Randomized and exchangeable improvements of Markov's, Chebyshev's and Chernoff's inequalities

Aaditya Ramdas, Tudor Manole

Abstract. We present simple randomized and exchangeable improvements of Markov's inequality, as well as Chebyshev's inequality and Chernoff bounds. Our variants are never worse and typically strictly more powerful than the original inequalities. The proofs are short and elementary, and can easily yield similarly randomized or exchangeable versions of a host of other inequalities (for example, martingale inequalities by Doob and Ville) that employ Markov's inequality as an intermediate step. We point out some simple statistical applications involving tests that combine dependent e-values. In particular, we uniformly improve the power of universal inference, and obtain tighter betting-based nonparametric confidence intervals. Simulations reveal nontrivial gains in power (and no losses) in a variety of settings.

Key words and phrases: Randomization, exchangeability, e-values, universal inference, Markov, Chernoff, Ville, Doob, Chebyshev, Bernstein, Hoeffding, martingales.

1. INTRODUCTION

Consider a standard probability space $(\Omega, \mathcal{F}, P)$. Markov's inequality (MI) states that for any real-valued random variable X defined on this space and constant $a > 0$, we have

$$(1) \qquad P(|X| \geqslant 1/a) \leqslant a \cdot \mathbb{E}[|X|].$$

If X is nonintegrable, then the inequality trivially holds, so the reader may implicitly assume going forward that we deal with integrable X, without loss of generality. We now present three inequalities that are all strictly stronger than Markov's inequality, with an eye towards applications and improvements of other inequalities.

1.1 The exchangeable Markov inequality (EMI)

The following stronger version of Markov's inequality was recently noted by Manole and Ramdas (2023). Given its seemingly basic and fundamental nature, it may exist elsewhere in the literature. We will refer to it as the *exchangeable Markov inequality (EMI)*.

Recall that $X_1, \ldots, X_n$ are called exchangeable if the joint distribution of $(X_1, \ldots, X_n)$ equals that of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ for any permutation π of $\{1, \ldots, n\}$. A sequence $X_1, X_2, \ldots,$

is called exchangeable if $X_1, \ldots, X_n$ are exchangeable for any $n \geqslant 1$. Of course, iid random variables are exchangeable, but so are identical copies of the same random variable. Also, exchangeable random variables have the same expectation.

THEOREM 1.1 (Exchangeable Markov Inequality). *For any exchangeable sequence $X_1, X_2, \ldots$ sequence of integrable random variables and for any $a > 0$,*

$$
\begin{aligned}
(2) \quad & P\left(\exists t \geqslant 1 : \left| \frac{1}{t} \sum_{i=1}^{t} X_i \right| \geqslant 1/a \right) \\
& \leqslant P\left(\exists t \geqslant 1 : \frac{1}{t} \sum_{i=1}^{t} |X_i| \geqslant 1/a \right) \leqslant a \cdot \mathbb{E}[|X_1|].
\end{aligned}
$$

The same claim holds for a finite set of exchangeable random variables.

This is clearly a strictly stronger statement than Markov's inequality which effectively makes an identical claim only at $t = 1$ (or when all random variables are identical).

The proof of Theorem 1.1 is short (the first inequality is obvious, so we focus on the second). The exchangeability of $X_1, X_2, \ldots$ implies that the process $\left(\sum_{i=1}^{n} |X_i| / n \right)_{n \geqslant 1}$ forms a nonnegative reverse martingale; indeed, this follows from the fact that it can be rewritten as $(\mathbb{E}[|X_1| \| \mathcal{E}_n])_{n \geqslant 1}$, where $(\mathcal{E}_n)_{n \geqslant 1}$ denotes the exchangeable filtration generated by $X_1, X_2, \ldots$ (defined for instance in Manole and

Carnegie Mellon University, Pittsburgh, USA (e-mail: aramdas@cmu.edu). Massachusetts Institute of Technology, Boston, USA (e-mail: tmanole@mit.edu).

Ramdas (2023) and references therein). The claim then follows by invoking the time-reversed Ville inequality (recapped in Theorem 5.1 below, for completeness).

We point out immediately that the above inequality results in an improvement to subsampled universal inference (Wasserman et al., 2020), where repeated sample splitting gives rise to exchangeable "split likelihood ratios" on different subsets of the data. We return to this topic in more detail in Section 7.

A useful corollary of the EMI is as follows. Let $X_1, \ldots, X_n$ be any set of (potentially nonexchangeable) arbitrarily dependent random variables. Let π be a uniformly random permutation of $\{1, \ldots, n\}$. Then, for any $a > 0$,

(3)
$$P\left(\sup_{1 \leqslant t \leqslant n} \frac{1}{t} \sum_{i=1}^{t} |X_{\pi(i)}| \geqslant 1/a\right) \leqslant a \cdot \frac{\mathbb{E}[|X_1| + |X_2| + \cdots + |X_n|]}{n}.$$

Here, the original random variables are effectively made exchangeable by the random permutation, thus allowing us to invoke the original inequality.

We state a final variant of the inequality. Suppose we take N arbitrarily dependent random variables and put them in a bag. Suppose $X_{\pi(1)}, \ldots, X_{\pi(n)}$ be n samples drawn uniformly at random with or without replacement from this bag. Then, we have

(4)
$$P\left(\sup_{1 \leqslant t \leqslant n} \frac{1}{t} \sum_{i=1}^{t} |X_{\pi(i)}| \geqslant 1/a\right) \leqslant a \cdot \frac{\mathbb{E}[|X_1| + |X_2| + \cdots + |X_N|]}{N}.$$

This holds because the sampling process induces the exchangeability required for (2) to be invoked on the otherwise non-exchangeable random variables.

The aforementioned three variants of EMI are all relatively weak, in the sense that they do not improve with increasing t, n, N: there is no concentration of measure really happening, and indeed there cannot really be any since we have assumed so little about the underlying random variables (a first moment for each, and either an exchangeable or an arbitrary dependence structure). Hence we are under no illusions that these can be "much" stronger than the original Markov's inequality.

1.2 The uniformly-randomized Markov inequality (UMI)

Another improvement of Markov's inequality is the following *uniformly-randomized Markov inequality (UMI)*.

THEOREM 1.2 (Uniformly-randomized Markov Inequality). *Let X be a nonnegative random variable, and let $U \sim \text{Unif}(0, 1)$ be independent of X. Then, for any $a > 0$,*

(5)
$$P(X \geqslant U/a) = \mathbb{E}[\min(aX, 1)] \leqslant a \cdot \mathbb{E}[X].$$

The equality above is nontrivial even for nonintegrable X. Further, that equality becomes an inequality $\leqslant$ if U is stochastically larger than uniform. Last, if X is bounded, that is $X \in [0, C]$ almost surely for some $C > 0$, then the inequality in (5) holds with equality for any $a \leqslant 1/C$.

One may consider the first equality in (5) as Markov's *equality*. The proof is simple:

(6) $P(X \geqslant U/a) = \mathbb{E}[P(U \leqslant aX \mid X)] = \mathbb{E}[\min(aX, 1)]$,

yielding the claim. As mentioned above, if $X \leqslant 1/a$ almost surely, then (5) holds with equality, removing all looseness in Markov's inequality. Further, the statement is nontrivial for nonintegrable X. For example, when X is a standard Cauchy distribution and $a = 0.05$, we obtain $P(|X| \geqslant U/a)$ actually *equals* $(20\pi + \log(401) - 40\tan^{-1}(20))/(20\pi) \approx 0.127$, but Markov's inequality only states that $P(|X| \geqslant 1/a)$ is *at most* 0.127.

Clearly, UMI is strictly stronger than MI. Noting that $\mathbb{E}[U] = 1/2$ gives an intuitive idea of the extent of the gain. The proof is so simple that the above inequality may have been previously noted by other authors. But it appears to not be commonly taught or broadly known, and we have not found any source containing it so far. Since the UMI also holds if U is stochastically larger than uniform, Theorem 1.2 also gives an alternative proof to Markov's inequality (by choosing $U = 1$).

REMARK 1.3.   *What happens when Markov's inequality holds with equality? This only happens in a few rare cases, but it does happen, and in these cases UMI also holds with equality. For example, take the discrete distribution with point masses at 0 and 2, with probability half each, so that the mean is one. For $a = 1/2$, we have $\{X \geqslant 2\} = \{X = 2\}$ and thus $P(X \geqslant 2) = 1/2$, meaning Markov's holds with equality. In this case, UMI also holds with equality, because the event $\{X \geqslant 2U\}$ is almost surely equal to the event $\{X = 2\}$, since $U > 0$ almost surely. To summarize, Markov's inequality is only tight for a discrete random variable taking values in $\{0, 1/a\}$, while the UMI holds with equality for any random variable taking values in $[0, 1/a]$.*

We again note that the UMI improves the recent method of universal inference (Wasserman et al., 2020), by rejecting the null hypothesis when the split likelihood ratio (or its variants like the crossfit or subsampled likelihood ratio) exceed U/α rather than $1/\alpha$. We return to this "randomized universal inference" in Section 7.

Another particular application of UMI was recently and independently noted as a passing remark, for a different context, in a recent paper by Ignatiadis et al. (2024). It is known that, given an e-value E (defined in Section 6), the mapping from $E \mapsto 1/E$ is an admissible e-to-p calibrator (Vovk and Wang, 2021) (that converts an e-value into a p-value). But this admissibility result only applies

to deterministic maps. After proving that P/E is a valid p-value where P is a p-value and $E \perp P$, Ignatiadis et al. (2024) note that this implies U/E is also a valid (randomized) e-to-p calibrator that converts an e-value E into a p-value using independent randomization U, thus rendering $1/E$ as inadmissible if randomization is allowed. We expand on this type of application in Section 6.

Last, it turns out that the UMI and EMI can be combined into a single statement that implies both; this is postponed to Theorem 5.3.

1.3 The additively-randomized Markov inequality (AMI)

Instead of multiplicative randomization (as done in the previous subsection), we may consider additive randomization. In this context, Huber (2019) recently proved an interesting "smoothed Markov" inequality: for any nonnegative X and constant $\epsilon > 0$,

$$(7) \qquad P(X + B \geqslant \epsilon) \leqslant \frac{\mathbb{E}[X]}{2\epsilon},$$

where B is an independent uniform random variable on $[-\epsilon, \epsilon]$. However, while the right-hand side halves the bound of Markov's inequality (as his paper title suggests), the left-hand side is not directly comparable. Indeed, when calculating $P(X \geqslant \epsilon - B)$, the right-hand side $\epsilon - B$ is not always bigger than ϵ. Hence, it appears that the bound is not in general comparable to Markov's inequality. Huber derives many interesting consequences of this inequality, including extensions to Chebyshev's and Chernoff's inequalities, and once again they appear incomparable to the original inequalities.

We present the following result, which (unlike Huber's) is stronger than Markov's inequality.

PROPOSITION 1.4 (Additively-randomized Markov Inequality). *Let $\epsilon > 0$. Given a nonnegative random variable X, and an independent random variable $A \sim Unif(0, \epsilon)$, it holds that*

$$(8) \qquad P(X \geqslant \epsilon - A) \leqslant \frac{\mathbb{E}[X]}{\epsilon}.$$

The proof is simple. Since $P(A \geqslant \epsilon - x) = x/\epsilon$ for $x \leqslant \epsilon$, the left-hand side simplifies to

$$P(A \geqslant \epsilon - X) = \mathbb{E}[P(A \geqslant \epsilon - X|X)] = \mathbb{E}[\min(X, \epsilon)]/\epsilon,$$

implying the claim.

It is also not hard to see that our additively-randomized Markov inequality (Proposition 1.4) is actually equivalent to our earlier uniformly-randomized Markov inequality (Theorem 1.2). To see this, write $A = \epsilon U$, where U is uniform on $[0, 1]$. Then $P(X \geqslant \epsilon - A) = P(X \geqslant \epsilon(1 - U)) = P(X \geqslant \epsilon U')$, where $U' = 1 - U$ is also uniform on $[0, 1]$. Writing $a = 1/\epsilon$ equates the two claims.

In Appendix A, we further discuss the relationship of the above bound to Huber's smoothed Markov inequality. We show, in particular, how the two types of bounds can be used to derive each other, despite having different interpretations and implications.

In the rest of this paper, we will continue to use our multiplicative version, because we think the resulting expressions are cleaner, but readers may find the additive version more useful in some settings, which is our reasoning for recording Proposition 1.4 as a separate result.

1.4 Contributions and paper outline

Having already introduced multiple new generalizations of Markov's inequality, we next point out several new concentration inequalities that result out of their use and/or combination. Section 2 derives randomized and exchangeable improvements of Chebyshev's inequality (with Appendix C.1 containing an improvement to Cantelli's inequality, as a way of exemplifying a more general proof technique). Section 3 does the same for Hoeffding's inequality, and points out that the same techniques improve any Chernoff bound (extensions of the Bernstein and empirical Bernstein inequalities are in Appendix C.2). Section 4 shows how to derive a randomized improvement of Ville's inequality for forward supermartingales. Then, Section 5 randomizes the reverse Ville's inequality (for reverse submartingales), and as a consequence derives an inequality that combines the strengths of the EMI and UMI into a single inequality (Theorem 5.2). After presenting some of these improvements (and omitting others for brevity), we describe some statistical applications. Section 6 produces more powerful tests using arbitrarily dependent e-values. Section 7 uniformly improves different versions of the recent universal inference (Wasserman et al., 2020) methodology. Section 8 improves betting-based tests and confidence intervals that are an exciting development in nonparametric statistics (Waudby-Smith and Ramdas, 2024). Section 9 explores the (often large) improvements in power obtained in a variety of simulations. Section 10 contains an extended discussion, including concerns about reproducibility of randomized tests, and the possibility of avoiding external randomization entirely by utilizing the internal randomness of the data. Section 11 contains a brief conclusion.

2. RANDOMIZED & EXCHANGEABLE CHEBYSHEV INEQUALITY

In the rest of this section and paper, U is always a uniform random variable on $[0, 1]$, independent of all other random variables.

2.1 Uniformly-randomized Chebyshev inequality

For any random variable X having variance (at most) σ^2, Chebyshev's inequality states that

$$(9) \qquad P(|X - \mathbb{E}X| \geqslant k\sigma) \leqslant 1/k^2.$$

The proof is transparent: one just squares both terms within the probability on the left-hand side, and applies Markov's inequality. More generally, consider $n \geqslant 1$ i.i.d. random variables $X_1, \ldots, X_n$ with variance σ^2, and define $\bar{X}_n := (X_1 + \cdots + X_n)/n$. Chebyshev's inequality implies that for any $k > 0$,

$$(10) \qquad P\left(|\bar{X}_n - \mathbb{E}X| \geqslant k\frac{\sigma}{\sqrt{n}}\right) \leqslant 1/k^2.$$

Our *uniformly-randomized Chebyshev inequality* reads as follows:

THEOREM 2.1 (Uniformly-randomized Chebyshev Inequality). *Let $X_1, \ldots, X_n$ be i.i.d. random variables with variance σ^2, and let $U \sim \text{Unif}(0,1)$ be an independent random variable. Then, for any $k > 0$,*

$$(11) \qquad P\left(|\bar{X}_n - \mathbb{E}X| \geqslant k\sigma\sqrt{\frac{U}{n}}\right) \leqslant 1/k^2.$$

As before, the same result holds when U is stochastically larger than uniform.

This is clearly a tighter claim than the original: the probability of exceeding a random (but always smaller) threshold is identical. Noting that $\mathbb{E}[\sqrt{U}] = 2/3$ while $\mathbf{Var}(\sqrt{U}) = 1/18$ gives an intuitive idea of the extent of the gain: the obtained confidence intervals will be 2/3 as wide using our improved inequality.

The proof is similar to the UMI (6):

$$P\left(|\bar{X}_n - \mathbb{E}X| \geqslant k\sigma\sqrt{\frac{U}{n}}\right) = \mathbb{E}\left[P\left(U \leqslant \frac{n|\bar{X}_n - \mathbb{E}X|^2}{k^2\sigma^2}\Big|X\right)\right]$$
$$\leqslant \mathbb{E}\left[\frac{n|\bar{X}_n - \mathbb{E}X|^2}{k^2\sigma^2}\right] = 1/k^2.$$

The sole inequality exists, despite U being exactly uniform, because $\frac{n|\bar{X}_n - \mathbb{E}X|^2}{k^2\sigma^2}$ could be larger than one. If one replaces the resulting inequality by $\mathbb{E}\left[\frac{n|\bar{X}_n - \mathbb{E}X|^2}{k^2\sigma^2} \wedge 1\right]$, then the inequality would turn to equality, but the subsequent equality ($= 1/k^2$) would turn into an inequality.

It is clear that the i.i.d. assumption above, and in the following sections, is made for convenience, and can be weakened to give many variants of the above. It is also possible to prove a randomized version of Cantelli's inequality (Cantelli, 1929), which is a one-sided analogue of Chebyshev's inequality. We present such a result in Appendix C.

REMARK 2.2. *We thank an anonymous referee for pointing out that by similar lines as above, it is also possible to derive the following variant of the uniformly-randomized Chebyshev inequality:*

$$(12) \qquad P\left(|\bar{X}_n - \mathbb{E}X| \geqslant k^2\sigma\frac{U}{\sqrt{n}}\right) \leqslant 1/k^2.$$

Although this bound has a poor dependence on k, it has a more favourable dependence on U, since $\mathbb{E}[U] = 1/2 < \mathbb{E}[\sqrt{U}]$. A simple derivation shows that equation (12) is sharper in expectation than the uniformly-randomized Chebyshev inequality (11) when the value of k is sufficiently small, specifically $k \leqslant 4/3$.

2.2 Exchangeable Chebyshev inequality

If the data are exchangeable rather than i.i.d., we have the following claim:

THEOREM 2.3 (Exchangeable Chebyshev Inequality). *If $X_1, X_2, \ldots$ is a sequence of exchangeable random variables with variance at most σ^2, then for any $k > 1$,*

$$(13) \qquad P\left(\sup_{m \geqslant 1}|\bar{X}_m - \mathbb{E}X| \geqslant k\sigma\right) \leqslant 1/k^2.$$

If the random variables are further assumed to be i.i.d., then for any $n \geqslant 1$, we have

$$(14) \qquad P\left(\sup_{m \geqslant n}|\bar{X}_m - \mathbb{E}X| \geqslant \frac{k\sigma}{\sqrt{n}}\right) \leqslant 1/k^2.$$

Note that (13) improves (9) by recovering it either at $m = 1$ or when all random variables are identical. Unlike the various bounds presented for the i.i.d. case, (13) does not improve as the sample size increases. But improvements are not possible without further assumptions, because when all X_i are identically equal to X, the statement reduces to a claim about X, with no role for concentration of measure.

The proof is simple: defining $R_m := |\bar{X}_m - \mathbb{E}X|^2/\sigma^2$, a short calculation invoking Jensen's inequality reveals that $(R_m)_{m \geqslant 1}$ is a nonnegative reverse submartingale with $\mathbb{E}[R_1] \leqslant 1$. The time-reversed Ville inequality (Theorem 5.1) then yields our claim. Essentially the same proof can be obtained by defining $Y_i = |X_i - \mathbb{E}X|^2/\sigma^2$. Applying the exchangeable Markov inequality, we infer that $P(\sup_{m \geqslant 1} \bar{Y}_m \geqslant k^2) \leqslant 1/k^2$, which implies (13) by Jensen's inequality. For the second part, one may apply the same proof as above, but now only considering the submartingale $(R_m)_{m \geqslant n}$ starting at time n. Since $\mathbf{Var}[R_n] = \sigma^2/n$ under the i.i.d. assumption, equation (14) follows.

REMARK 2.4. *The above inequalities are not to be confused with Kolmogorov's generalization of Chebyshev's inequality which states that if $X_1, X_2, \ldots, X_n$ are i.i.d. (not*

just exchangeable) random variables with variance at most σ^2, then for any $k > 1$ and $n \geqslant 1$,

$$(15) \quad P\left(\sup_{m \leqslant n} |\sum_{i=1}^{m} (X_i - \mathbb{E}X)| \geqslant k\sigma \sqrt{n}\right) \leqslant 1/k^2.$$

Technically, the i.i.d. assumption above can be weakened to just independence, and even to a martingale dependence assumption, but we omit these for simplicity.

3. RANDOMIZED AND EXCHANGEABLE CHERNOFF BOUNDS

We present randomized and exchangeable variants of Chernoff bounds below. We remark that every Chernoff bound, including matrix concentration bounds, or self-normalized concentration inequalities, are improved via the same technique presented below (Howard et al., 2020), but it is impractical to develop every one of these, so we just pursue three of them: the Hoeffding bound below, Bernstein's inequality in Appendix C.2, and the empirical Bernstein inequality in Appendix C.3.

3.1 Uniformly-randomized Hoeffding inequality

Recall that X is called σ-subGaussian if for any constant λ,

$$(16) \quad \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leqslant \exp(\lambda^2\sigma^2/2).$$

Hoeffding's inequality states that if X is σ-subGaussian, then for any $\epsilon > 0$,

$$(17) \quad P(X - \mathbb{E}[X] \geqslant \sigma\epsilon) \leqslant \exp(-\epsilon^2/2).$$

More generally, considering n i.i.d. σ-subGaussian random variables $X_1, \ldots, X_n$, the above inequality implies that

$$(18) \quad P\left(\bar{X}_n - \mathbb{E}[X] \geqslant \sigma\epsilon\right) \leqslant \exp(-n\epsilon^2/2).$$

Setting the right-hand side to equal $\alpha \in (0, 1)$, it can be rewritten as

$$(19) \quad P\left(\bar{X}_n - \mathbb{E}[X] \geqslant \sigma\sqrt{\frac{2\log(1/\alpha)}{n}}\right) \leqslant \alpha.$$

An identical bound can also be derived on the other tail. As we shall soon see, the proof proceeds by multiplying both sides within the probability by $\lambda > 0$, exponentiating both sides, then applying Markov's inequality, and finally tuning λ.

Our *uniformly-randomized Hoeffding inequality* reads as follows.

THEOREM 3.1 (Uniformly-randomized Hoeffding Inequality). *Let $X_1, \ldots, X_n$ be i.i.d. σ-subGaussian random variables, and let $U \sim Unif(0,1)$ be an independent random variable. Then, for all $\alpha \in (0,1)$,*
(20)
$$P\left(\bar{X}_n - \mathbb{E}[X] \geqslant \sigma\sqrt{\frac{2\log(1/\alpha)}{n}} + \sigma\frac{\log(U)}{\sqrt{2n\log(1/\alpha)}}\right) \leqslant \alpha.$$

The above bound also holds in non-i.i.d. settings under a certain martingale dependence structure, where each X_i is σ-subGaussian conditional on $X_1, \ldots, X_{i-1}$, in which case the left-hand side $\mathbb{E}[X]$ is replaced by $\sum_{i=1}^{n} \mathbb{E}[X_i|X_1, \ldots, X_{i-1}]/n$. We omit the details for simplicity; see Howard et al. (2020) for details.

The original Hoeffding inequality (19) is recovered by replacing U with 1. Since $\log U < 0$, this is a strictly tighter bound than the original. Note that $\mathbb{E}[\log U] = -1$ and $\mathbf{Var}(\log U) = 1$, perhaps giving an intuitive idea of the extent of the gain.

The proof is simple: definition (16) implies that the random variable $\exp(\lambda \sum_{i=1}^{n}(X_i - \mathbb{E}[X]) - \frac{\lambda^2}{2}\sigma^2 n)$ has expected value at most one. Since it is also nonnegative, we can apply our randomized Markov's inequality. Setting $\lambda = \sqrt{\frac{2\log(1/\alpha)}{\sigma^2 n}}$, and rearranging terms, yields the above claim. (Note that λ cannot be a function of U for the proof to work.)

REMARK 3.2. *The bound in (20) can be rewritten by setting $\epsilon = \sqrt{2\log(1/\alpha)/n}$ as:*

$$(21) \quad P\left(\bar{X}_n - \mathbb{E}[X] \geqslant \sigma\epsilon + \frac{\sigma\log U}{\epsilon n}\right) \leqslant \exp(-n\epsilon^2/2),$$

which is easier to compare to (18). Also, a $(1-\alpha)$-confidence interval for the mean $\mathbb{E}[X]$ is

$$(22) \quad \bar{X}_n \pm \left[\sigma\sqrt{\frac{2\log(2/\alpha)}{n}} + \sigma\frac{\log(U)}{\sqrt{2n\log(2/\alpha)}}\right].$$

The usual half-width of Hoeffding's inequality simply involves the first term above, that is $\sigma\sqrt{\frac{2\log(2/\alpha)}{n}}$. Since $\mathbb{E}[\log U] = -1$, the expected improvement in width is $\sigma/\sqrt{2n\log(2/\alpha)}$, and the expected relative improvement in width is the ratio of the aforementioned quantities (latter divided by former), given by $1/(2\log(2/\alpha))$. This equals $1/(2\log(40)) \approx 0.14$ when $\alpha = 0.05$, meaning we expect to get about a 14% improvement in width over Hoeffding's inequality. Indeed, we observe exactly this factor of 14% in our later simulations.

REMARK 3.3. *The Chernoff technique, while attractive due to its simplicity, leads to inefficient tail inequalities. The original work of Hoeffding (1963, Theorem 1) provides a simple sharpening of the bound (18) for bounded random variables. A variety of other strictly tighter tail bounds have been developed in the literature—see for instance Bentkus (2002); Bentkus et al. (2006). Each of these tail bounds can be derived by applying MI to a deterministic function of $\bar{X}_n$, and can therefore be randomized using the UMI in the same way that we have done here. We do not pursue this avenue further in order to keep our exposition simple. Nevertheless, in Appendix C,*

we derive a general inequality which could be used to derive such sharper randomized tail bounds, and which also unifies Theorems 1.2, 2.1 and 3.1.

We end by pointing out the possibility that the interval (22) can be empty. Despite this, it maintains its frequentist coverage as discussed in Section 10.3, but we nevertheless propose a practical fix in Section 10.5.

3.2 Exchangeable Hoeffding Inequality

If the data are exchangeable rather than i.i.d. σ-subGaussian, then one can verify via Jensen's inequality that $\bar{X}_n$ is still σ-subGaussian. In fact, we have the following stronger claim:

THEOREM 3.4 (Exchangeable Hoeffding Inequality). *If $X_1, X_2, \ldots$ are an exchangeable sequence of σ-subGaussian random variables, then for any $\epsilon > 0$,*

$$(23) \qquad P\left(\sup_{m \geqslant 1} |\bar{X}_m - \mathbb{E}X| \geqslant \epsilon\sigma\right) \leqslant \exp(-\epsilon^2/2).$$

If the random variables are further assumed to be i.i.d., then we have for any $n \geqslant 1$,

$$(24) \qquad P\left(\sup_{m \geqslant n} |\bar{X}_m - \mathbb{E}X| \geqslant \frac{\epsilon\sigma}{\sqrt{n}}\right) \leqslant \exp(-\epsilon^2/2).$$

Equation (23) improves (17) by recovering it either at $m = 1$ or when all random variables are identical. The proof mimics that of Theorem 2.3, except by using $Y_i = \exp(\lambda(X_i - \mathbb{E}X) - \lambda^2\sigma^2/2)$, which has expectation at most one.

REMARK 3.5. *The above bound is not to be confused with the Azuma-Hoeffding inequality which states that if $X_1, X_2, \ldots$ are i.i.d. (not merely exchangeable) σ-subGaussian random variables, then for any $\epsilon > 0$,*

$$(25) \qquad P\left(\sup_{m \leqslant n} \sum_{i=1}^{m} (X_i - \mathbb{E}X) \geqslant \epsilon\sigma\sqrt{n}\right) \leqslant \exp(-\epsilon^2/2).$$

Of course, the above inequality holds under a certain type of martingale dependence structure, that we omit for simplicity. The Azuma-Hoeffding inequality, despite its fame, is itself loose and was uniformly improved by Howard et al. (2020).

4. A RANDOMIZED IMPROVEMENT OF VILLE'S INEQUALITY

There is a fundamental inequality for nonnegative supermartingales called Ville's inequality, whose time-reversed version led to the exchangeable Markov inequality. In fact, both Ville's and reverse Ville's inequalities are themselves stronger statements than Markov's inequality.

Here we ask the question: is there a randomized improvement of these inequalities? The answer is subtle—both yes and no in some sense—and we return to answer this question in Section 4.2 after giving a brief introduction to these inequalities below.

4.1 Ville's inequality for forward supermartingales

Ville (1939) proved that if $X_1, X_2, \ldots$ form a nonnegative supermartingale (with respect to any filtration) then for any constant $a > 0$,

$$(26) \qquad P\left(\sup_{t \geqslant 1} X_t \geqslant 1/a\right) \leqslant a \cdot \mathbb{E}[X_1].$$

Markov's inequality, of course, replaces $\sup_{t \geqslant 1} X_t$ with just X_1, and thus (26) is strictly stronger. This inequality appears for example in the foundational books of game-theoretic probability Shafer and Vovk (2005, 2019) and much of the recent literature on time-uniform concentration inequalities (Howard et al., 2020, 2021), in which it is called Ville's inequality.

The proof is simple. Define the stopping time $\tau := \inf\{t \geqslant 1 : X_t \geqslant 1/a\}$, where $\inf \emptyset = \infty$. For any fixed m, Markov's inequality implies

$$(27) \quad P(\tau \leqslant m) = P(X_{\tau \wedge m} \geqslant 1/a) \leqslant a \cdot \mathbb{E}[X_{\tau \wedge m}] \leqslant a \cdot \mathbb{E}[X_1],$$

where the second inequality follows by Doob's optional stopping theorem. Letting $m \to \infty$ and using the bounded convergence theorem yields $P(\tau < \infty) \leqslant a \cdot \mathbb{E}[X_1]$, proving (26).

4.2 Randomizing Ville's inequality

Howard et al. (2021, Lemma 3) implies that there are actually three equivalent statements of Ville's inequality: if $M = (M_t)_{t \geqslant 0}$ is a nonnegative supermartingale with respect to a filtration $\mathbb{F} = (\mathcal{F}_t)_{t \geqslant 0}$, and $\mathbb{E}[M_0] = 1$, then for any $a > 0$, the following three statements hold and imply each other:

(28a)
$$P(\exists t \geqslant 0 : M_t \geqslant 1/a) \leqslant a.$$

$$(28b) \qquad P(M_\tau \geqslant 1/a) \leqslant a \text{ for any } \mathbb{F}\text{-stopping time } \tau.$$

$$(28c) \qquad P(M_T \geqslant 1/a) \leqslant a \text{ for any } \mathcal{F}_\infty\text{-measurable}$$

$$(28d) \qquad \qquad \text{random time } T.$$

Despite the fact that the above three statements imply each other, it turns out that (only) the second of these inequalities can be randomized to yield the following result.

THEOREM 4.1. *For any nonnegative supermartingale M with $\mathbb{E}[M_0] \leqslant 1$, we have*

$$(29) \qquad P(M_\tau \geqslant U/a) \leqslant a \text{ for every } \mathbb{F}\text{-stopping time } \tau,$$

where U is (stochastically larger than) uniform on $[0, 1]$ and is independent of $\mathbb{F}$ (and thus independent of M and τ).

The proof is simple: the optional stopping theorem for nonnegative supermartingales implies that for any stopping time without restriction, $\mathbb{E}[M_\tau] \leqslant 1$, to which we apply the UMI (5).

The entire discussion above does not require time to be discrete: all the statements hold in continuous time as well. In continuous time, it is easy to observe that equation (28a) cannot be improved by randomization: the inequality actually holds with equality for $\exp(\lambda B_t - \lambda^2 t/2)$, where λ is any nonzero constant and $(B_t)_{t \geqslant 0}$ is a standard Brownian motion (see for instance Durrett (2019), Exercise 7.5.2).

We now present a corollary of Theorem 4.1 that has direct implications for sequential testing.

COROLLARY 4.1.1 (Randomized Ville Inequality). *For any nonnegative supermartingale M with $\mathbb{E}[M_0] \leqslant 1$, and any $\mathbb{F}$-stopping time τ, we have*

$$(30) \qquad P(\exists t < \tau : M_t \geqslant 1/a \text{ or } M_\tau \geqslant U/a) \leqslant a,$$

where U is (stochastically larger than) uniform on $[0, 1]$ and is independent of $\mathbb{F}$ (and thus independent of M and τ).

Given any stopping time τ, (30) is obtained by applying Theorem 4.1 to the stopping time $\tau' := \min\{\tau, \gamma\}$, where $\gamma := \inf\{t : M_t \geqslant 1/a\}$. We now discuss the implications of the above randomized Ville inequality.

4.3 Implications for safe anytime-valid inference

Ville's inequality plays a central role in modern sequential statistics (Howard et al., 2020, 2021), and in particular within "game-theoretic statistics and safe, anytime-valid inference" (Ramdas et al., 2020, 2023). In the latter context, the above randomized variant of Ville's inequality (29) can improve power in a concrete way. To understand why, first note that one usually constructs M to be a nonnegative supermartingale under the null hypothesis, such that it increases to infinity under the alternative. Then, $\tau := \inf\{t : M_t \geqslant 1/\alpha\}$ is a stopping time of special importance, since Ville's inequality implies that we can reject the null if $\tau < \infty$, while controlling the type-I error at level α.

However, this stopping rule is not literally followed due to its potentially unbounded nature. One may terminate an experiment even if the above stopping time has not been reached: in a simulation, we have bounded computational resources, which means we only really allow for rejection before some maximum time $t_{\max}$, and in real experiments, one may terminate due (for example) budget constraints. This motivates the following three-step rule for sequential testing with the randomized Ville's inequality:

- Collect data and continuously monitor the test statistic process M (that is guaranteed to be a nonnegative supermartingale under the null).

- If M ever crosses $1/a$, stop and reject the null. Else, stop at any $\mathcal{F}$-stopping time τ.
- Draw an independent random variable U that is (stochastically larger than) uniform on $[0, 1]$. Reject the null if $M_\tau \geqslant U/\alpha$.

Theorem 4.1 implies that the above rule yields a bona fide level-α sequential test that is valid under continuous monitoring and adaptive stopping. This is clearly more powerful than the usual rule employed in the aforementioned papers (and references cited therein), which only reject the null when $M_\tau \geqslant 1/\alpha$.

We note in passing that Theorem 4.1 also applies to a larger class of processes than nonnegative supermartingales. These are called "e-processes", and play a particularly key role in sequential composite null testing. However, we omit the details for brevity, and refer instead to the aforementioned survey by Ramdas et al. (2023).

We end with the following note. Thanks to a duality between sequential tests and sequential estimation using confidence sequences (CSs), the above observations also have implications for constructing CSs. A CS is a time-uniform or anytime-valid generalization of a confidence interval. Formally, a $(1 - \alpha)$-CS for a parameter θ is a sequence $(C_n(\alpha))_{n \geqslant 1}$ of confidence intervals (one for each sample size n) that are valid at arbitrary stopping times, meaning that it satisfies $P(\theta \in C_\tau(\alpha)) \geqslant 1 - \alpha$ for any $\mathcal{F}$-stopping time τ, or equivalently it satisfies $P(\forall n \geqslant 1 : \theta \in C_n(\alpha)) \geqslant 1 - \alpha$. Since they are usually (or in fact, essentially always, as per Ramdas et al. (2020); Waudby-Smith and Ramdas (2024)) obtained by inverting a family of sequential tests based on Ville's inequality, our randomized Ville's inequality improves the CI at the final stopping time. To clarify, we have the following result for any CS: For any $\mathcal{F}$-stopping time τ, we have

$$(31) \qquad P(\exists t < \tau : \theta \notin C_\tau(\alpha) \text{ or } \theta \notin C_\tau(\alpha/U)) \leqslant \alpha,$$

where U is a uniform that is independent of the data and thus the stopping time (one may imagine it to be drawn after stopping). To summarize, if we ever stop a sequential experiment in which we were tracking a CS, the very last confidence interval that we report can be at level α/U, and this would still have an overall miscoverage of at most α.

5. RANDOMIZING THE TIME-REVERSED VILLE INEQUALITY

Directly inspired by Ville (1939), the time-reversed Ville inequality was first proved by Doob (1940) for reverse martingales, and for instance by Lee (1990) and Christofides and Serfling (1990) for reverse submartingales. We also refer to Manole and Ramdas (2023) for two self-contained proofs of this result. The statement is given as follows.

THEOREM 5.1 (Time-reversed Ville inequality). *Let $(X_t)_{t=1}^{\infty}$ be a nonnegative reverse submartingale with respect to a reverse filtration $(\mathcal{E}_t)_{t=1}^{\infty}$. Then, for any $a > 0$,*

$$(32) \qquad P\left(\sup_{t \geq 1} X_t \geq 1/a\right) \leq a \cdot \mathbb{E}[X_1].$$

We provide a self-contained proof in Appendix B, which we briefly outline here. The proof is similar to that of Ville's inequality, with changes to account for the reversed nature of the process. Given $m \geq 1$, define $\tau := \sup\{1 \leq t \leq m : X_t \geq 1/a\}$, where $\sup \emptyset = -\infty$. Markov's inequality implies

$$(33) \quad P(\tau \geq 1) = P(X_{\tau \vee 1} \geq 1/a) \leq a \cdot \mathbb{E}[X_{\tau \vee 1}] \leq a \cdot \mathbb{E}[X_1].$$

To prove the last inequality, notice that the process $Y_t = X_{m-t+1}$, $1 \leq t \leq m$, is a forward submartingale, and $\eta := m - \tau + 1$ is a stopping time with respect to the same filtration as $(Y_t)_{t=1}^{m}$, thus by the optional stopping theorem,

$$\mathbb{E}[X_{\tau \vee 1}] = \mathbb{E}[Y_{\eta \wedge m}] \leq \mathbb{E}[Y_m] = \mathbb{E}[X_1].$$

Noting that $\{\tau \geq 1\} = \{\sup_{1 \leq t \leq m} X_t \geq 1/a\}$ yields

$$P\left(\sup_{1 \leq t \leq m} X_t \geq 1/a\right) \leq a \cdot \mathbb{E}[X_1].$$

Sending $m \to \infty$ yields our claim.

5.1 The uniformly-randomized time-reversed Ville inequality

Given a reverse filtration $(\mathcal{E}_t)_{t=1}^{\infty}$, we will say that τ is a reverse stopping time if it satisfies $\{\tau = t\} \in \mathcal{E}_t$ for all $t \geq 1$. We then have the following statement.

THEOREM 5.2 (Uniformly-randomized time-reversed Ville inequality). *Let $(X_t)_{t=1}^{\infty}$ be a nonnegative reverse submartingale, and τ a reverse stopping time, both with respect to a reverse filtration $(\mathcal{E}_t)_{t=1}^{\infty}$. Let $U \sim Unif(0,1)$ be independent of $(\mathcal{E}_t)$. Then, for any $a > 0$,*

$$(34) \qquad P(X_\tau \geq U/a) \leq a \cdot \mathbb{E}[X_1].$$

Theorem 5.2 is a uniformly-randomized analogue of the time-reversed Ville inequality in Theorem 5.1. This last has been used for instance by Manole and Ramdas (2023) to derive nonparametric sequential goodness-of-fit and two-sample hypothesis tests using divergences between empirical probability distributions as test statistics. When the validity of these sequential tests is only needed at arbitrary reverse stopping times τ, their power can immediately be improved with uniform-randomization by Theorem 5.2, in much the same way as we described in Section 4.3.

The proof of Theorem 5.2 is straightforward: by reasoning as in the proof of Theorem 5.1, above we have $\mathbb{E}[X_\tau] \leq \mathbb{E}[X_1]$, thus the claim follows by applying UMI to X_τ.

5.2 A uniformly randomized variant of the EMI

We can use Theorem 5.2 to obtain the following variant of the EMI which we refer to as the *exchangeable and uniformly-randomized Markov inequality (EUMI)*.

THEOREM 5.3 (EUMI). *Let $X_1, \ldots, X_n$ be a set of exchangeable random variables. Then, for any $a \in (0,1)$,*

$$(35)$$
$$P\left(X_1 \geq U/a \text{ or } \exists t \leq n : \left|\sum_{i=1}^{t} X_i/t\right| \geq 1/a\right) \leq a \cdot \mathbb{E}|X_1|,$$

where U is a uniform random variable on $[0,1]$ that is independent of $X_1, \ldots, X_n$.

The proof follows by defining the stopping time $\tau := 1 \vee \sup\{1 \leq t \leq n : \sum_{i=1}^{t} X_i/t \geq 1/a\}$. Then, Theorem 5.2 implies that

$$(36) \qquad P\left(\left|\frac{1}{\tau}\sum_{i=1}^{\tau} X_i\right| \geq U/a\right) \leq a \cdot \mathbb{E}|X_1|,$$

which is mathematically equivalent to (35). It is easy to see that an analogous statement holds for infinite sequences of exchangeable random variables, by removing the upper bound on t in the definition of τ.

Equation (35) is one way of combining the strengths of the EMI and UMI. Note that the EUMI is stronger than both the UMI and the EMI, and indeed implies both of them. This inequality has important implications for (more powerful) statistical testing, which we discuss in the following sections.

6. CONSTRUCTING RANDOMIZED TESTS WITH E-VALUES

Despite e-values not being defined formally yet, we already used them implicitly in the proofs of all preceding theorems: indeed the random variable Y_i in the proofs of Theorem 2.3 and Theorem 3.4 are e-values, as is $X/\mathbb{E}[X]$ in the proof of Markov's inequality.

Before defining e-values formally, we give a brief summary of what's to come, in order to orient the reader. Let E_1 and E_2 be arbitrarily dependent e-values for testing a given hypothesis. In order to achieve a level α test, the natural way to combine them is to average them into a combined e-value and check if $(E_1 + E_2)/2 \geq 1/\alpha$, whose validity is guaranteed by Markov's inequality.

However, a simple and uniform improvement is as follows: choose $i \in \{1,2\}$ randomly with equal probability, and first check if $E_i \geq 1/\alpha$, and if not then check whether $(E_1 + E_2)/2 \geq 1/\alpha$. This is also a level-$\alpha$ test, due to the EMI. Another valid randomized level-α test is to reject if $(E_1 + E_2)/2 \geq U/\alpha$ where U is an independent uniform random variable, whose validity is guaranteed by the UMI. We expand on some statistical applications of these ideas below.

6.1 A brief review of e-values

For a set of distributions $\mathcal{P}$, an e-value[1] for the null hypothesis $H_0 : P \in \mathcal{P}$ is a nonnegative random variable X such that

$$\mathbb{E}_P[X] \leqslant 1 \text{ for all } P \in \mathcal{P}.$$

Since e-values are likely to be small under the null (and hopefully large under the alternative), a level-α test is given by:

$$(37) \qquad \text{reject } H_0 \text{ if } X \geqslant 1/\alpha.$$

This test controls type-1 error (nonasymptotically and hence uniformly over $\mathcal{P}$) by Markov's inequality (1). Without knowing any further details about the distribution of X, this rule does not appear to be improvable in general.

However, we note that with a little external randomization, the rule (37) can be made (usually strictly) more powerful. The UMI (5) implies that if U is an independent uniform random variable on $[0, 1]$, then the rule

$$(38) \qquad \text{reject } H_0 \text{ if } X \geqslant U/\alpha,$$

controls the type-I error at level α. In other words, $\min(U/X, 1)$ is a valid p-value; this fact was independently recently pointed out in Ignatiadis et al. (2024).

We will see that such uniform randomization can effectively be used to improve on (arguably natural) ways to combine e-values to yield tests, but a new type of test is opened up as a consequence of the EMI (2), and the EUMI (35).

---

[1]If $\mathcal{P} = \{P\}$ is a singleton, then all optimal e-values take the form of dQ/dP, that is likelihood ratios of Q against P, for some (implicit or explicit) alternative Q. So, technically e-values have been around for 100 years masquerading as likelihood ratios (and Bayes factors). The recent christening of the term "e-value" is simply to recognize the importance of a more general concept that has utility much beyond point nulls. Indeed, beyond the singleton case, e-values can be viewed as nonparametric/composite generalizations of likelihood ratios to complex settings involving nonparametric and composite nulls and alternatives. Even for this setting, e-values have technically been around for over 50 years (Robbins, 1970), appearing in the form of stopped nonnegative supermartingales and implicitly within the proofs of Chernoff bounds (Howard et al., 2020). The concept appears to have simply floated around without a unified name for 50 years, until several authors—who had a priori used different terms for the same concept—simultaneously decided to converge to the terminology "e-value" a few years ago (Shafer, 2021; Vovk and Wang, 2021; Grünwald et al., 2024; Ramdas et al., 2020; Wasserman et al., 2020). Research on e-values has blossomed recently, without acknowledgment of understanding of its roots. The reader may see Ramdas et al. (2023) for a recent survey on game-theoretic statistics and safe anytime-valid inference, which provides a broader context in which e-values arise naturally, and for many examples of composite, nonparametric e-values, as well as details of the connection to betting scores and the wealth of a gambler betting against the null.

6.2 Combining multiple arbitrarily dependent e-values to test $\mathcal{P}$

Suppose we have constructed K arbitrarily dependent e-values $X_1, \ldots, X_K$ for the same null hypothesis $H_0 : P \in \mathcal{P}$. These may or may not be exchangeable. A natural way to form a test is to define

$$\bar{X}_K := (X_1 + \cdots + X_K)/K,$$

which is also an e-value, and thus to

$$(39) \qquad \text{reject } H_0 \text{ if } \bar{X}_K \geqslant 1/\alpha.$$

Said differently,

$$(40) \qquad p = \min(1/\bar{X}_K, 1)$$

is a p-value. In fact, Vovk and Wang (2021, Appendix G) prove that among symmetric e-to-p merging functions, $\min(1/\bar{X}_K, 1)$ is the optimal choice. However, the above optimality result precludes the use of randomization, meaning that it can potentially be dominated by randomized rules. We show that this is indeed the case. To prepare for the result, given a permutation π of $\{1, \ldots, K\}$, define

$$\bar{X}_t^\pi := \frac{1}{t} \sum_{i=1}^{t} X_{\pi(i)}.$$

Note that $\bar{X}_K^\pi = \bar{X}_K$.

PROPOSITION 6.1. *Let $X_1, \ldots, X_K$ be arbitrarily dependent e-values for the null hypothesis $H_0 : P \in \mathcal{P}$. If the X_i are not exchangeable, let π be a uniformly random permutation of $\{1, \ldots, K\}$, otherwise let π be the identity permutation. Then, the following rules control the type-I error at level α:*

$$(41) \qquad 1. \text{ reject } H_0 \text{ if } \bar{X}_K \geqslant U/\alpha,$$

$$(42) \qquad 2. \text{ reject } H_0 \text{ if } \sup_{t \leqslant K} \bar{X}_t^\pi \geqslant 1/\alpha,$$

$$(43) \qquad 3. \text{ reject } H_0 \text{ if } X_{\pi(1)} \geqslant U/\alpha \text{ or } \sup_{t \leqslant K} \bar{X}_t^\pi \geqslant 1/\alpha.$$

Further, each of these rules is more powerful than (39).

The proof follows directly from the UMI (Theorem 1.2) for (41), from the EMI (Theorem 1.1) for (42), and from EUMI (Theorem 5.3) for (43). Despite rule (42) being strictly less powerful than rule (43), we state it separately for ease of reference below. These three rules can be alternatively written as forming one of the following three p-values:

$$(44) \qquad 1. \ p_1 = \min(U/\bar{X}_K, 1)$$

$$(45) \qquad 2. \ p_2 = \min(\inf_{1 \leqslant t \leqslant K} (\bar{X}_t^\pi)^{-1}, 1)$$

$$(46) \qquad 3. \ p_3 = \min(\inf_{1 \leqslant t \leqslant K} (\bar{X}_t^\pi)^{-1}, U/X_{\pi(1)}, 1),$$

and rejecting the null when that p-value is smaller than α.

Note that if one wants to combine K arbitrarily dependent e-values to obtain a combined e-value, averaging is still optimal. But if one wants to form a p-value or test at level α, options like (44), (45), (46) dominate (40).

6.3 Combining m-way independent e-values

We now briefly extend the previous subsection's ideas beyond arbitrary dependence. Suppose that $X_1, \ldots, X_K$ are m-way independent e-values for some null H_0, meaning that every subset of m of them is jointly independent. $m = 1$ corresponds to arbitrary dependence as discussed previously, and $m = K$ corresponds to full joint independence.

A natural way to test using m-way independent[2] e-values is to use U-statistics and

$$\text{reject if } \bar{X}_K := \frac{1}{\binom{K}{m}} \sum_{\substack{A \subset \{1,\ldots,K\} \\ \text{s.t. } |A|=m}} \prod_{i \in A} X_i \geqslant 1/\alpha.$$

However, a better idea is to randomly sample (with or without replacement) a size-m subset $A_b \subset \{1, \ldots, K\}$ one-by-one for $b = 1, 2, \ldots$, calculate the average for the subsets sampled thus far, and reject as soon as any average exceeds $1/\alpha$:

$$\text{reject as soon as } \frac{1}{b} \sum_{s=1}^{b} \prod_{i \in A_s} X_i \geqslant 1/\alpha \text{ for some } b \geqslant 1.$$

This is a level-α test because of the EMI (2). Instead, we could also

$$\text{reject if } \bar{X}_K := \frac{1}{\binom{K}{m}} \sum_{\substack{A \subset \{1,\ldots,K\} \\ \text{s.t. } |A|=m}} \prod_{i \in A} X_i \geqslant U/\alpha,$$

for an independent uniform U. As before, while the latter two tests (often strictly) dominate the first one, it is a priori unclear which of the two options will be more powerful, so we explore this in the simulations later. Last, one can use the EUMI to combine the strengths of both the earlier rules, but we omit this for brevity.

7. IMPROVING THE POWER OF UNIVERSAL INFERENCE

Universal inference (Wasserman et al., 2020) is a simple and extremely broadly applicable test for any composite null hypothesis, that is nonasymptotically valid without regularity conditions. At its heart is a randomized

---

[2]Even under full independence, such combinations may be more robust and stable than the product $\prod_{i=1}^{K} X_i$. The latter will have the largest value if all e-values exceed one, but equals zero if even a single e-value was unluckily equal to zero. Of course there are ways to get around the zero issue, like calculating $\prod_{i=1}^{K} (X_i/2 + 1/2)$, but that is besides the current point.

method for constructing an e-value. To describe the simplest version of their idea, consider a setting where we have i.i.d. data $Y_1, \ldots, Y_n \sim P$ and we would like to test the null $H_0 : P \in \mathcal{P}$, perhaps against an alternative (implicit or explicit) $H_1 : P \in \mathcal{Q}$. We first partition the data at random into two (possibly unequal) datasets D_0 and D_1. Using D_1, we come up with any estimator/guess $\widehat{Q}$ in $\mathcal{Q}$. The split likelihood ratio is defined as

$$(47) \qquad X = \inf_{P \in \mathcal{P}} \prod_{i \in D_0} \frac{d\widehat{Q}}{dP}(Y_i),$$

where we assume for simplicity that $Q \ll P$ for any $P, Q \in \mathcal{P} \cup \mathcal{Q}$. In other words, it is the likelihood ratio of a particular alternative $\widehat{Q}$ (picked from D_1) against the maximum likelihood estimator under the null. Wasserman et al. (2020) prove that X is an e-value for $\mathcal{P}$, and they

$$(48) \qquad \text{reject } H_0 \text{ when } X \geqslant 1/\alpha.$$

There are many other variants, for example using profile likelihoods to handle nuisance parameters, smoothed likelihoods to avoid encountering an infinite likelihood, relaxed likelihoods in case calculating the maximum likelihood is infeasible, and so on. Universal inference is named such because it provides a simple and universally applicable baseline method that works for testing any null, without making any regularity assumptions (unlike the generalized likelihood ratio test, whose threshold is often unknown for singular or complex $\mathcal{P}$).

Of course, the downside is that the method is conservative in general. In parametric settings without nuisance parameters, when the usual likelihood ratio test applies, universal inference is typically loose (in asymptotic efficiency, say) by a small constant factor of about 2 to 4, achieving the right rate in sample size n, dimensionality and level α, and appropriate notions of signal-to-noise ratio (Dunn et al., 2023). In nonparametric settings, sometimes no other test exists, so the conservativeness of universal inference remains unclear (Dunn et al., 2024).

We mention two ways to gain back some of the conservativeness. The first is to simply replace (48) with the UMI (1.2) to yield "uniformly-randomized universal inference": let U be an independent $U[0, 1]$ random variable, then we may

$$(49) \qquad \text{reject } H_0 \text{ when } X \geqslant U/\alpha,$$

yielding a strictly more powerful test than universal inference, that still controls type-I error at level α.

The above use of uniform randomization is reminiscent of a somewhat similar use in the context of permutation or randomization tests, which are typically conservative by default, but can be made to be exact by the use of external randomization; see also Section 10.6 on avoiding the use of external randomization U.

The second way to gain back some of the constant factors is to recall the randomness associated with sample splitting, and derandomize the approach by averaging. One natural subsampling approach proposed by Wasserman et al. (2020) just repeats the calculation of X a total of B times, each time on a different independent random split of the data. Call the resulting exchangeable e-values as $X_1, \ldots, X_B$. They propose to

$$(50) \qquad \text{reject } H_0 \text{ when } \frac{X_1 + \cdots + X_B}{B} \geqslant 1/\alpha.$$

Dunn et al. (2023) prove that such derandomization does improve power. Given the discussion in preceding sections, using the UMI (5) to

$$(51) \qquad \text{reject } H_0 \text{ when } \frac{X_1 + \cdots + X_B}{B} \geqslant U/\alpha,$$

also yields a level-α test. Alternatively, one may

$$(52) \qquad \text{reject } H_0 \text{ when } \sup_{1 \leqslant b \leqslant B} \frac{X_1 + \cdots + X_b}{b} \geqslant 1/\alpha,$$

whose validity follows from the EMI (2).

Several of the above observations also apply to other related tests, for example using the reverse information projection e-value (Grünwald et al., 2024). We summarize the above observations below for easier reference.

PROPOSITION 7.1 ((De)randomized universal inference). *Let X be the split likelihood ratio statistic defined in (47) (or the crossfit likelihood ratio defined in Wasserman et al. (2020), or the reverse information projection e-value (Grünwald et al., 2024)). Consider the test that*

$$\text{rejects } H_0 \text{ when } X \geqslant U/\alpha.$$

Similarly, for subsampling-based universal inference, consider the rules (51) or (52), or a third rule that
(53)
rejects H_0 when either $X_1 \geqslant U/\alpha$ or $\sup_{1 \leqslant b \leqslant B} \dfrac{X_1 + \cdots + X_b}{b} \geqslant 1/\alpha$.

All the above rules are more powerful than universal inference (48), and control type-I error at level α without regularity conditions.

An important point worth remarking is that B has to be fixed in advance if we want to average the e-values to calculate a single e-value. However, in (52) or (53), *B does not have to be fixed in advance* since technically those suprema hold from 1 to ∞, not just 1 to B as stated, and thus they hold even if B is chosen adaptively (as a stopping time, say). In other words, instead of fixing B in advance, one can simply calculate one X_b at a time, calculate running averages as you go along, stop when you want, and reject if at any step the average crosses $1/\alpha$ (or, in the case of (53), stop as early as the first step if $X_1 \geqslant U/\alpha$).

8. (DE)RANDOMIZING NONPARAMETRIC BETTING-BASED TESTS

Consider a simple special case of a nonparametric testing problem from Waudby-Smith and Ramdas (2024). Let $Y_1, \ldots, Y_n$ be drawn i.i.d. from an unknown distribution P on $[0, 1]$, having mean μ. Suppose we want to test the null $H_0 : \mu = 0.5$, against an alternative $H_1 : \mu > 0.5$. Define the initial wealth of a gambler who wishes to bet against this null as $M_0 = 1$, and let their wealth evolve as

$$(54) \quad M_t = \prod_{i=1}^{t} (1 + \lambda_i(Y_i - 0.5)) = M_{t-1} \cdot (1 + \lambda_t(Y_t - 0.5)),$$

where $\lambda_i \in [0, 2]$ is a random variable (representing the gambler's bet) that can be chosen based on $Y_1, \ldots, Y_{i-1}$, meaning that it is "predictable" with respect to the filtration $\mathcal{F}_t := \sigma(Y_1, \ldots, Y_t)$. It is easy to check that under the null, $(M_t)_{t \geqslant 0}$ is a nonnegative martingale with initial value one (and in fact for each fixed t, M_t is an e-value). Ville's inequality (Section 4.1) implies that, under the null, $P\left(\sup_{0 \leqslant t \leqslant n} M_t \geqslant 1/\alpha\right) \leqslant \alpha$, and thus a level-$\alpha$ test is obtained by

$$(55) \qquad \text{rejecting the null if } \sup_{0 \leqslant t \leqslant n} M_t \geqslant 1/\alpha.$$

Said differently, $\inf_{0 \leqslant t \leqslant n}(1/M_t)$ is a p-value. The experiments in Waudby-Smith and Ramdas (2024) demonstrate that these tests perform excellently in practice (when using appropriate rules[3] for updating λ_i at each step), and the resulting confidence intervals obtained by inverting such a test are usually much shorter than a plethora of competing methods. We note immediately an improvement delivered by the randomized Ville inequality: we may

$$(56) \qquad \text{reject the null if } \sup_{0 \leqslant t \leqslant n} M_t \geqslant 1/\alpha \text{ or } M_n \geqslant U/\alpha,$$

for an independent uniform random variable U.

However, there is something slightly unsettling about this test: the p-value and test depend on the random order $Y_1, \ldots, Y_n$ of processing the points one by one, since the bets λ_t depend on $Y_1, \ldots, Y_{i-1}$. Since the chosen ordering was random to begin with, the p-value and test are symmetric functions of the data in a distributional sense (indeed, one can randomly scramble the data before running the test to enforce this), but there is some sense in which one may hope that the "algorithmic randomness" introduced by processing the data along *one* random ordering method can somehow be removed. In fact, this is also an

---

[3]One can show that if the alternative is true, then it is possible to bet smartly (meaning derive automated rules to predictably set λ_t) so that the gambler's wealth M_t grows exponentially fast, with the exponent automatically adapting to both the unknown signal $\mu - 0.5$, and the unknown variance $\mathbb{E}[(Y - \mu)^2]$. The authors also derive new exponential "empirical Bernstein" inequalities that can achieve the same effect.

issue with the oracle test: there is an optimal choice $\lambda^*(P)$ that could be used at every step, but the supremum over t still makes the resulting test or p-value dependent on the order of processing the bag of data.

Waudby-Smith and Ramdas (2024) describe one way to remove the effect of the arbitrary ordering. Noting that the final wealth M_n is an e-value (because of the nonnegative martingale property of the wealth process under the null), one can repeat the procedure B times on different independent random permutations of the original data, and only note the final wealth at time n on each permutation of the data, denoted $M_n^1, \ldots, M_n^B$. Then, they proposed to

$$(57) \qquad \text{reject the null if } \frac{M_n^1 + \cdots + M_n^B}{B} \geqslant 1/\alpha,$$

which for large enough B effectively becomes a symmetric function of the data.

Noting the parallels between the above rule and (50), one can instead gain more power by using (51) or (52) instead, meaning to either

$$(58) \qquad \text{reject the null if } \frac{M_n^1 + \cdots + M_n^B}{B} \geqslant U/\alpha,$$

or to

$$(59) \qquad \text{reject the null if } \sup_{1 \leqslant b \leqslant B} \frac{M_n^1 + \cdots + M_n^b}{b} \geqslant 1/\alpha.$$

The first method derandomizes by averaging (thus removing the effect of data ordering), but then again randomizes the threshold using the UMI. The second derandomizes in a more sophisticated manner, using the EMI. The following rule combines the two techniques
(60)

reject the null if $M_n^1 \geqslant U/\alpha$ or $\sup_{1 \leqslant b \leqslant B} \frac{M_n^1 + \cdots + M_n^b}{b} \geqslant 1/\alpha,$

using the EUMI. We formalize these observations below.

PROPOSITION 8.1. *Let* $Y_1, \ldots, Y_n$ *be independent random variables supported in [0,1], with identical mean* μ. *Let* $\pi_1, \ldots, \pi_B$ *be permutations of* $\{1, \ldots, n\}$ *that are sampled uniformly at random (with or without replacement), and let*

$$M_n^b = \prod_{i=1}^{t} (1 + \lambda_i(Y_{\pi_b(i)} - 0.5)), \quad b = 1, \ldots, B.$$

Then, the test (56) *for* $H_0 : \mu = 1/2$ *controls the type-I error at level* α, *and is more powerful than the original rule* (55). *Likewise, the tests* (58), (59) *and* (60) *control the type-I error and are more powerful than the rule* (57).

Since one cannot take a supremum over both t and B, it is a priori unclear which of the tests (55) and (59) is more powerful (or (56) versus (60)). We examine such questions in the simulations that follow.

We note that when inverting these tests to form confidence intervals for the mean μ, as done in Waudby-Smith and Ramdas (2024), the same U can be used across all the tests.

9. EXPERIMENTS

We perform a simulation study to illustrate the extent to which the UMI, EMI, and EUMI can increase the power of the aforementioned methodologies. Code for reproducing this simulation study is publicly available[4]. We choose the level $\alpha = .05$ across all simulations. The parameter B appearing in Sections 7–8 is always taken to be 100.

9.1 Confidence Intervals for a Gaussian Mean

We begin with a toy example to compare the tightness of our randomized tail bounds. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the $\mathcal{N}(0,1)$ distribution. We compare the width of the uniformly-randomized Hoeffding confidence interval (20) for the mean $\mathbb{E}[X_1]$, to that of the traditional Hoeffding interval (19). As a benchmark, we also compare them to the exact confidence interval $\bar{X}_n \pm z_{\alpha/2}/\sqrt{n}$, where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard Gaussian distribution. The average length and coverage of these three intervals is reported in Figure 1, across ten values of $n \in [100,2000]$.

It can be seen that the randomized Hoeffding interval (based on the UMI) has length lying between that of the traditional Hoeffding interval and the exact interval. By reasoning similarly as in Remark 3.2, the expected relative improvement in length of the exact interval over the UMI interval is approximately 17%. On the other hand, we have already stated that the relative improvement in length of the randomized Hoeffding bound over its classical counterpart is 14%. Both of these expected length ratios are confirmed by our simulation study, and show that the UMI interval sits roughly halfway between the Hoeffding and exact intervals, both in length and coverage. As discussed in Remark 3.3, the Hoeffding interval can itself be sharpened using more sophisticated tail bounds, and we expect that randomized versions of such inequalities would lead to even tighter intervals.

9.2 Testing with a set of arbitrarily dependent e-values

Assume that $X_1, \ldots, X_K$ are $K = 100$ test statistics which are not necessarily independent, and which are distributed as $\mathcal{N}(\mu, 1)$ for some $\mu \in \mathbb{R}$. We would like to combine them to test the null hypothesis $H_0 : \mu \leqslant 0$. One approach is to define the transformed statistics

$$(61) \qquad E_j = \exp(X_j - 0.5), \quad j = 1, \ldots, K,$$
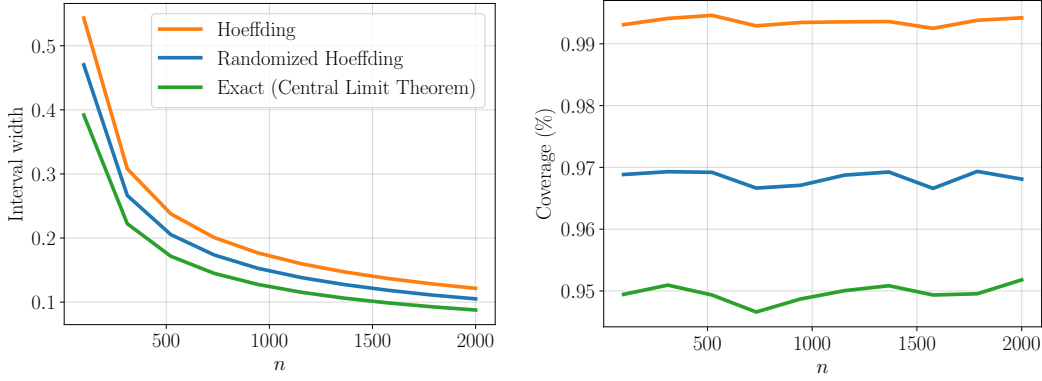

FIG 1. *Average length and coverage of the three confidence intervals across 20,000 replications for each sample size. The uniformly-randomized Hoeffding interval sits halfway between the traditional Hoeffding interval and the exact interval, both in terms of coverage and length.*

which are e-values for the null hypothesis H_0. We can combine them to test H_0 using the rule (39) based on averaging the e-values and applying MI ("Av+MI"), the rule (41) based on UMI, the rule (42) based on EMI, or the rule (43) based on on EUMI.

We simulate the power of these four approaches by drawing $(X_1, \ldots, X_K)$ with a Toeplitz-structured covariance matrix, where $\mathbf{Cov}[X_i, X_j] = \rho^{|i-j|}$ for all $1 \leqslant i, j \leqslant K$ and some $\rho \in [0,1]$. In Figure 2, we report their proportion of rejections across ten equally-spaced values of the true mean $\mu \in [0,4]$ and of the correlation parameter $\rho \in [0,1]$. The results are based on 500 replications from each model under consideration.

It can be seen that the tests based on the EMI, UMI, and EUMI improve upon the test based on the MI by at least 10% (in absolute power) for most combinations (μ, ρ). In some cases, the improvement in power is as high as 60% for the UMI and EUMI. The gain made by the UMI is similar to that of the EUMI, and both have similar behavior across all values of ρ. In contrast, the gains made by the EMI are most pronounced for small values of ρ. This is to be expected, since the cumulative averages of e-values in equation (42) are highly correlated when ρ is large; in fact, they are identical in the limit $\rho = 1$, in which case the rejection rules based on MI and EMI coincide.

We report the individual rejection proportions of all four methods in Figure D.1 of Appendix D. Therein, we also report simulation results under $K = 2$ rather than $K = 100$; see Figures D.2–D.3. When $K = 2$, it can be seen that the EUMI provides a more pronounced improvement over the UMI, with an absolute gain in power as high as 10% for several combinations (μ, ρ).

9.3 Universal Inference for model selection

We compare the methods presented in Section 7 for the problem of testing the number of components in a Gaussian mixture model. It is well-known that the parametric family of Gaussian mixtures does not satisfy the regularity conditions required for (twice the negative logarithm

of) the likelihood ratio statistic to admit its traditional χ^2 limiting distribution (Ghosh and Sen, 1984; Dacunha-Castelle and Gassiat, 1999; Chen and Li, 2009). In contrast, the method of universal inference based on the *split* likelihood ratio statistic, and its variants presented in Section 7, are valid without any regularity conditions, and are therefore natural candidates for this problem.

Although the limiting distribution of the likelihood ratio statistic is unknown or intractable for general Gaussian mixtures, it admits a simple expression when the underlying mixing proportions are known (Goffinet et al., 1992). We will assume this to be the case so that we can use the likelihood ratio test (LRT) as a benchmark, but we emphasize that the LRT cannot easily be used to derive a valid test for more general Gaussian mixtures, where the universal inference method would remain valid.

Let

$$(62) \quad X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} 0.25 \cdot \mathcal{N}(\mu_1, 1) + 0.75 \cdot \mathcal{N}(\mu_2, 1),$$

where the only unknown parameters are $\mu_1, \mu_2 \in \mathbb{R}$, and consider the problem of testing whether the above mixture has one vs. two components, i.e.

$$(63) \quad H_0 : \mu_1 = \mu_2, \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2.$$

By Theorem 1 of Goffinet et al. (1992), if λ denotes the likelihood-ratio statistic for these hypotheses, then $-2\log\lambda$ admits the limiting distribution $\max(0, Z)^2$, for $Z \sim \mathcal{N}(0,1)$, thus a valid level-α test for H_0 is to reject if $-2\log\lambda > q_{1-2\alpha}$, the latter being the $1 - 2\alpha$ quantile of the χ_1^2 distribution. We will refer to this as the LRT test. We compare its numerical performance to that of universal inference (UI; (48)), uniformly-randomized UI (UMI-UI; (49)), subsampling UI (SUI; (50)), uniformly-randomized SUI (UMI-SUI; (51)), exchangeable SUI (EMI-SUI; (52)), and exchangeable, uniformly-randomized SUI (EUMI-SUI; (53)).

Figure 3 reports the empirical power of these procedures based on 500 samples of size $n = 500$ from model (62), under ten equally-spaced values of $\mu :=$
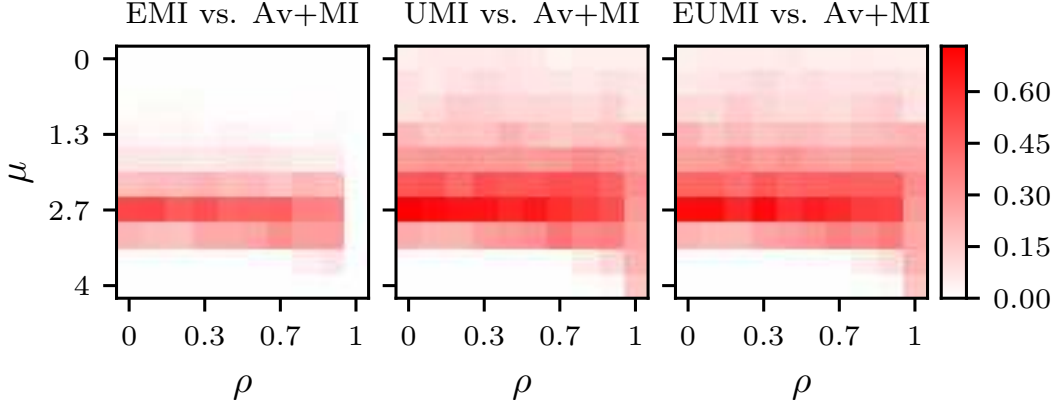
FIG 2. *Comparison of the rejection proportions π_{AvMI}, π_{EMI}, π_{UMI}, and π_{EUMI}. of the procedures* (39), (41), (42), *and* (43), *for rejecting the null hypothesis $H_0 : \mu \leq 0$ based on the e-values* (61). *For varying values of ρ and μ, the left-hand side plot represents the difference $\pi_{\text{EMI}} - \pi_{\text{AvMI}}$, the middle plot represents $\pi_{\text{UMI}} - \pi_{\text{AvMI}}$, and the right-hand side plot represents $\pi_{\text{EUMI}} - \pi_{\text{AvMI}}$. The procedures based on the UMI, EMI, and EUMI are strictly more powerful than the naive procedure based on MI, in some cases leading to an absolute increase in power of 60%.*

$-\mu_1 = \mu_2 \in [0, 1]$. We observe that the power of the UMI-UI method uniformly dominates that of the original UI method, and similarly, the UMI-SUI, EMI-SUI, and EUMI-SUI methods dominate their SUI counterpart. The methods UMI-SUI and EUMI-SUI exhibit similar performance, and their absolute increase in power compared to SUI is on the order of 15% for some values of μ. Although all methods based on the split LRT are markedly more conservative than the LRT, we recall that they can be used in arbitrary mixture models, while the LRT cannot.
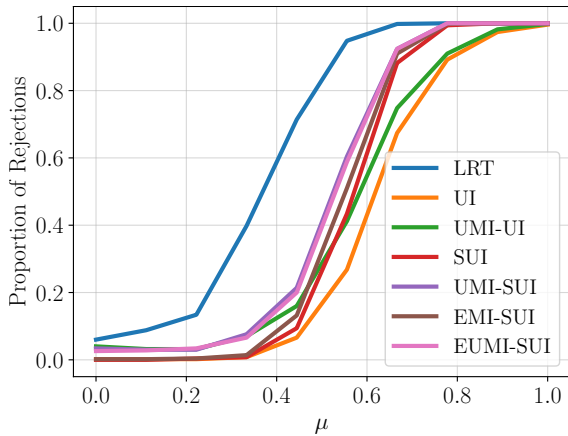

FIG 3. *Empirical power of the seven tests for the null hypothesis H_0 in equation* (63). *While the benchmark test provided by the LRT is most powerful, the variants of Universal Inference based on the EMI, UMI, and EUMI are more powerful than their MI counterparts.*

9.4 Testing the mean of a bounded random variable by betting

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a Beta(a, b) distribution, for some $a, b > 0$. Let $\mu = \mathbb{E}[X_1] = a/(a + b)$, and consider the problem of testing the null hypothesis

$H_0 : \mu = 1/2$ using the procedures defined in Section 8. We form the wealth statistic M_n in equation (54) based on a predictable sequence $(\lambda_i)_{i=1}^n$ chosen according to the LBOW betting strategy described in Waudby-Smith and Ramdas (2024), and compare the rejection rules in equations (55), (57), (58), (59),(60), based respectively on Ville's inequality, averaging followed by MI ("Av+MI"), UMI, EMI, and EUMI. We take $a = 20$, thus the null hypothesis reduces to $H_0 : b = 20$, and we compare the empirical power of these methods for varying values of $b \in [19, 20.8]$ and $n \in [100, 2000]$, based on 500 replications for each pair (n, b).
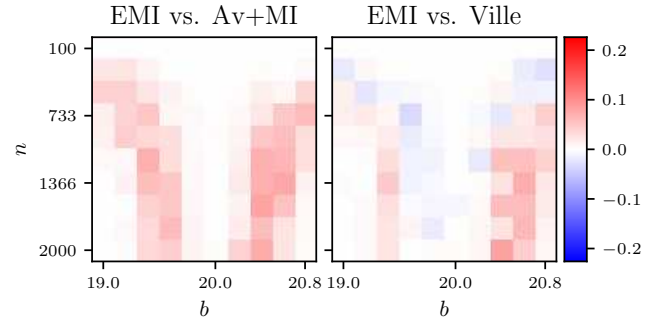

FIG 4. *Comparison of the rejection proportions π_{AvMI}, π_{Ville}, and π_{EMI} of the respective procedures* (57), (55) *and* (59), *for rejecting the null hypothesis $H_0 : b = 20$ based on the statistic M_n. For varying values of b and n, the left-hand side plot represents the difference $\pi_{\text{EMI}} - \pi_{\text{AvMI}}$, and the right-hand side plot represents the difference $\pi_{\text{EMI}} - \pi_{\text{Ville}}$. The procedure based on the EMI provides a modest improvement over that based on MI, but does not dominate the procedure based on Ville's inequality.*

Figure 4 compares the EMI-based procedure to those based on the MI and Ville's inequality. It can be seen that the EMI yields a modest improvement in power—on the order of 5%—compared to the MI, across the majority of

choices of b and n. In some cases, it yields an improvement of similar order over the procedure based on Ville's inequality, but does not uniformly dominate this method, as could have been anticipated from the discussion in Section 8. In contrast, in Figure 5, it can be seen that the procedure based on the UMI uniformly dominates both the MI and the procedure based on Ville's inequality, with a gain in absolute power as high as 20% in many cases. The performance of the EUMI-based procedure is nearly identical to that of the UMI, thus we defer this result to Figure D.4 of Appendix D.
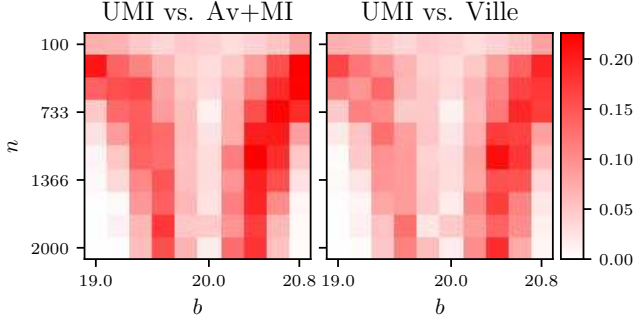

FIG 5. *Comparison of the rejection proportions* π_{AvMI}, π_{Ville}, *and* π_{UMI} *of the respective procedures* (57), (55) *and* (58), *for rejecting the null hypothesis* $H_0 : b = 20$ *based on the statistic* M_n. *For varying values of b and n, the left-hand side plot represents the difference* $\pi_{\mathrm{UMI}} - \pi_{\mathrm{AvMI}}$, *and the right-hand side plot represents the difference* $\pi_{\mathrm{UMI}} - \pi_{\mathrm{Ville}}$. *In contrast to the EMI-based procedure reported in Figure 4, the UMI-based procedure uniformly dominates both procedures based on MI and Ville's inequality.*

In Figure D.5 of Appendix D, we report the individual power of these five methods. In particular, it can be seen therein that the power of all methods is at most approximately 5% under the null hypothesis $b = 20$. It can further be seen that all methods achieve perfect (or nearly perfect) power for values of b near the boundaries of the interval $[19, 20.8]$, and for large values of n. This explains why the UMI, EMI, and EUMI do not to appear to provide a substantial improvement over the other methods in this regime.

10. DISCUSSION

This paper presented the uniformly-randomized and exchangeable Markov inequalities, along with extensions to Chebyshev and Chernoff bounds, and some example statistical applications involving universal inference and testing by betting. We now begin a relatively long discussion, mixing technical observations with some philosophical thoughts.

10.1 Markov's inequality as a derandomization of UMI

Let us begin by recalling the following implication of Markov's inequality from Section 6. For a nonnegative,

integrable X, we have that $X/\mathbb{E}[X]$ is an e-value, and thus $p^* := \mathbb{E}[X]/X$ is a p-value, meaning that

$$P(\mathbb{E}[X]/X \leqslant a) \leqslant a.$$

In this vein, UMI implies that $U\mathbb{E}[X]/X$ is also a valid p-value.

What if one tries to derandomize this statement? Suppose we draw B independent uniforms $U_1, \ldots, U_B$ and calculate B such p-values (where $p_b = U_b \mathbb{E}[X]/X$). Then, we get B dependent, exchangeable, p-values. An old result by Rüschendorf (1982) implies that twice the average of arbitrarily dependent p-values are also p-values, meaning that $\bar{p}_B := 2(p_1 + \cdots + p_B)/B$ is a p-value, and this factor of 2 cannot be improved in general (also see Vovk and Wang (2020)). Choi and Kim (2023) recently showed that the factor of 2 cannot be improved even assuming that the p-values are exchangeable (which is true in our setting).

Now note that as $B \to \infty$, $\bar{p}_B$ converges to p^*, because the uniforms average out to $1/2$. Thus the gain made by the uniform randomization is exactly offset by the factor of 2 lost by combining p-values. In other words, one can view Markov's inequality as a derandomized version of our UMI.

Further note that other forms of randomization do not appear to help. A result by Rüger (1978) shows that $\widetilde{p}_B := 2 \cdot \mathrm{median}(p_1, \ldots, p_B)$ is also a p-value and the factor of 2 cannot be improved (assume B is odd for simplicity). In fact, the same work also showed that $p_B^k := p_{(k)}B/k$ is a p-value for any fixed $k \leqslant B$, where $p_{(k)}$ is the k-th smallest p-value. Remarkably, $\widetilde{p}_B$ and p_B^k also converge to p^* as $B \to \infty$. This appears to be a perfectly-designed coincidence, but perhaps on more reflection a simple explanation of this phenomenon may be found. For now, it adds further justification to the title of this subsection.

10.2 On the role of external randomization in statistics

At a high level, there appear to be (at least) three reasons that external randomization is used in statistics:

1. **To save computation.** A classic example of this would be the permutation test. When applied to (say) a problem like two-sample or independence testing, the deterministic permutation test needs $n!$ permutations, where n is the number of data points. The variant that is typically used in practice, however, involves the permutations being uniformly sampled from the set of all permutations. This also results in a valid p-value, and the randomization is introduced solely to save computational effort. Another example is a risk-limiting election audit, where ballots are sampled in a random order, allowing one to possibly stop the audit early (with a guarantee on the error); without this randomization, one must look at every single ballot in the audit. A last example would be the use of stochastic

gradient descent in minimizing convex objectives because calculating a full gradient may be too expensive.

2. **To enable inference that is (essentially) impossible otherwise.** An example of this would be differential privacy; it is provably impossible to guarantee privacy without randomization (adding noise to summary statistics before releasing them). The same impossibility also arises in online learning against adversaries (e.g.: adversarial multi-armed bandits), and in calibrated probabilistic forecasting. Such examples also arise in the Monte Carlo literature, for example MCMC algorithms are used to enable sample from distributions that would otherwise be (analytically and computationally) intractable. The bootstrap and subsampling methods would also be examples. Universal inference also falls under this umbrella, since for many problems, we do not know of any other computational or analytical tool that could replace it. Another example is the knockoffs method for (fixed-X or model-X) conditional independence testing, and the related conditional randomization test. Nonexchangeable conformal prediction provides a contemporary case in point. Finally, sample splitting is used in a variety of contexts to enable assumption-lean inference (such as in post-selection inference).

3. **For more powerful inference.** An example of this would again be the permutation test. Whether used in its deterministic or random form, it produces a discrete p-value. Randomization can be introduced to convert it into a continuous p-value that is almost surely smaller than the discrete p-value (and exactly uniformly distributed under the null), thus improving (strictly) power.

The randomization used by UMI falls into the last category: the only purpose of introducing randomization is to improve power.

On a different note, there are at least two types of randomized procedures:

1. **Can be computationally derandomized.** For many procedures, expending more computation can render them "effectively deterministic" in the sense that some concentration of measure kicks in, so that the stochastic result concentrates around some deterministic limiting quantity. Examples include the bootstrap, subsampling, permutation and Monte Carlo methods (including MCMC), the conditional randomization test, (subsampling-based) universal inference, and stochastic gradient descent for convex optimization.

2. **Cannot be computationally derandomized.** This includes adversarial multi-armed bandits, probabilistic forecasting, differentially private inference, nonexchangeable conformal prediction, and many methods based on sample splitting.

UMI falls into the first category, but as discussed in the previous subsection, the power benefits of UMI vanish when it is derandomized, because it reduces exactly to Markov's inequality.

10.3 Frequency interpretation of the tests & confidence intervals

Chebyshev's inequality (10) implies that for any $\alpha \in (0, 1)$,

$$(64) \qquad \bar{X}_n \pm \frac{\sigma}{\sqrt{\alpha n}}$$

is a $(1 - \alpha)$ confidence interval for $\mathbb{E}X$. This means that when we construct infinitely many such intervals for different problems (with independent data from distributions with potentially different means and variances), at least 95% of those confidence intervals will cover the corresponding means. In contrast, our randomized Chebyshev's inequality (10) implies that

$$(65) \qquad \bar{X}_n \pm \frac{\sigma \sqrt{U}}{\sqrt{\alpha n}}$$

is also $(1 - \alpha)$ confidence interval for $\mathbb{E}X$. We highlight that it has *exactly* the same frequency interpretation as above. Despite being randomized and (almost surely) strictly tighter than (64), when we construct infinitely many such intervals for different problems (with independent data from distributions with different means and variances), at least 95% of those confidence intervals will cover the corresponding means.

Our e-value based tests (38), like our more powerful variant of universal inference (51), also have the same frequency interpretation as the nonrandomized tests (37). There is a simple way to interpret our use of randomization. When $\alpha = 0.05$, nonrandom thresholding rules reject the null when the e-value exceeds $1/\alpha = 20$. Meaning that if the e-value equals 10, we do not reject, while if it equals 20, we do. In our randomized setting, we simply view an e-value of 10 as having half the evidence as that of an e-value equaling 20, so we reject it with probability one half. Similarly, an e-value of 19.999 would not get rejected with the usual nonrandomized rules, but would get rejected with very high probability in our scheme. In other words, the rejection probability is exactly proportional to the required evidence for a definite rejection, resulting in a "smoothed" test, as opposed to a sharp 0-1 decision.

10.4 The potential for p-hacking, and ideas to overcome it

Despite the frequency interpretation discussed in the previous subsection, we recognize the potential risks for "p-hacking" (or hacking confidence intervals), where a naive or dishonest practitioner may reconstruct our uniformly randomized confidence interval many times (redrawing U), and pick the one that suits them (in order to report a narrow enough interval for their purposes), even reporting the random seed for "reproducibility". This of course is not valid. Thus our intervals must be employed with care. We suggest a few points that should be kept in mind for practical applications:

- If the interval constructions are coded up as part of some automated software that constructs thousands of such intervals (and perhaps acts on them) without any human involvement, then the above frequency interpretations will be preserved and (65) and (22) yield bona fide, valid confidence intervals.
- In Section 10.5, we point out that it may make sense to sometimes truncate the intervals, which increases interpretability, avoids contradicting intuition, and also reduces the extent to which p-hacking is possible.
- In Section 10.6, we point out that sometimes external randomization is not needed at all, and one can use randomness intrinsic in the data itself.
- Finally, we note that there may be opportunities to systematize the use of external randomness. One could create a central repository of uniform random numbers, and you can request a fixed number of them, but then have to use all of them. Independently, one could report a file with all the uniform random numbers used in an analysis, and these should pass a battery of uniformity tests (though this could itself be p-hacked, it is now a much higher bar).

10.5 Truncation to avoid empty (or tiny) confidence intervals

Since $\log(U)$ has its smallest possible value being $-\infty$, the interval (22) could sometimes simply be the empty interval. This will only happen to (much) less than an α fraction of constructed intervals, since the $(1 - \alpha)$ coverage property does still hold; indeed a direct calculation shows that an empty interval is constructed if and only if $\log U \leqslant -2\log(2/\alpha)$, which happens with probability at most $\alpha^2/4$. Rare as it may be, this phenomenon may not be very useful or intuitive in practice. Thus we suggest the following alternative:

$$(66) \quad \bar{X}_n \pm \left[\max\left(\sigma \sqrt{\frac{2\log(2/\alpha)}{n}} + \sigma \frac{\log(U)}{\sqrt{2n\log(2/\alpha)}} \, , \, \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} \right) \right],$$

where z_{1-a} is the right a-quantile of the standard Gaussian distribution. In short, whenever the randomized Hoeffding interval becomes smaller than the asymptotic interval based on the central limit theorem, we resort to reporting the latter. The interval in (66) is never shorter than what the CLT reports, almost surely shorter than the original Hoeffding interval, and is nonasymptotically valid.

Despite the fact that the Chebyshev interval is nonempty almost surely, we recognize that observing an extremely short interval, even if by chance due to randomization U, may also be troubling. One simple fix is to output the interval:

$$(67) \quad \bar{X}_n \pm \frac{\sigma \max(\sqrt{U}, 1/2)}{\sqrt{\alpha n}}$$

Of course, $1/2$ can be replaced by any other constant. The above interval is still nonasymptotically valid at level α, is almost surely tighter than Chebyshev's inequality (that is, the interval obtained from it) but never improves on it by more than a factor of 2. Indeed, the expected ratio of widths is $\mathbb{E}[\max(\sqrt{U}, 1/2)] = 17/24 \approx 0.71$, only a mild increase from the $2/3$ value obtained earlier.

In short, enlarging the interval by truncating the random improvement may be a suitable practical middle ground.

10.6 Using internal randomization in lieu of external randomization

Several of the bounds in this paper were formulated in terms of external randomization U. However, we note that in some situations, we can avoid the use of U entirely, while maintaining the gist of the original statements.

To elaborate, recall that many statistics in the paper, like the sample mean $\bar{X}_n$, are only functions of the order statistics of the data $\mathbf{X} := \{X_{(1)}, \ldots, X_{(n)}\}$ (equivalently, of the unordered bag of data, or of the empirical distribution). For a real x and finite set S of reals, define

$$\text{rank}(x; S) := \sum_{i \in S} 1(x_i \leqslant x)/|S|.$$

Now, note that if $X_1, \ldots, X_n$ are i.i.d. from a continuous univariate distribution, then

$$\mathbf{X} \perp \text{rank}(X_n; \mathbf{X}).$$

Further, the aforementioned rank is uniformly distributed on the discrete set $\{1/n, 2/n, \ldots, 1\}$. Thus, the rank stochastically dominates U and can therefore be used in its place in our earlier bounds. In essence, the leftover information in the data ordering, like the rank of X_n within the set $\mathbf{X}$, can be utilized "for free" without affecting the distribution of the underlying main statistic (like $\bar{X}_n$).

Thus, to use two examples whose expressions were recalled in the previous subsection, both

$$(68) \quad \bar{X}_n \pm \left[\sigma \sqrt{\frac{2\log(2/\alpha)}{n}} + \sigma \frac{\log(\text{rank}(X_n; \mathbf{X}))}{\sqrt{2n\log(2/\alpha)}} \right],$$

and

$$(69) \qquad \bar{X}_n \pm \frac{\sigma \sqrt{\text{rank}(X_n; \mathbf{X})}}{\sqrt{\alpha n}}$$

are valid $(1 - \alpha)$ confidence intervals for the mean (of a σ-subGaussian distribution, or a distribution with variance at most σ^2, respectively).

The same observation also applies to universal inference, because the likelihood calculation in the split likelihood ratio statistic is only a function of the unordered set $\{X_1, \ldots, X_m\}$ (where m is the size of the first split), and hence the rank of X_m within that set can be used in place of U in (49).

Perhaps such uses of the uniformly-randomized Markov's inequality, which are entirely "intrinsic" to the data itself, may be more palatable to those who have reasons to not prefer to use "extrinsic" randomization by using U.

10.7 Other potential applications

Despite our applications being focused on universal inference and "betting-based" inference, the use of the exchangeable Markov inequality was really enabled by two properties:

- The original underlying problem statement has a certain symmetry, for example the data are i.i.d. or exchangeable.
- The original method did not respect the above symmetry, by employing sample splitting, or by processing the data one at a time in a random order.

To be clear, there were some benefits to deviating from the original symmetric problem statement: in the case of universal inference, it enabled constructing a test without regularity conditions, and in the case of testing bounded means, it enabled a powerful test that was adaptive to the underlying unknown variance of the data. The loss of the problem symmetry could be regained by "algorithmic derandomization", that is repeating the same procedure and averaging the resulting e-values. It is in this latter step that the exchangeable Markov inequality kicks in and delivers more power to the final test.

The above bullet points apply to several other problems, for example Shekhar and Ramdas (2023) design betting-based tests for nonparametric two-sample testing, and the same techniques would apply to that problem as well. Similarly, Waudby-Smith and Ramdas (2024) derived the only known closed-form empirical Bernstein inequality that converges in width *exactly* to Bernstein's inequality, both of which can be improved with our uniform randomization technique, while the former can also be improved with the exchangeable Markov inequality. We omit the details for brevity.

The large improvements delivered by the uniformly-randomized Markov's inequality may be unsettling to some readers, which is why we presented a version that only uses the data itself for randomization in the previous section. We anticipate more applications and discussions about when such techniques may be appropriate (or not) to emerge with time.

11. SUMMARY

This paper revisited several standard inequalities (by Markov, Chebyshev, Chernoff, Ville, Doob) and proved that they can be improved by suitably employing a simple randomization technique involving a single independent uniform random variable, or by exploiting exchangeability of the underlying random variables (or both). Since these are standard building blocks for nonasymptotic inference, those inferential tools are thus improved. We provide several contemporary examples, including universal inference, betting-based concentration, and combining e-values, where we show large improvements in power. The preceding discussion attempts to address some concerns that may arise with the use of such techniques.

Acknowledgments

SUPPLEMENTARY MATERIAL

Additional Simulation Results
Additional simulation results deferred from Section 9.

Code for Simulation Study
Python code for reproducing the simulation study presented in Section 9.

REFERENCES

V. Bentkus. A Remark on Bernstein, Prokhorov, Bennett, Hoeffding, and Talagrand Inequalities. *Lithuanian Mathematical Journal*, 42 (3):262–269, 2002.

V. Bentkus, N. Kalosha, and M. Van Zuijlen. On domination of tail probabilities of (super) martingales: Explicit bounds. *Lithuanian Mathematical Journal*, 46(1):1–43, 2006.

F. P. Cantelli. Sui confini della probabilita. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 47–60, 1929.

J. Chen and P. Li. Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, pages 2523–2542, 2009.

W. Choi and I. Kim. Averaging p-values under exchangeability. *Statistics & Probability Letters*, 194:109748, 2023.

T. C. Christofides and R. J. Serfling. Maximal Inequalities for Multi-dimensionally Indexed Submartingale Arrays. *The Annals of Probability*, 18(2):630–641, 1990.

D. Dacunha-Castelle and E. Gassiat. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *The Annals of Statistics*, 27(4):1178–1209, 1999.

J. L. Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3): 455–486, 1940.

R. Dunn, A. Ramdas, S. Balakrishnan, and L. Wasserman. Gaussian universal likelihood ratio testing. *Biometrika*, 110(2):319–337, 2023.

R. Dunn, A. Gangrade, L. Wasserman, and A. Ramdas. Universal inference meets random projections: a scalable test for log-concavity. *Journal of Computational and Graphical Statistics*, pages 1–13, 2024.

R. Durrett. *Probability: Theory and Examples*, volume 49. Cambridge University Press, 2019.

B. Ghosh. Probability inequalities related to Markov's theorem. *The American Statistician*, 56(3):186–190, 2002.

J. K. Ghosh and P. K. Sen. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. Technical report, North Carolina State University. Dept. of Statistics, 1984.

B. Goffinet, P. Loisel, and B. Laurent. Testing in normal mixture models when the proportions are known. *Biometrika*, 79(4):842–846, 1992.

P. Grünwald, R. De Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society, Series B (with discussion)*, 2024.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301): 13–30, 1963.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *Annals of Statistics*, 49:1055–1080, 2021.

M. Huber. Halving the bounds for the Markov, Chebyshev, and Chernoff inequalities using smoothing. *The American Mathematical Monthly*, 126(10):915–927, 2019.

N. Ignatiadis, R. Wang, and A. Ramdas. E-values as unnormalized weights in multiple testing. *Biometrika*, 111(2):417–439, 2024.

A. J. Lee. *U-Statistics: Theory and Practice*. CRC Press, 1990.

T. Manole and A. Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 69(7):4641–4658, 2023.

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*, 2020.

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

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

B. Rüger. Das maximale signifikanzniveau des tests:"lehne h o ab, wenn k unter n gegebenen tests zur ablehnung führen". *Metrika*, 25:171–178, 1978.

L. Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

G. Shafer. Testing by betting: a strategy for statistical and scientific communication (with discussion and response). *Journal of the Royal Statistic Society A*, 184(2):407–478, 2021.

G. Shafer and V. Vovk. *Probability and finance: it's only a game!*, volume 491. John Wiley & Sons, 2005.

G. Shafer and V. Vovk. *Game-theoretic foundations for probability and finance*, volume 455. John Wiley & Sons, 2019.

S. Shekhar and A. Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70:1178–1203, 2023.

P. Tchebichef. Sur les valeurs limites des intégrales. *Journal de mathématiques pures et appliquées*, 19:157–160, 1874.

J. Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, 1939.

V. Vovk and R. Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

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

APPENDIX A: FURTHER DETAILS ON THE RELATIONSHIP TO Huber (2019)

We have seen that the additively randomized Markov's inequality in Proposition 1.4 is equivalent to the (multiplicatively) uniformly randomized Markov's inequality in Theorem 1.2, in the sense that they can be used to derive each other. We also saw that the result (7) by Huber (2019) makes a different claim, that is neither stronger nor weaker than Markov's inequality.

Despite the above facts, we can show that all three results are mathematically equivalent, meaning that they can all be used to derive each other. To see this, first let us recall (7) below for simplicity:

$$P(X + B \geqslant \epsilon) \leqslant \mathbb{E}[X]/(2\epsilon),$$

where $B \sim U[-\epsilon, \epsilon]$. Now, rewrite the left-hand side as

$$
\begin{aligned}
P(X \geqslant \epsilon - B) &= P(X \geqslant 2\epsilon - (\epsilon + B)) \\
&= P(X \geqslant 2\epsilon - 2A) = P(X/2 \geqslant \epsilon - A),
\end{aligned}
$$

(70)

where we used the fact that $\epsilon + B$ is distributed as $U[0, 2\epsilon]$, which has the same distribution as $2A$, where $A = U[0, \epsilon]$ was defined in Proposition 1.4.

We can either apply Huber's result to the left-hand side of (70) or Proposition 1.4 to the right-hand side of (70) to see that the two results imply each other. Nevertheless, the "take-home message" behind these inequalities is quite different. Indeed, as suggested by the title of his paper, Huber's focus is on halving the bounds of Markov's inequality with his two-sided additive randomization B (with implications for shape-constrained settings), while ours focus has been on improving Markov's inequality and the statistical implications of such an improvement.

APPENDIX B: PROOF OF THE TIME-REVERSED VILLE INEQUALITY (THEOREM 5.1)

In this appendix, we provide a detailed proof of the time-reversed Ville inequality (Theorem 5.1), which is at the heart of the proofs of the EMI (Theorem 1.1) and EUMI (Theorem 5.3). As previously mentioned, proofs of this result appear under varying assumptions in the works of Doob (1940), Lee (1990), Christofides and Ser-fling (1990), and Manole and Ramdas (2023). In what follows, we provide a new self-contained proof which easily lends itself to deriving the EUMI.

Let $m \geqslant 1$. Notice first that the process $(Y_t)_{t=1}^m$ defined by $Y_t = X_{m-t+1}$ is a forward submartingale with respect to the forward filtration $\mathcal{G}_t = \mathcal{E}_{m-t+1}$, $1 \leqslant t \leqslant m$. Indeed, (Y_t) is adapted to $(\mathcal{G}_t)$, and

$$\mathbb{E}[Y_{t+1}|\mathcal{G}_t] = \mathbb{E}[X_{m-t}|\mathcal{E}_{m-t+1}] \geqslant X_{m-t+1} = Y_t,$$

for all $t = 1, \ldots, m-1$. It must then also follow that $(Y_t)_{t=1}^m$ is a forward submartingale with respect to the filtration

$$\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t) = \sigma(X_{m-t+1}, \ldots, X_m), \qquad 1 \leqslant t \leqslant m.$$

With these preliminaries in place, we turn to proving the claimed inequality. Given $m \geqslant 1$, define $\tau := \sup\{1 \leqslant t \leqslant m : X_t \geqslant 1/a\}$, where $\sup \emptyset = -\infty$. Markov's inequality implies

(71) $\qquad P(\tau \geqslant 1) = P(X_{\tau \vee 1} \geqslant 1/a) \leqslant a \cdot \mathbb{E}[X_{\tau \vee 1}].$

Now define $\eta := m - \tau + 1$, so that $X_{\tau \vee 1} = Y_{\eta \wedge m}$. For all $t = 1, \ldots, m$, we have

$$\{m - \tau + 1 = t\} = \{\tau = m - t + 1\} \in \mathcal{F}_t$$

thus η and $\eta \wedge m$ are stopping times with respect to $(\mathcal{F}_t)$. Furthermore, the process $(Y_t)_{t=1}^m$ is trivially uniformly integrable, hence by Doob's optional stopping theorem for submartingales,

$$\mathbb{E}[X_{\tau \vee 1}] = \mathbb{E}[Y_{\eta \wedge m}] \leqslant \mathbb{E}[Y_m] = \mathbb{E}[X_1].$$

Returning to equation (71) and noting that $\{\tau \geqslant 1\} = \{\sup_{1 \leqslant t \leqslant m} X_t \geqslant 1/a\}$, we have thus shown

$$P\left(\sup_{1 \leqslant t \leqslant m} X_t \geqslant 1/a\right) \leqslant a \cdot \mathbb{E}[X_1].$$

By the bounded convergence theorem, sending $m \to \infty$ yields our claim. $\qquad \square$

APPENDIX C: A GENERAL RANDOMIZED TAIL BOUND

We state and prove a simple randomized tail bound which contains Theorems 1.2, 2.1 and 3.1 as special cases, and can be used to derive randomized variants of other tail bounds in the literature.

PROPOSITION C.1. *Let X be a random variable taking values in a set $\mathcal{X} \subseteq \mathbb{R}$, and let $I \subseteq \mathbb{R}_+$ be an interval. Let $f : \mathcal{X} \to I$ and $g : I \to \mathcal{X}$ be nondecreasing Borel-measurable functions such that $f(g(z)) \geqslant z$ for any $z \in I$. Then, given a random variable $U \sim \text{Unif}(0,1)$ independent of X, and $x > 0$, it holds that*

$$\mathbb{P}(X \geqslant g(Uf(x))) \leqslant \frac{\mathbb{E}[f(X)]}{f(x)}.$$

The proof is exactly as before. Notice first that $Uf(x)$ takes values in I since U is supported in $[0, 1]$, thus the quantity $g(Uf(x))$ is well-defined. Furthermore,

$$P(X \geqslant g(Uf(x))) = P(f(X) \geqslant f(g(Uf(x))))$$
$$\leqslant P(f(X) \geqslant Uf(x))$$
$$= \mathbb{E}[P(U \leqslant f(X)/f(x)|X)] \leqslant \mathbb{E}f(X)/f(x),$$

which proves the claim. $\qquad \square$

As an example, we next use Proposition C.1 to derive uniformly-randomized analogues of Cantelli's inequality, Bernstein's inequality, and of the empirical-Bernstein inequality.

C.1 Uniformly-randomized Cantelli inequality

We begin by deriving a uniformly-randomized analogue of Cantelli's inequality (Cantelli, 1929), which is a one-sided version of Chebyshev's inequality[5]. In contrast to (10), it states that

(72) $\qquad P(X - \mathbb{E}X \geqslant k\sigma) \leqslant \frac{1}{k^2 + 1}.$

It can be improved by uniform randomization as follows:

(73) $\qquad P(X - \mathbb{E}X \geqslant \sqrt{U}(k\sigma + \sigma/k) - \sigma/k) \leqslant \frac{1}{k^2 + 1},$

which we call the *uniformly-randomized Cantelli inequality*. To see that (73) is a stronger statement than (72), rewrite the left-hand side as

$$P(X - \mathbb{E}X \geqslant k\sigma - (1 - \sqrt{U})(k\sigma + \sigma/k)),$$

and note that $(1 - \sqrt{U})$ is positive.

The proof of (73) is a simple consequence of Proposition C.1. Let $x = \sigma k$ and $u = \sigma/k$. Taking $f(y) = (y + u)^2$ for all $y \in \mathbb{R}$ and and $g(z) = \sqrt{z} - 1$ for all $z \in \mathbb{R}_+$, we have

$$P\left(X - \mathbb{E}X \geqslant \sqrt{U}(k\sigma + \sigma/k) - \sigma/k\right)$$
$$= P(X - \mathbb{E}X \geqslant g(Uf(x)))$$
$$\leqslant \frac{\mathbb{E}[f(X - \mathbb{E}X)]}{f(x)}$$
$$= \frac{\sigma^2 + u^2}{(x + u)^2} = \frac{1 + 1/k^2}{2 + k^2 + 1/k^2} = \frac{1}{1 + k^2},$$

---

[5]Despite its name, Cantelli's inequality apparently originated in Chebyshev's much earlier work (Tchebichef, 1874); see Ghosh (2002). In fact, it is commonly accepted that Markov's inequality itself had already been proven by Chebyshev, who was Markov's doctoral advisor.

as claimed.

C.2 Uniformly-randomized Bernstein inequality

We next derive a uniformly-randomized analogue of the classical Bernstein inequality (see for instance Proposition 2.10 of Wainwright (2019)). We will assume that the random variable X satisfies the Bernstein condition, namely

$$(74) \qquad \left| \mathbb{E}(X - \mathbb{E}X)^k \right| \leq \frac{1}{2} k! \sigma^2 b^{k-2}, \quad \text{for } k = 2, 3, \dots$$

for some $\sigma, b > 0$. In particular, if X is a bounded random variable with variance at most σ^2 and satisfying $|X - \mathbb{E}X| \leq 1$, then the above condition holds for $b = 1/3$. Bernstein's inequality states that, under condition (74), for all $\epsilon > 0$,

$$(75) \qquad P(X - \mathbb{E}X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2(\sigma^2 + \epsilon b)} \right).$$

Setting the right-hand side equal to $\alpha \in (0, 1)$, the above implies

$$P\left(X - \mathbb{E}X \geq \sqrt{2\sigma^2 \log(1/\alpha)} + 2b\log(1/\alpha) \right) \leq \alpha.$$

Consequently, given n i.i.d. samples of X, the following is a $(1 - \alpha)$-confidence interval for $\mathbb{E}X$:

$$(76) \qquad \bar{X}_n \pm \left[\sqrt{\frac{2\sigma^2 \log(2/\alpha)}{n}} + \frac{2b\log(2/\alpha)}{n} \right].$$

It is well known that the i.i.d. assumption can be relaxed into a martingale dependence assumption requiring neither the independence aspect nor the identically distributed aspect, but we omit this generalization for simplicity.

Our *uniformly-randomized Bernstein inequality* reads

$$(77)$$
$$P\Big(X - \mathbb{E}X \geq \Big(\sigma^2 + b\sqrt{\sigma^2 \log(1/\alpha)} 2b^2 \log(1/\alpha) \Big) \log U$$
$$+ \sqrt{2\sigma^2 \log(1/\alpha)} + 2b\log(1/\alpha) \Big) \leq \alpha,$$

where U is a random variable that is independent of X and (stochastically larger than) uniform on $[0, 1]$. Since $\log U < 0$ almost surely, the above inequality provides a strict and almost sure improvement of Bernstein's inequality (and it recovers Bernstein's inequality by substituting $U = 1$). Recall also that $\mathbb{E}[\log U] = -1$ and **Var**$[\log U] = 1$, giving an idea of the extent of the improvement.

To prove (77), let $|\lambda| \leq b^{-1}$, $f(x) = \exp(\lambda x)$ for all $x \in \mathbb{R}$, and $g(z) = \log z/\lambda$ for all $z > 0$. By Proposition C.1, we have for all $x > 0$,

$$P\left(X - \mu \geq \frac{\log U}{\lambda} + x \right) \leq e^{-\lambda x} \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}]$$

$$\leq \exp\left\{ \frac{\lambda^2 \sigma^2}{2(1 - b|\lambda|)} - \lambda x \right\},$$

where we used the fact that $\mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \leq \exp(\lambda^2 \sigma^2 / 2(1 - b|\lambda|))$ for all $|\lambda| \leq b^{-1}$ under the Bernstein condition (cf. Proposition 2.10 of Wainwright (2019)). Now, letting $\lambda = (bx + \sigma^2)^{-1}$ and simplifying the above expression, we obtain

$$(78)$$
$$P\left(X - \mu \geq (bx + \sigma^2) \log U + x \right) \leq \exp\left\{ -\frac{x^2}{2(\sigma^2 + bx)} \right\},$$

which can be viewed as another form of our randomized Bernstein inequality. Setting $x = \sqrt{2\sigma^2 \log(1/\alpha)} + 2b\log(1/\alpha)$ leads to the claimed inequality.

Combining the above with Lieb's inequality, one directly obtains a randomized matrix-Bernstein inequality as well; we omit the details for brevity.

C.3 Uniformly-randomized empirical Bernstein inequality

Bernstein's inequality is not always practically applicable due to the need to know σ. When the data are bounded, one can construct so-called empirical Bernstein (EB) inequalities that only depend on the data. There are several such EB inequalities in the literature, but we present below a randomized variant of a recent EB inequality by Waudby-Smith and Ramdas (2024), because it is the only one that we are aware of whose corresponding confidence interval width exactly matches the first order term in (76).

Going forward, suppose that the i.i.d. data lie in $[0, 1]$; this is done for simplicity and without loss of generality. Define

$$\psi(\lambda) := (-\log(1 - \lambda) - \lambda) \quad \text{for } \lambda \in [0, 1),$$

and the instantaneous empirical variance as

$$v_t := (X_t - \widehat{\mu}_{t-1})^2, \quad \text{where } \widehat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t + 1}$$

The following is then a $(1 - \alpha)$-CI for $\mathbb{E}X$:

$$(79) \qquad \frac{\sum_{t=1}^n \lambda_t X_t}{\sum_{t=1}^n \lambda_t} \pm \left[\frac{\log(2/\alpha) + \sum_{t=1}^n v_t \psi(\lambda_t)}{\sum_{t=1}^n \lambda_t} \right],$$

where $\lambda_t \in (0, 1)$ is a function of $X_1, \dots, X_{t-1}$ that is set as follows:

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

The proof follows by observing that

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

is a nonnegative supermartingale with initial value $M_0 = 1$, and thus M_n is an e-value. Applying Markov's inequality to M_n, and rearranging, yields one side of (79), and a union bound with $-\lambda_i$ in place of λ_i yields the other side.

Instead, applying UMI in place of Markov's inequality, we obtain the uniformly-randomized empirical Bernstein confidence interval:

$$(80) \quad \frac{\sum_{t=1}^n \lambda_t X_t}{\sum_{t=1}^n \lambda_t} \pm \left[\frac{\log(2/\alpha) + \log U + \sum_{t=1}^n v_t \psi(\lambda_t)}{\sum_{t=1}^n \lambda_t} \right],$$

which is almost surely tighter than (79) since $\log U < 0$ with probability one.

REMARK C.2. *Denoting the expression in (79) by C_n, it turns out that $\bigcap_{i \leqslant n} C_i$ is also a valid $(1 - \alpha)$-confidence interval. This is obtained by applying Ville's inequality to the supermartingale M, in place of Markov's inequality. It may be a priori unclear whether $\bigcap_{i \leqslant n} C_i$ is tighter than (80) or not, but Figure 5 suggests a clear win for UMI over Ville.*