{for some } C \geq 0 \textit{ and for all } n \geq 1. \tag{64}$$

Then $a_n b_n \to 0$. Further, if (A_n) is a sequence of random variables such that $A_n \to 0$ almost surely, then $A_n b_n \to 0$ almost surely.

The proof is trivial, since $|A_n b_n| \leq C|A_n|$ which converges to zero almost surely. □

Now, we prove that a modified variance estimator is consistent.

LEMMA 6. *Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} Q \in \mathcal{Q}^\mu$ with $\mathrm{Var}(X_i) = \sigma^2$. Then the modified variance estimator*

$$\widehat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2$$

converges to σ^2, Q-almost surely.

PROOF. By direct substitution,

$$\begin{aligned}\widehat{\sigma}_n^2 &:= \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu + \mu - \widehat{\mu}_{i-1})^2 \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2}_{\xrightarrow{a.s.} \sigma^2} - \underbrace{\frac{2}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})(\widehat{\mu}_{i-1} - \mu)}_{(\star)} + \underbrace{\frac{1}{n} \sum_{i=1}^n (\mu - \widehat{\mu}_{i-1})^2}_{(\star\star)}.\end{aligned}$$

Now, note that $\widehat{\mu}_{i-1} - \mu \xrightarrow{a.s.} 0$ and $|X_i - \widehat{\mu}_{i-1}| \leq 1$ for each i. Therefore, by Lemma 5, $(X_i - \widehat{\mu}_{i-1})(\widehat{\mu}_{i-1} - \mu) \xrightarrow{a.s.} 0$, and by Lemma 4, $(\star) \xrightarrow{a.s.} 0$. Furthermore, we have that $(\mu - \widehat{\mu}_{i-1})^2 \xrightarrow{a.s.} 0$ and so by another application of Lemma 4, we have $(\star\star) \xrightarrow{a.s.} 0$. This completes the proof of Lemma 6. □

Next, let us analyze the second term in the numerator in the margin of $C_n^{\text{PrPl-EB}(n)}$,

$$\frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}. \tag{65}$$

LEMMA 7. *Under the same assumptions as Lemma 6,*

$$\sum_{i=1}^n v_i \psi_E(\lambda_i) \xrightarrow{a.s.} \log(2/\alpha).$$

Proof. Recall that $\frac{\psi_E(\lambda)}{\psi_H(\lambda)} \xrightarrow{\lambda \to 0} 1$, and $\widehat{\sigma}_t^2 \xrightarrow{t \to \infty} \sigma^2$. By definition of λ_i, we have that $\lambda_i \xrightarrow{a.s.} 0$ and thus we may also write

$$\frac{\psi_E(\lambda_i)}{\psi_H(\lambda_i)} = 1 + R_i \quad \text{and} \tag{66}$$

$$\sqrt{\frac{\sigma^2}{\widehat{\sigma}_t^2}} = 1 + R_i' \tag{67}$$

for some $R_i, R_i' \xrightarrow{a.s.} 0$. Thus, we rewrite the left hand side of the claim as

$$\begin{aligned}
\sum_{i=1}^n v_i \psi_E(\lambda_i) &= \sum_{i=1}^n v_i \psi_H(\lambda_i) \frac{\psi_E(\lambda_i)}{\psi_H(\lambda_i)} = \sum_{i=1}^n v_i (\lambda_i^2/8)(1 + R_i) \\
&= \sum_{i=1}^n v_i \cdot \frac{2\log(2/\alpha)}{8 \widehat{n\sigma}_{i-1}^2} \cdot (1 + R_i) \\
&= \sum_{i=1}^n v_i \cdot \frac{2\log(2/\alpha)}{8 n \sigma^2} \cdot (1 + R_i') \cdot (1 + R_i) \\
&= \sum_{i=1}^n 4(X_i - \widehat{\mu}_{i-1})^2 \cdot \frac{2\log(2/\alpha)}{8 n \sigma^2} \cdot (1 + R_i + R_i' + R_i R_i').
\end{aligned}$$

Defining $R_i'' = R_i + R_i' + R_i R_i'$ for brevity, and noting that $R_i'' \to 0$ almost surely, the above expression becomes

$$\begin{aligned}
\sum_{i=1}^n v_i \psi_E(\lambda_i) &= \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 \cdot \frac{\log(2/\alpha)}{n \sigma^2} \cdot (1 + R_i'') \\
&= \frac{\log(2/\alpha)}{\sigma^2} \left[\frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 \cdot (1 + R_i'') \right] \\
&= \frac{\log(2/\alpha)}{\sigma^2} \left[\underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2}_{\xrightarrow{a.s.} \sigma^2 \text{ by Lemma 6}} + \underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_{i-1})^2 R_i''}_{\xrightarrow{a.s.} 0 \text{ by Lemma 5}} \right] \xrightarrow{a.s.} \log(2/\alpha),
\end{aligned}$$

which completes the proof of Lemma 7. □

Now, consider the denominator in (65).

Lemma 8. *Continuing with the same notation,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i \xrightarrow{a.s.} \sqrt{\frac{2\log(2/\alpha)}{\sigma^2}}.$$

PROOF. Let R_i' be as in (67). Then,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{\frac{2\log(2/\alpha)}{n\widehat{\sigma}_{i-1}^2}} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{\frac{2\log(2/\alpha)}{n\sigma^2}} \cdot (1 + R_i') \\
&= \sqrt{\frac{2\log(2/\alpha)}{\sigma^2}} \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n (1 + R_i')}_{\xrightarrow{a.s.} 1 \text{ by Lemma 4}} \xrightarrow{a.s.} \sqrt{\frac{2\log(2/\alpha)}{\sigma^2}},
\end{aligned}$$

completing the proof of Lemma 8. □

We are now able to combine Lemmas 7 and 8 to prove the main result.

PROPOSITION 9. *Denoting the half-width of $C_n^{\text{PrPl-EB}(n)}$ as W_n, and assuming the data are drawn iid from a distribution $Q \in \mathcal{Q}^\mu$ with variance σ^2, we have*

$$\sqrt{n}W_n \;\equiv\; \sqrt{n}\left(\frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}\right) \xrightarrow{a.s.} \sigma\sqrt{2\log(2/\alpha)}. \tag{68}$$

Thus, the width is asymptotically proportional to the standard deviation.

PROOF. By direct rearrangement of the left hand side, we see that

$$\begin{aligned}
\sqrt{n}\left(\frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}\right) &= \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\frac{1}{\sqrt{n}}\sum_{i=1}^n \lambda_i} \\
&\xrightarrow{a.s.} \frac{\log(2/\alpha) + \log(2/\alpha)}{\sigma^{-1}\sqrt{2\log(2/\alpha)}} = \sigma\sqrt{2\log(2/\alpha)},
\end{aligned}$$

which completes the proof of Proposition 9. □

E.4. aGRAPA sublevel sets need not be intervals: a worst-case example

In the proof of Theorem 3, we demonstrated that the hedged capital process with predictable plug-in bets yielded convex confidence sets, making their construction more practical. However, this proof was made simple by taking advantage of the fact that the sequences before truncation $(\dot{\lambda}_t^+)_{t=1}^\infty$ and $(\dot{\lambda}_t^-)_{t=1}^\infty$ did not depend on $m \in [0,1]$. This raises the natural question, of whether there are betting-based confidence sets which are nonconvex when these sequences depend on m. Here, we provide a (somewhat pathological) example of the aGRAPA process with nonconvex sublevel sets.

Consider the aGRAPA bets,

$$\lambda_t^{\text{aGRAPA}} := \frac{\widehat{\mu}_{t-1} - m}{\widehat{\sigma}_{t-1}^2 + (\widehat{\mu}_{t-1} - m)^2} \text{ where } \widehat{\mu}_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1},\ \widehat{\sigma}_t^2 := \frac{1/20 + \sum_{i=1}^t (X_i - \widehat{\mu}_i)^2}{t+1}. \tag{69}$$

Furthermore, suppose that the observed variables are $X_1 = X_2 = 0$. Then it can be verified that

$$\begin{aligned}\mathcal{K}_2^{\text{aGRAPA}}(m) &= \left(1 + \lambda_1^{\text{aGRAPA}}(X_1 - m)\right)\left(1 + \lambda_2^{\text{aGRAPA}}(X_2 - m)\right) \\ &= \left(1 + \frac{1/2 - m}{1/20 + (1/2 - m)^2}(-m)\right)\left(1 + \frac{1/4 - m}{0.05625 + (1/4 - m)^2}(-m)\right),\end{aligned}$$

which does not yield convex sublevel sets. For example, $\mathcal{K}_2^{\text{aGRAPA}}(0.08) < 0.85$ and $\mathcal{K}_2^{\text{aGRAPA}}(0.4) < 0.85$ but $\mathcal{K}_2^{\text{aGRAPA}}(0.03) > 0.85$. In particular, the sublevel set,

$$\left\{m \in [0, 1] : \mathcal{K}_2^{\text{aGRAPA}}(m) < 0.85\right\}$$

is not convex. In our experience, however, situations like the above do not arise frequently. In fact, we needed to actively search for these examples and use a rather small "prior" variance of 1/20 which we would not use in practice. Furthermore, the sublevel set given above is at the 0.85 level while confidence sets are compared against $1/\alpha$ which is always larger than 1 and typically larger than 10. We believe that it may be possible to restrict $(\lambda_t^{\text{aGRAPA}})_{t=1}^{\infty}$ and/or the confidence level, $\alpha \in (0, 1)$ in some way so that the resulting confidence sets are convex. One reason to suspect that this may be possible is because of the intimate relationship between $\lambda_t^{\text{aGRAPA}}$, λ_t^{GRAPA}, and the optimal hindsight bets, λ^{HS}. Specifically, we show in Section E.6 that the optimal hindsight capital $\mathcal{K}_t^{\text{HS}}$ is exactly the empirical likelihood ratio [Owen, 2001] which is known to generate convex confidence sets for the mean [Hall and La Scala, 1990]. We leave this question as a direction for future work.

E.5. Betting confidence sequences for non-iid data

The CSs presented in this paper are valid under the assumption that each observation is bounded in $[0, 1]$ with conditional mean μ. That is, we require that $X_1, X_2, \ldots$ are $[0, 1]$-valued with $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = \mu$ for each t, which includes familiar regimes such as independent and identically-distributed (iid) data from some common distribution P with mean μ. Despite the generality of our results, we made matters simpler by focusing the simulations in Section C on the iid setting. For the sake of completeness, we present a simulation to examine the behavior of our CSs in the presence of some non-iid data.

In this setup, we draw the first several hundred or thousand observations independently from a Beta(10, 10) — a distribution whose mean is 1/2 but whose variance is small (≈ 0.012) — while the remaining observations are independently drawn from a Bernoulli(1/2) whose mean is also 1/2 but with a maximal variance of 1/4. We chose to start the data off with low-variance observations in an attempt to "trick" our betting strategies into adapting to the wrong variance. Empirically, we find that the hedged capital (Theorem 3) and ConBo (Corollary 1) CSs start off strong, adapting to the small variance of a Beta(10, 10). After several Bernoulli(1/2) observations, the CSs remain tight, but seem to shrink less rapidly. Nevertheless, we find that the hedged capital and ConBo CSs greatly outperform the Hoeffding (Proposition 1) and empirical Bernstein (Theorem 2) predictable plug-in CSs (see Figure 20). Regardless of empirical performance, all methods considered produce *valid* CSs for μ.

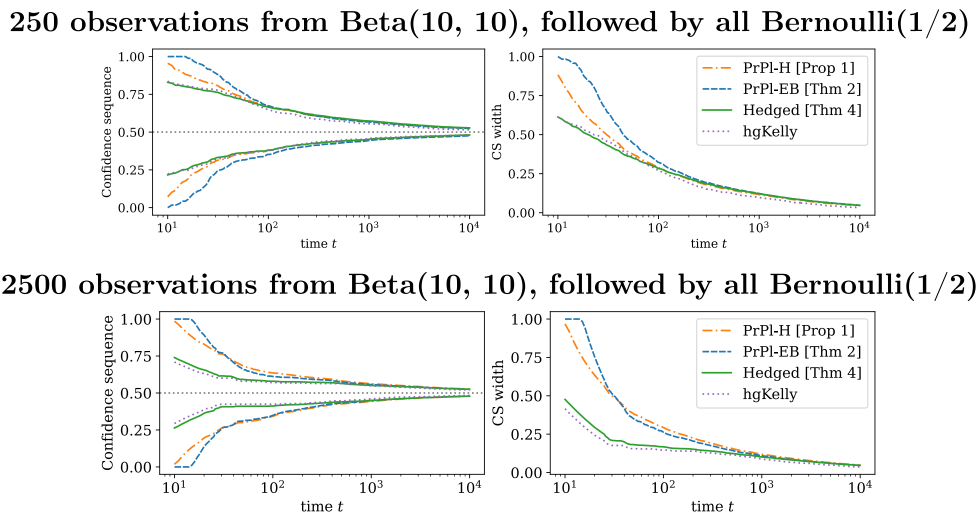

Figure 20. CSs for the true mean $\mu = 1/2$ for non-iid data. In top pair of plots, the first 250 observations were independently drawn from a Beta(10, 10) while the subsequent observations are drawn from a Bernoulli(1/2). The bottom pair of plots is similar, but with 2500 initial draws from a Beta(10, 10) instead of 250. In both cases, the betting-based CSs (Hedged and ConBo) tend to outperform those based on supermartingales.

E.6. Owen's empirical likelihood ratio and Mykland's dual likelihood ratio

Let $x_1, \dots, x_t \in [0, 1]$ and recall the optimal hindsight capital process $\mathcal{K}_t^{\mathrm{HS}}(m)$,

$$\mathcal{K}_t^{\mathrm{HS}}(m) := \prod_{i=1}^{t} (1 + \lambda^{\mathrm{HS}}(x_i - m)) \quad \text{where } \lambda^{\mathrm{HS}} \text{ solves } \quad \sum_{i=1}^{t} \frac{x_i - m}{1 + \lambda^{\mathrm{HS}}(x_i - m)} = 0.$$

Now, let $\mathcal{Q}^m \equiv \mathcal{Q}^m(x_1^t)$ be the collection of discrete probability measures with support $\{x_1, \dots, x_t\}$ and mean m. Let $\mathcal{Q} \equiv \mathcal{Q}(x_1^t) := \bigcup_{m \in [0,1]} \mathcal{Q}^m$ and define the empirical likelihood ratio [Owen, 2001],

$$\mathrm{EL}_t(m) := \frac{\sup_{Q \in \mathcal{Q}} \prod_{i=1}^{t} Q(x_i)}{\sup_{Q \in \mathcal{Q}^m} \prod_{i=1}^{t} Q(x_i)}.$$

Owen [2001] showed that the numerator equals $(1/t)^t$ and the denominator equals

$$\prod_{i=1}^{t} (1 + \lambda^{\mathrm{EL}}(x_i - m))^{-1} \quad \text{where } \lambda^{\mathrm{EL}} \text{ solves } \quad \sum_{i=1}^{t} \frac{x_i - m}{1 + \lambda^{\mathrm{EL}}(x_i - m)} = 0.$$

Notice that the above product is exactly the reciprocal of $\mathcal{K}_t^{\mathrm{HS}}$ and that $\lambda^{\mathrm{EL}} = \lambda^{\mathrm{HS}}$. Therefore for each $m \in [0, 1]$,

$$\mathrm{EL}_t(m) = (1/t)^t \mathcal{K}_t^{\mathrm{HS}}(m).$$

Furthermore, given the connection between the empirical and dual likelihood ratios for independent data [Mykland, 1995], the hindsight capital process is also proportional to the dual likelihood ratio in this case.

F. An extended history of betting and its applications

(This is an expanded version of Section 6 and Figure 9.)

The use of betting-related ideas in probability, statistics, optimization, finance and machine learning has evolved in many different parallel threads, emanating from different influential early works and thus having different roots and evolutions. Since these threads have had little interaction for many decades now, we consider it worthwhile to mention them in some detail. Two notes of caution:

- We anticipate missing some authors and works in our broad strokes below, but a thorough coverage would be better suited to a longer survey paper on the topic. For example, we entirely skip the field of mathematical finance, since betting is literally a foundation of the entire field (and theoretical and applied progress on martingales, betting strategies, and related topics has been phenomenal).
- Many of the authors listed below have used the language of betting in their works explicitly, but others have not — and may even prefer (or have preferred) *not* to do so. Thus, our references should be treated with a pinch of salt, as some connections that we draw to betting may be more apparent in hindsight (to us) than foresight (to the authors).

If we had to pick the most critical early authors without whom our work would have been impossible, it would be Ville, Wald, Kelly and Robbins; later influences on us have been via Lai, Cover, Shafer, Vovk, Grunwald and the second author's own earlier works [Howard et al., 2020, 2021]. These authors stand out below.

Probability. Ville's 1939 PhD thesis [Ville, 1939] contained an important and rather remarkable result of its time that connected measure-theoretic probability with betting, and indeed brought the very notion of a martingale into probability theory. In brief, Ville proved that for every event of measure zero, there exists a betting strategy for which a gambler's wealth process (a nonnegative martingale) grows to infinity if that event occurs. For example, the strong law of large numbers (SLLN) and the law of the iterated logarithm (LIL) are two classic measure-theoretic statements that occur on all sequences of observations, except for a null set according to some underlying probability measure (where the two null sets for the two laws are different). Ville proved that it is possible to bet on the next outcome such that if the LIL were false for that particular sequence of observations, then the gambler's wealth would grow in an unbounded fashion.

Doob's monumental papers and book Doob [1953] in the following decades stripped martingales of their betting roots and presented them as some of the most powerful tools of measure-theoretic probability theory, with applications to many other branches of mathematics. (However, betting could be viewed as instances of "Doob's martingale transform".) These betting roots were revived in the 1960s with the renewed interest in algorithmic definitions of randomness, due to Kolmogorov, Martin-Löf [1966] and many others.

More recently, Shafer and Vovk [2001, 2019] have produced two seminal books that aim bring betting and martingales to the front and center of probability and finance, aiming to derive much (if not all) of probability theory from purely game-theoretic principles based on betting strategies. The product martingale wealth process that appears in our work also appears in theirs (indeed, it is a fundamental process), but Shafer and Vovk did not explore the topics in our paper (confidence sequences, explicit computationally efficient betting strategies, sampling without replacement, thorough numerical simulations, and so on). Indeed, their book has a thorough treatment of probability and finance, but with respect to statistical inference, there is little explicit methodology for practice. Perhaps they were aware of such a statistical utility, but they did not explicitly recognize or demonstrate the excellent power of betting in practice (when properly developed) for problems such as ours.

Statistical inference. Using the power of hindsight, we now know that Wald's influential work on the sequential probability ratio test was implicitly based on martingale techniques Wald [1945]. Wald derived many fundamental results that he required from scratch without having the general language that was being set up by Doob in parallel to his work. In the case of testing a simple null $H_0 : \theta = \theta^*$ against a composite alternative $H_0 : \theta \neq \theta^*$, Wald [1945, Eq (10:10)] suggests forming the likelihood ratio process $\prod_{i=1}^{n} f_{\theta_{i-1}}(X_i) / \prod_{i=1}^{n} f_{\theta^*}(X_i)$, where θ_{i-1} is a mapping from $X_1, \dots, X_{i-1}$ to Θ; in other words, θ_{i-1} is predictable. In the language of our paper, this is a predictable plug-in, and the first appearance of betting-like ideas in the statistical literature. However, beyond this passing equation in a parametric setup, the idea appears to have lain dormant.

Robbins (along with students and colleagues Siegmund, Darling, and Lai) quickly realized the power of Wald's and Ville's ideas as well as martingales more generally, and pursued a rather broad agenda around sequential testing and estimation, including the introduction and extensive study of confidence sequences and the method of mixtures [Darling and Robbins, 1967c,a,b, Robbins and Siegmund, 1968, 1969, 1970, 1972, 1974, Lai, 1976]. Robbins and Siegmund also analyzed Wald's "betting" test, and proved in some generality that its behavior is similar to a mixture likelihood ratio test [Robbins and Siegmund, 1974, Section 6]. Most of Wald's and Robbins' work was parametric, but Robbins did explicitly study the sub-Gaussian setting in some detail [Robbins, 1970]. Building on a vast literature of Chernoff-style concentration inequalities that exploded after Robbins' time, Howard et al. [2020, 2021] recently extended mixture methods of Robbins to derive confidence sequences under a large class of nonparametric settings using exponential supermartingales. Howard et al. [2020, 2021] recognized Wald's betting idea, but did not develop it nonparametrically beyond a brief mention in the paper as a direction for future work. The current work takes this natural next step in some thorough detail.

Information and coding theory. Soon after the seminal work of Shannon [1948], another researcher at AT&T Bell Labs, John Larry Kelly Jr. wrote a paper titled "A New Interpretation of Information Rate" which explicitly connected betting with the new field of information theory, complementing the work of Shannon [Kelly Jr, 1956]. In short, he proved that it is possible to bet on the symbols in a communication channel at odds consistent with their probabilities in order to have a gambler's wealth grow exponentially, with the exponent equaling the rate of transmission over the channel. More explicitly, given a sequence of Bernoulli random variables with probability $p > 1/2$, Kelly proved that betting a $(2p - 1)$ fraction of your current wealth on the next outcome being 1 is the unique strategy that maximizes the expected log wealth of the gambler.

When the probability p changes at each step in an unknown manner, the "universal coding" work of Krichevsky and Trofimov [1981] showed that a mixture method involving the Jeffreys prior and maximum likelihood can achieve nearly the optimal wealth in hindsight, with the expected log wealth of their strategy only being worse than the optimal oracle log-wealth by a factor that is logarithmic in the number of rounds; these observations work for any discrete alphabet, not just a binary. Cover's interest in these techniques spans several decades [Cover, 1974, 1984, 1987, Bell and Cover, 1980, 1988], culminating in his famous universal portfolio algorithm [Cover, 1991], that today forms a standard textbook topic in information theory.

There are other parts of information/coding theory that could be seen as related in some ways to betting through the use of (what are now called) e-variables: these include the topics of prequential model selection and minimum description length; see works by Rissanen [1984, 1998], Dawid [1984, 1997], Grünwald [2007], Grünwald et al. [2019], Li [1999] and references therein.

Online learning and sequential prediction under log loss. In the 1990s, the problems studied by Krichevsky, Trofimov, and Cover continued to be extended — often dropping the information theoretic context — under the title of sequential prediction under the logarithmic loss. In the active subfield of online learning, the previous results were effectively "regret bounds" against potentially adversarial sequences of observations, with a chapter devoted to the problem in the book on prediction, learning and games by Cesa-Bianchi and Lugosi [2006]. More recently, Orabona and colleagues such as Pal and Jun have found powerful implications of these ideas in deriving parameter-free algorithms for online convex optimization [Orabona and Pal, 2016, Orabona and Tommasi, 2017, Jun et al., 2017, Jun and Orabona, 2019].

Rakhlin and Sridharan [2017] found that deterministic regret inequalities can be used to derive concentration inequalities for martingales, connecting the two rich fields. Later, Jun and Orabona [2019] also derive concentration inequalities using their betting-based regret bounds, with explicit bounds derived in the sub-Gaussian and bounded settings. However, because regret bounds could be tight in rate but are typically loose in constants, the resulting concentration inequalities are not tight in practice. Thus, we view this line of work as important and complementary to our explorations, which are different in their motivation, derivation and practicality.

Typically, none of these lines of literature have cited the others. For example, the important paper of Rakhlin and Sridharan [2017] does not mention the work of Ville, Wald or Robbins, or even of Vovk and Shafer. Similarly, despite the books of Shafer and Vovk having a wonderful coverage of the history of probability and martingales stemming back hundreds of years, even their recent 2019 book [Shafer and Vovk, 2019] does not cite the coding theory and online learning literature very much, including the works of Orabona and coauthors [Orabona and Pal, 2016, Orabona and Tommasi, 2017, Jun et al., 2017, Cutkosky and Orabona, 2018, Jun and Orabona, 2019], Krichevsky and Trofimov [1981], or Rakhlin and Sridharan [2017]. Recent work of Orabona and colleagues also in turn has no mention of the books of Shafer and Vovk [2001, 2019], or works of Ville, Wald, Robbins, Howard, their coauthors and other recent authors. The work of Howard et al. [2020, 2021] does cite the Wald and Robbins literatures, as well as the books of Shafer and Vovk and pioneering work of Ville, but does not form connections to information/coding theory nor to online learning. The excellent book of Cesa-Bianchi and Lugosi [2006] does not cite Ville, the seminal martingale works of Robbins, or the 2001 book by Shafer and Vovk.[‖]

The reason for the lack of intersection of these parallel threads is likely manifold, and definitely far from malicious: (a) these works were and continue to be published in different literatures, (b) these works had different goals in mind, meaning that they were addressing different problems and often using different techniques, (c) our understanding of these literatures and their relationships is constantly evolving and far from complete; it is likely that no author has a command over all these parallel literatures, and indeed this should not be expected.

In the preface of their 2006 book, Cesa-Bianchi and Lugosi write

> Prediction of individual sequences, the main theme of this book, has been studied in various fields, such as statistical decision theory, information theory, game theory, machine learning, and mathematical finance. Early appearances of the problem go back as far as the 1950s, with the pioneering work of Blackwell, Hannan, and others. Even though the focus of investigation varied across these fields, some of the main principles have been discovered independently. Evolution of ideas remained parallel for quite some time. As each community developed its own vocabulary, communication became difficult. By the mid-1990s, however, it became clear that researchers of the different fields had a lot to teach each other. When we decided to write this book, in 2001, one of our main purposes was to investigate these connections and help ideas circulate more fluently. In retrospect, we now realize that the interplay among these many fields is far richer than we suspected. ... Today, several hundreds of pages later, we still feel there remains a lot to discover. This book just shows the first steps of some largely unexplored paths. We invite the reader to join us in finding out where these paths lead and where they connect.

Thus it is clear that Cesa-Bianchi and Lugosi already foresaw that there were many connections between the fields that have been unstated, underappreciated, undiscovered and underutilized. The connections we briefly point out above between these literatures, both historical and modern, are themselves new in their own right (not existing in any of the aforementioned books or papers) and may be considered a small contribution of this paper. A more thorough investigation of these connections may be the topic of a future survey paper, or indeed, a book on these topics.

[‖] Authors like like Rissanen [1984, 1998] and Dawid [1984, 1997] are not cited in most of these works, perhaps because the connections of their works to betting are indirect.