Correlated Binomial Process

Moïse Blanchard MOISEB@MIT.EDU
Massachusetts Institute of Technology

Doron Cohen DORONV@POST.BGU.AC.IL
Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Ben-Gurion University of the Negev

Editors: Shipra Agrawal and Aaron Roth

Abstract

Cohen and Kontorovich (COLT 2023) initiated the study of what we call here the Binomial Empirical Process: the maximal empirical mean deviation for sequences of binary random variables (up to rescaling, the empirical mean of each entry of the random sequence is a binomial hence the naming). They almost fully analyzed the case where the binomials are independent, which corresponds to all random variable entries from the sequence being independent. The remaining gap was closed by Blanchard and Voráček (ALT 2024). In this work, we study the much more general and challenging case with correlations. In contradistinction to Gaussian processes, whose behavior is characterized by the covariance structure, we discover that, at least somewhat surprisingly, for binomial processes covariance does not even characterize convergence. Although a full characterization remains out of reach, we take the first steps with nontrivial upper and lower bounds in terms of covering numbers.

Keywords: empirical process; subgaussian; concentration; high dimension; convergence

1. Introduction

We consider the fundamental problem of estimating the mean of a high-dimensional and potentially infinite-dimensional distribution μ from independent and identically distributed (iid) samples $X^{(1)}, \dots, X^{(n)} \sim \mu$. More precisely, we aim to characterize the maximum deviation from the empirical mean $\hat{p} := n^{-1} \sum_{i=1}^{n} X^{(i)}$ to the true mean $p = \mathbb{E}X^{(1)}$. We mainly focus on distributions with support on the hypercube, that is, for $X = (X_j)_{j \geq 1} \sim \mu$ each entry X_j is a Bernoulli random variable — as these already capture most of the relevant high-dimensional phenomena. The object of interest is the expected maximum deviation:

$$\Delta_n(\mu) \quad := \quad \mathbb{E} \left\| \hat{p} - p \right\|_\infty, \tag{1}$$

and we aim to study the behavior of $\Delta_n(\mu)$ as the number of samples n grows.

Estimating the mean of a distribution μ on $\mathbb{R}^d$ from independent draws is among the most basic problems of statistics. Much of the earlier theory has focused on obtaining efficient estimators $\hat{m}_n$ of the true mean m and analyzing the decay of $\|\hat{m}_n - m\|_2$ as a function of sample size n, dimension d, and various moment assumptions on X (Catoni, 2012; Devroye et al., 2016; Lugosi and Mendelson, 2019a,b; Cherapanamjeri et al., 2019, 2020; Diakonikolas et al., 2020; Hopkins, 2020; Lugosi and Mendelson, 2021; Lee and Valiant, 2022). For d-dimensional distributions μ on $\{0,1\}^d$, the classical Glivenko-Cantelli theorem ensures the convergence of the maximum deviation $\Delta_n(\mu)$ to 0, which can be quantified by the Dvoretsky-Kiefer-Wolfowitz inequality via

$\Delta_n(\mu) \lesssim \sqrt{\ln(d+1)/n}$. This worst-case bound can however be overly pessimistic. In particular, when the dimension is infinite the bound becomes vacuous: distributions-dependent bounds are necessary. Inspired by these observations and by Thomas (2018), Cohen and Kontorovich (2023a) introduced the problem of characterizing the distribution-dependent behavior of the empirical deviation $\Delta_n(\mu) = \mathbb{E}\,\|\hat{p} - p\|_\infty$ for distributions $\mu \in \{0,1\}^{\mathbb{N}}$. The ℓ_∞ norm is in some sense the most interesting of the ℓ_r norms; indeed, for $r < \infty$, $\Delta_n^{(r)} := \mathbb{E}\,\|\hat{p} - p\|_r^r$ decomposes into a sum of expectations and the condition $\Delta_n^{(r)} \to 0$ reduces to one of convergence of the appropriate series. The uniform convergence implied by the $\|\cdot\|_\infty$ norm as well as the dependence on the particular distribution μ led the authors to refer to this problem as *local Glivenko-Cantelli*.

Cohen and Kontorovich (2023a) obtained an almost complete understanding of the behavior of $\Delta_n(\mu)$ in the case of product measures, that is, for $X = (X_j)_{j\geq 1} \sim \mu$ all entries X_j are independent. In this case, μ is determined entirely by its mean p. To emphasize that these are product measures, we write $\Delta_n(p)$ for $\Delta_n(\mu)$. Restricting $p \in [0,1]^{\mathbb{N}}$ to the range $[0, \frac{1}{2}]$ and requiring that $p_j \downarrow 0$ as $j \to \infty$ (which incurs no loss of generality, as shown ibid.) they defined the functional

$$T(p) \;:=\; \sup_{j\in\mathbb{N}} \frac{\log(j+1)}{\log(1/p_j)}$$

and showed that $\Delta_n(p) \to 0$ iff $T(p) < \infty$. They also characterized up to constants the asymptotic convergence of $\Delta_n(p)$ when $T(p) < \infty$ via the functional

$$S(p) \;:=\; \sup_{j\in\mathbb{N}} p_j \log(j+1),$$

establishing that $\Delta_n(p)$ decays as $\sqrt{S(p)/n}$. Additional finite-sample bounds provided therein were tightened by Blanchard and Voráček (2024) as follows:

$$\Delta_n(p) \asymp 1 \wedge \left(\sqrt{\frac{S(p)}{n}} + \sup_{j\geq 1} \frac{\log(j+1)}{n \log\left(2 + \frac{\log(j+1)}{np_j}\right)} \right), \qquad \text{if } n \cdot \sup_{j\geq 1} 2jp_j > 1, \tag{2}$$

$$\Delta_n(p) \asymp \frac{1}{n} \wedge \sum_{j\geq 1} p_j, \qquad \text{otherwise.} \tag{3}$$

In this paper, we tackle the much more challenging case of *general* distributions μ on $\{0,1\}^{\mathbb{N}}$: entries of $X \sim \mu$ may have arbitrary correlations. While the upper bounds from Eq (2) and (3) mostly hold up to minor changes (Blanchard and Voráček, 2024), these are in general very loose since they do not account for possible correlations between entries of $X \sim \mu$. One can easily find examples for which these bounds are vacuous but $\Delta_n(\mu)$ still converges to 0 as n grows. Our goal is to understand for which forms of correlations in μ the quantity $\Delta_n(\mu)$ decays as n grows, and hence the empirical mean is an accurate estimator of the mean.

Our contributions. The decoupling result in Theorem 1 shows that when the pairwise correlations are negative, the behavior of $\Delta_n(\mu)$ remains as in the independent case, up to universal constants. This might lead one to conjecture that the pairwise correlations suffice to characterize the decay of Δ_n. Theorem 3 decisively shatters this conjecture, by exhibiting two processes μ, ν with identical pairwise covariances but for which $\Delta_n(\mu)$ decays to 0 while $\Delta_n(\nu)$ does not. While characterizing

the measures μ for which $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$ remains a challenging open problem, we take nontrivial first steps in this direction. In particular, we define two metrics and show how covering numbers with respect to these provide upper and lower bounds on $\Delta_n(\mu)$. More specifically, we give necessary and sufficient conditions on these covering numbers for $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$ to hold and show that among the covering-number-based bounds, this is essentially the best one can do. Some of the techniques developed for the results may be of independent interest, including decoupling (Proposition 10), subgaussian increments (Lemma 18), a recursive argument to extract entries with large deviations (Theorem 4), and probability distributions with tree structures (Proposition 8).

Terminology and related works. Because we focus on distributions μ on $\{0,1\}^{\mathbb{N}}$, the rescaled empirical mean $Y = n\hat{p}$ is a vector of binomial random variables $Y_j \sim \text{Binomial}(n, p_j)$. When the distribution μ is a product measure, Y is exactly the sequence of *independent* binomials $Y_j \sim \text{Binomial}(n, p_j)$; however, for general distributions μ, these binomials are *correlated*. With this perspective, the empirical deviation of the mean $\bar{Y} := \hat{p} - p = n^{-1}Y - p$ is a centered, normalized process, which we refer to as the *binomial empirical process*. The quantity $\Delta_n(\mu) = \mathbb{E}\sup_{j\in\mathbb{N}} |\tilde{Y}_j|$ is then simply the expected uniform absolute deviation of the binomial process Y.

A few remarks are in order. First, the binomial empirical process Y does not have an arbitrary structure since its entries are coupled via the joint distribution μ on $\{0,1\}^{\mathbb{N}}$: we have $Y = \hat{p} - p$, where p is the true mean and $\hat{p}$ is the empirical mean based on n iid copies of μ.

Second, the study of the binomial empirical process differs from that of *Bernoulli processes* as studied by Talagrand and others. We refer the reader to Chapter 5 of Talagrand (2014). Briefly, one defines a sequence of mutually independent symmetric Bernoulli variables ε_i with $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$. To any $T \subset \ell^2(\mathbb{N})$ we assign the value

$$b(T) = \mathbb{E}\sup_{t\in T} \sum_{i\in\mathbb{N}} t_i \varepsilon_i. \tag{4}$$

The celebrated Bednorz-Latała Theorem (Bednorz and Latała, 2014) characterizes $b(T)$ up to universal constants in terms of the combinatorial structure of T involving Gaussian width and ℓ_1 diameter. In principle, one may be able to relate the two objects of interest, $b(T)$ and $\Delta_n(\mu)$, via some mapping of μ and n to T and vice versa. Despite repeated attempts, we were not successful in doing this; in fact, we were not able to extract any information from the Bednorz-Latała Theorem that would shed light on either of the two open problems above.

Notation. The measure-theoretic subtleties of defining distributions on $\{0,1\}^{\mathbb{N}}$ are addressed in Cohen and Kontorovich (2023a). If μ is a probability measure on $\{0,1\}^{\mathbb{N}}$, we say that $\tilde{\mu}$ is its *product version* if $\tilde{\mu}$ is a product measure on $\{0,1\}^{\mathbb{N}}$ that agrees with μ on all of the marginals, that is, if $X \sim \mu$ and $\tilde{X} \sim \tilde{\mu}$, then the entries X_j and $\tilde{X}_j$ are equal in distribution and the $\{\tilde{X}_j : j \in \mathbb{N}\}$ are mutually independent. In this case, we say that $\tilde{X}$ is the *independent version* of X. A probability measure μ on $\{0,1\}^{\mathbb{N}}$ induces the metrics ξ and ρ on $\mathbb{N}$ as follows. Putting

$$p_i := \mathop{\mathbb{E}}_{X\sim\mu}[X_i], \qquad r_{ij} := \mathop{\mathbb{E}}_{X\sim\mu}[X_i X_j], \qquad i,j \in \mathbb{N}, \tag{5}$$

we define

$$\xi(i,j) := \mathbb{P}(X_i \neq X_j) = p_i + p_j - 2r_{ij}, \tag{6}$$

$$\rho(i,j) := \frac{2}{\sqrt{3}} \wedge \sqrt{\frac{2}{\log \frac{2}{\xi(i,j)}}}, \tag{7}$$

where we set $\rho(i,j) = 0$ whenever $\xi(i,j) = 0$. It is straightforward to verify that both are metrics; for the former, we have $\xi(i,j) = \mathbb{E}\,|X_i - X_j|$ and the latter is verified as such in Lemma 19.

For any two quantities $a, b \in (0,\infty)$, we use the shorthand $a \lesssim b$ when there exists a universal constant $C > 0$ (independent of any parameters of the problem) such that $a \le Cb$. Likewise, $a \gtrsim b$ if $b \lesssim a$ and $a \asymp b$ if both $a \lesssim b$ and $a \gtrsim b$ hold. The floor and ceiling functions, $\lfloor t \rfloor$, $\lceil t \rceil$, map $t \in \mathbb{R}$ to its closest integers below and above, respectively; also, $s \vee t := \max\{s,t\}$, $s \wedge t := \min\{s,t\}$, $[s]_+ := 0 \vee s$ and $[s]_- := 0 \vee (-s)$. Unspecified constants such as c, c' may change value from line to line. We use superscripts to denote distinct random vectors and subscripts to denote indices within a given vector. Thus, if $X^{(1)}, , \ldots, X^{(n)}$ are independent copies of X, then $X_j^{(i)}$ denotes the jth entry of the ith copy.

2. Main results

Our first result is a decoupling inequality comparing a negatively correlated binomial process to its independent version. If μ is a probability measure on $\{0,1\}^{\mathbb{N}}$ and $\tilde{\mu}$ its *product version* as defined above (the two agree on the marginals and $\tilde{\mu}$ is a product measure), then "decoupling from above", i.e., $\Delta_n(\mu) \lesssim \Delta_n(\tilde{\mu})$, holds without any structural assumptions on μ and the proof is quite straightforward (Chollete et al., 2023, Proposition 3.1). The other direction is far less trivial and is clearly not true in general:

Theorem 1 (Decoupling from below) *Let μ be a probability measure on $\{0,1\}^{\mathbb{N}}$ with negatively correlated coordinates (i.e., $X \sim \mu$ verifies $\mathbb{E}[X_i X_j] \le \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ for $i, j \in \mathbb{N}$) and $\tilde{\mu}$ its* product version. *Then*

$$\Delta_n(\mu) \;\ge\; \frac{1}{4}\Delta_n(\tilde{\mu}), \qquad n \ge 1.$$

The proof is given in Section 3. In particular, together with the tight bounds Eq (2) and (3) from Blanchard and Voráček (2024) for independent Bernoulli sequences, we directly obtain tight non-asymptotic bounds for negatively correlated Bernoulli sequences.

Corollary 2 *Let μ be a probability measure on $\{0,1\}^{\mathbb{N}}$ with negatively correlated coordinates. Then, Eq (2) and (3) continue to hold.*

In the previous result, only the lower bounds are worsened by a factor $\frac{1}{4}$ compared to those for independent Bernoulli sequences in Eq (2) and (3). It is worth noting that the upper bounds are exactly the same as for the independent case—these only require union bounds and Markov's inequality, hence also hold for general distributions, as noted in Blanchard and Voráček (2024, Corollary 3).

Based on Corollary 2 as well as Gaussian process theory, one might plausibly conjecture that the behavior of $\Delta_n(\mu)$ is determined by the covariance structure of $X \sim \mu$. It is therefore at least somewhat surprising that this is very much not the case:

Theorem 3 (Covariance does not characterize Δ_n) *There exist probability measures μ, ν on $\{0,1\}^{\mathbb{N}}$ that agree on their covariance matrices,*

$$\mathop{\mathbb{E}}_{X\sim\mu}[X_i X_j] \;=\; \mathop{\mathbb{E}}_{X\sim\nu}[X_i X_j], \qquad i, j \in \mathbb{N},$$

while $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$ and $\Delta_n(\nu) \underset{n\to\infty}{\longrightarrow} \frac{1}{2}$.

The proof is given in Section 4. It turns that the two distributions we construct also share the exact same third-order moments; that is,

$$\mathop{\mathbb{E}}_{X\sim\mu}[X_iX_jX_k] \quad = \quad \mathop{\mathbb{E}}_{X\sim\nu}[X_iX_jX_k], \qquad i,j,k \in \mathbb{N}.$$

In particular, this shows that third-order moments still fail to characterize the convergence of Δ_n to 0 (the proof of this claim appears in Appendix A). We conjecture that similar examples could be constructed to show that for any $k \geq 1$, knowledge of the kth order moments does not suffice to characterize the decay of Δ_n.

Although a complete understanding of the behavior of Δ_n as a function of μ so far remains out of reach, we take nontrivial steps in this direction by providing upper and lower bounds in terms of the covering numbers. To state our bounds, we fix a probability measure μ on $\{0,1\}^{\mathbb{N}}$. We define the metric spaces $(\mathbb{N},\xi)$ and $(\mathbb{N},\rho)$ as in Eq (6) and (7). For $\varepsilon \in (0,1]$, we will denote by $\mathcal{N}_\xi(\varepsilon)$ and $\mathcal{N}_\rho(\varepsilon)$ the ε-covering numbers of $(\mathbb{N},\xi)$ and $(\mathbb{N},\rho)$ respectively. We recall that the covering number $\mathcal{N}(\epsilon)$ for a metric space (M,d) is the minimum cardinality of a covering set $S \subset M$, that is, such that for any $x \in M$, there exists $y \in S$ with $d(x,y) \leq \epsilon$. Recall that a subset of a metric space is *totally bounded* if its ε-covering numbers are finite for each $\varepsilon > 0$.

We start by showing that total boundedness is a necessary condition for convergence. The proof is deferred to Appendix B.

Theorem 4 *Let μ be a distribution on $\{0,1\}^{\mathbb{N}}$ be such that $(\mathbb{N},\xi)$ is not totally bounded and let $\varepsilon \in (0,1]$ be such that $\mathcal{N}_\xi(\varepsilon) = \infty$. Then, for all $n \geq 1$,*

$$\Delta_n(\mu) \geq \frac{\varepsilon^2}{6}.$$

Hence, if $\Delta_n(\mu) \to 0$, then $(\mathbb{N},\xi)$ is totally bounded.

The proof uses the following idea: given an ε-separated infinite subset $S \subset \mathbb{N}$, that is, one for which $\xi(i,j) = \mathbb{P}(x_i \neq X_j) > \varepsilon$ for all $i,j \in S$, we can show that with reasonable probability, the sequence $\{X_i : i \in S\}$ contains infinitely many realizations of both 0 and 1:

$$\mathbb{P}\left\{\min\left(\sum_{i\in S} X_i, \sum_{i\in S}(1-X_i)\right) = \infty\right\} \geq \epsilon$$

The proof then constructs a sequence of decreasing (random) infinite sets $S(n) \subset \mathbb{N}$ for $n \geq 1$ such that on the first n iid samples $X^{(1)},\dots,X^{(n)}$, all entries $i \in S(n)$ diverged from their mean significantly: $|\hat{p}_i - p_i| \gtrsim \varepsilon p_i$. These sets can be constructed recursively as follows. Given an infinite set $S(n) \subset A$, we can still apply the initial result to $S(n)$ because it is still ε-separated and infinite. Hence, for any $j > n$, with probability at least ε, the sequence $\{X_i^{(j)} : i \in S(n)\}$ contains infinitely many realizations of both 0 and 1. Whenever this event occurs for some samples $n' > n$ (after a geometric waiting time $n' - n \sim \mathcal{G}(\varepsilon)$), we can either select the entries $S(n') \subset S(n)$ as those with realization 0, or those with realization 1, whichever induces maximal deviation from the means. In both cases, $S(n')$ is still infinite hence we can continue the induction. The deviation on some entry $i \in S(n')$ is of order p_i after a waiting time $n' - n \approx 1/\varepsilon$, which corresponds to a $\approx p_i\varepsilon$ deviation in average; this can be replaced with $\approx \varepsilon^2$ with additional technicalities.

Although the necessary condition in Theorem 4 might at first appear somewhat weak, it turns out to be essentially the best one can do solely based on covering-number information:

Proposition 5 *Let $N : (0,1] \to \mathbb{N}$ be a non-increasing function. Then, there exists a distribution μ on $\{0,1\}^{\mathbb{N}}$ and a countable set E such that for any $\varepsilon \in (0, \frac{1}{2}] \setminus E$, one has $\mathcal{N}_\xi(\varepsilon) = N(\varepsilon)$, and*

$$\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0.$$

The proof is given in Appendix B. The previous result shows that if μ satisfies the necessary condition from Theorem 4, one can construct a distribution $\tilde{\mu}$ on $\{0,1\}^{\mathbb{N}}$ whose covering numbers agree with those of μ (up to a negligible countable set) for which $\Delta_n(\tilde{\mu})$ does converge to zero. Here, the exceptional set E exactly corresponds to the scales ε on which the covering number function $N(\cdot)$ is discontinuous: for the constructed distribution μ, the covering numbers are right-continuous, which may not necessarily be the case for a generic covering number function $\mathcal{N}_\xi(\cdot)$.

In light of Theorem 4, there is no loss of generality in assuming that all covering numbers $\mathcal{N}_\xi(\varepsilon)$ (a fortiori for $\mathcal{N}_\rho(\varepsilon)$) are finite. In this case, we have the following sufficient condition for the convergence of $\Delta_n(\mu)$ to 0, written in terms of the ξ-covering numbers.

Theorem 6 *Let μ be a distribution on $\{0,1\}^{\mathbb{N}}$ such that*

$$C_\mu := \int_0^1 \mathcal{N}_\xi(\varepsilon)\mathrm{d}\varepsilon < \infty. \tag{8}$$

Then, $\Delta_n(\mu) \to 0$ as $n \to \infty$. Further, in that case, for any $n \geq 1$,

$$\Delta_n(\mu) \leq C_1 \inf_{\varepsilon\in(0,1]} \varepsilon + \sqrt{\frac{\log(\mathcal{N}_\xi(\varepsilon)+1)}{n}} \leq C_2 \left(1 \wedge \sqrt{\frac{\log(n(1+C_\mu))}{n}}\right),$$

for universal constants $C_1, C_2 > 0$.

The proof, which we defer to Appendix B, proceeds via a chaining argument. For any $\epsilon > 0$, we denote by $\mathcal{S}_\xi(\epsilon)$ an ϵ-covering set with minimum cardinatity, that is $|\mathcal{S}_\xi(\epsilon)| = \mathcal{N}_\xi(\epsilon)$. We consider the directed graph such that a node in $\mathcal{S}_\xi(2^{-k})$ (covering set at level 2^{-k}) has as parent its nearest-neighbor within $\mathcal{S}_\xi(2^{-k+1})$. We show that the property $C_\mu < \infty$ ensures that with good probability, starting for any point at level $k > k_0$, the value at the node and its parent coincide exactly. This overall probability grows as k_0 is made larger. Conditional on this good event, it suffices to bound the deviation of entries in $\mathcal{S}_\xi(2^{-k})$ for $k \leq k_0$, for which convergence is trivial due to the finite number of points.

We note that the non-asymptotic upper bounds for $\Delta_n(\mu)$ of the previous result are in general not tight. In particular, under stronger assumptions, we can use tools from Gaussian process theory to exhibit subgaussian rates of convergence ($1/\sqrt{n}$ instead of $\sqrt{\log n/n}$) for $\Delta_n(\mu)$. More precisely, we show that $X \sim \mu$ is a subgaussian process on $\mathbb{N}$ with respect to the metric ρ (Lemma 18). Hence, a direct application of Dudley's theorem Van Handel (2014, Corollary 5.25) yields the following result.

Proposition 7 *Let μ be a distribution on $\{0,1\}^{\mathbb{N}}$ such that*

$$D_\mu := \int_0^1 \sqrt{\log \mathcal{N}_\rho(\varepsilon)}\mathrm{d}\varepsilon < \infty.$$

Then, for any $n \geq 1$, we have

$$\Delta_n(\mu) \leq \frac{24 D_\mu}{\sqrt{n}}.$$

The proof that $X \sim \mu$ is a subgaussian process (Lemma 18) along with the proof of the above proposition are deferred to Appendix C.

In light of Theorem 3, the condition from Eq (8) cannot also be a necessary condition for $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$, since it only involves pairwise correlations between coordinates of the distribution. However, we show that this is also essentially the tightest sufficient condition that can be stated solely in terms of the covering numbers $\mathcal{N}_\xi(\cdot)$. The proof can be found in Appendix B.

Proposition 8 *Let $N : (0, 1] \to \mathbb{N}$ be a non-increasing function such that*

$$\int_0^{1/2} N(\varepsilon)\mathrm{d}\varepsilon = \infty.$$

Then, there exists a distribution μ on $\{0, 1\}^{\mathbb{N}}$ and a countable set E such that for any $\varepsilon \in (0, \frac{1}{2}] \setminus E$, one has $\mathcal{N}_\xi(\varepsilon) = N(\varepsilon)$, and for any $n \geq 1$,

$$\Delta_n(\mu) = \frac{1}{2}.$$

Towards characterizing distributions μ for which $\Delta_n(\mu)$ decays to 0, we have the following sufficient condition, which subsumes the condition from Eq (8) and does not involve covering numbers. The proof is given in Appendix D.

Theorem 9 *Let μ be a distribution on $\{0, 1\}^{\mathbb{N}}$ such that $(\mathbb{N}, \xi)$ is totally bounded. Suppose that there exists $K \geq 1$ such that for any $\varepsilon > 0$, there exist events $(E_k)_{k\in\mathbb{N}}$ and a finite set $J \subset \mathbb{N}$ with*

- $\mathbb{P}(E_k) \leq \varepsilon, \ \forall k \in \mathbb{N}$,
- $\sup\limits_{k\in\mathbb{N}} \dfrac{\log(k+1)}{\log \frac{1}{\mathbb{P}(E_k)}} < \infty$,
- $\forall i \in \mathbb{N}, \exists j \in J, \exists \mathcal{K} \subset \mathbb{N}$, *such that* $|\mathcal{K}| \leq K$ *and* $\{X_i \neq X_j\} \subset \bigcup_{k\in\mathcal{K}} E_k$.

Then $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$.

While covering numbers $\mathcal{N}_\xi(\varepsilon)$ aim to account for the bad events when pairs of coordinates differ $\{X_i \neq X_j\}$ for $i, j \in \mathbb{N}$, the condition from Theorem 9 generalizes this covering number approach in two distinct ways: (i) Instead of considering the bad events $\{X_i \neq X_j\}$ only through the expectation $\xi(i, j) = \mathbb{P}(X_i \neq X_j)$, the condition allows for positive correlations of these bad events. Hence, instead of defining balls in the space $(\mathbb{N}, \xi)$, we directly define events E_k in the probability space of μ. (ii) Instead of covering a bad event $\{X_i \neq Z_j\}$ with a single event E_k (or ball in the covering number approach), we allow for the event to be covered by several events E_k. However, the number of covering events E_k for each bad event needs to be bounded (condition $|\mathcal{K}| \leq K$).

Both generalizations are important to define the exact characterization for distributions μ such that $\Delta_n(\mu) \to 0$. In particular, the necessity of the first generalization (i) is already exemplified by Theorem 3. In Section D, following the proofs of the latter claims, we provide illustrative examples of distributions demonstrating the necessity of these generalizations (distributions μ for which $\Delta_n(\mu) \to 0$ and that wouldn't satisfy the condition of Theorem 9 without such a generalization). These are meant to guide the intuition in future work towards the full characterization, which we leave as an open question.

Open problem. Characterize the distributions μ on $\{0,1\}^{\mathbb{N}}$ for which $\Delta_n(\mu) \to 0$ as $n \to \infty$. Is the sufficient condition from Theorem 9 also necessary to have $\Delta_n(\mu) \to 0$?

Open problem. How can the covering numbers $\mathcal{N}_\rho(\varepsilon)$ or $\mathcal{N}_\xi(\varepsilon)$ be calculated or bounded explicitly in terms of μ?

3. Proof of Theorem 1 (Decoupling from below)

Notation. Throughout this section, whenever $X_1, X_2, \ldots$ is a collection of (one-dimensional) random variables, we denote by $\tilde{X}_1, \tilde{X}_2, \ldots$ its *independent version*: the $\tilde{X}_i$s are mutually independent (and independent of the X_is), and each X_i is equivalent to $\tilde{X}_i$ in distribution. In this section, we will provide comparison results of the type $\mathbb{E}\max_{i\in[d]} X_i \gtrsim \mathbb{E}\max_{i\in[d]} \tilde{X}_i$ under negative pairwise correlation conditions on the X_i.

Bernoulli case. We begin with the case where $X_i \sim \mathrm{Bernoulli}(p_i)$, $i \in [d]$. Letting $Z = \sum_{i=1}^d X_i$ and $\tilde{Z} = \sum_{i=1}^d \tilde{X}_i$, we have

$$\mathbb{E}\max_{i\in[d]} X_i = \mathbb{P}(Z > 0), \qquad \mathbb{E}\max_{i\in[d]} \tilde{X}_i = \mathbb{P}(\tilde{Z} > 0). \tag{9}$$

Let us recall the notion of pairwise independence: for each $i \neq j \in [d]$, we have $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\,\mathbb{E}[X_j]$.

Proposition 10 *Let $X_i, \tilde{X}_i, Z, \tilde{Z}, p_i$ be as in (9) and assume additionally that the X_i are pairwise independent. Then*

$$\mathbb{P}(Z > 0) \geq \frac{1}{2}\mathbb{P}(\tilde{Z} > 0).$$

Proof By the Paley-Zygmund inequality,

$$\mathbb{P}(Z > 0) \geq \frac{(\mathbb{E}\,Z)^2}{\mathbb{E}[Z^2]}.$$

Now $\mathbb{E}\,Z = \sum_{i=1}^d p_i$, which we assume without loss of generality to be non-zero (otherwise $\mathbb{P}(Z > 0) = \mathbb{P}(\tilde{Z} > 0) = 0$), and by pairwise independence,

$$\mathbb{E}[Z^2] = \sum_{i=1}^d p_i + 2\sum_{1\leq i<j\leq d} p_i p_j = \sum_{i=1}^d p_i + \left(\sum_{i=1}^d p_i\right)^2 - \sum_{i=1}^d p_i^2 \leq \sum_{i=1}^d p_i + \left(\sum_{i=1}^d p_i\right)^2. \tag{10}$$

Hence,

$$\frac{(\mathbb{E}\,Z)^2}{\mathbb{E}[Z^2]} \geq \frac{\left(\sum_{i=1}^d p_i\right)^2}{\sum_{i=1}^d p_i + \left(\sum_{i=1}^d p_i\right)^2} = \frac{\sum_{i=1}^d p_i}{1 + \sum_{i=1}^d p_i}.$$

On the other hand, $\mathbb{P}(\tilde{Z} > 0)$ is readily computed:

$$\mathbb{P}(\tilde{Z} > 0) = 1 - \prod_{i=1}^d (1 - p_i).$$

Therefore, to prove the claim, it suffices to show that

$$G(p_1,\dots,p_d) := 2\sum_{i=1}^d p_i - \left(1+\sum_{i=1}^d p_i\right)\left(1-\prod_{i=1}^d (1-p_i)\right) \geq 0.$$

To this end, we write $G = S + P + SP - 1$, where $S = \sum_i p_i$ and $P = \prod_i (1-p_i)$. Now if $S \geq 1$ then obviously $G \geq 0$ and we are done. Otherwise, since $P \geq 1 - S$ trivially holds (which can be viewed as an application of the union bound), we have $G \geq S(1-S)$. In this case, $S < 1 \implies G \geq 0$. ■

Relaxing pairwise independence. An inspection of the proof shows that we do not actually need $\mathbb{E}[X_i X_j] = p_i p_j$, but rather only $\mathbb{E}[X_i X_j] \leq p_i p_j$. This condition is called *negative (pairwise) covariance* (Dubhashi and Ranjan, 1998).

Corollary 11 *Let $X_i, \tilde{X}_i, Z, \tilde{Z}, p_i$ be as in (9) and assume additionally that the X_i satisfy negative pairwise covariance: $\mathbb{E}[X_i X_j] \leq p_i p_j$ for $i \neq j$. Then*

$$\mathbb{P}(Z > 0) \geq \frac{1}{2}\mathbb{P}(\tilde{Z} > 0).$$

General positive random variables. Now let $X_1,\dots,X_d$ be non-negative integrable random variables and the $\tilde{X}_1,\dots,\tilde{X}_d$ are their independent copies: each $\tilde{X}_i$ is distributed identically to X_i and the $\tilde{X}_i$ are mutually independent.

Proposition 12 *Let $X_1,\dots,X_d$ be non-negative and integrable with independent copies $\tilde{X}_i$ as above. If additionally the X_i are pairwise independent, then*

$$\mathbb{E}\max_{i\in[d]} X_i \geq \frac{1}{2}\,\mathbb{E}\max_{i\in[d]} \tilde{X}_i.$$

Proof For $t > 0$ and $i \in [d]$, put $Y_i(t) = \mathbf{1}[X_i > t]$, $\tilde{Y}_i(t) = \mathbf{1}[\tilde{X}_i > t]$ and $Z(t) = \sum_{i=1}^d Y_i(t)$, $\tilde{Z}(t) = \sum_{i=1}^d Y_i(t)$. Then

$$\begin{aligned}
\mathbb{E}\max_{i\in[d]} X_i &= \int_0^\infty \mathbb{P}\left(\max_{i\in[d]} X_i > t\right) \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left(Z(t) > 0\right) \mathrm{d}t \\
&\geq \frac{1}{2}\int_0^\infty \mathbb{P}\left(\tilde{Z}(t) > 0\right) \mathrm{d}t \\
&= \frac{1}{2}\int_0^\infty \mathbb{P}\left(\max_{i\in[d]} \tilde{X}_i > t\right) \mathrm{d}t \\
&= \frac{1}{2}\,\mathbb{E}\max_{i\in[d]} \tilde{X}_i,
\end{aligned}$$

where Proposition 10 was invoked in the inequality step. ■

Relaxing pairwise independence. As before, the full strength of pairwise independence of the X_i is not needed. The condition $\mathbb{P}(X_i > t, X_j > t) \leq \mathbb{P}(X_i > t)\mathbb{P}(X_j > t)$ for all $i \neq j \in [d]$ and $t > 0$, called pairwise *negative upper orthant dependence* (Joag-Dev and Proschan, 1983, Definition 2.3), would suffice; it is weaker than pairwise independence.

Corollary 13 *Let $X_1, \ldots, X_d$ be non-negative and integrable with independent copies $\tilde{X}_i$ as above. If additionally the X_i verify $\mathbb{P}(X_i > t, X_j > t) \leq \mathbb{P}(X_i > t)\mathbb{P}(X_j > t)$ for all $i \neq j \in [d]$ and $t > 0$. Then*

$$\mathbb{E} \max_{i \in [d]} X_i \geq \frac{1}{2} \mathbb{E} \max_{i \in [d]} \tilde{X}_i.$$

Definition 14 (Joag-Dev and Proschan (1983), Definition 2.1) *Random variables $X_1, X_2, ..., X_k$ are said to be* negatively associated *(NA) if for every pair of disjoint subsets A_1, A_2 of $\{1, 2, ..., k\}$,*

$$\mathrm{Cov}\left[f_1(X_i,\ i \in A_1), f_2(X_j,\ j \in A_2)\right] \leq 0$$

whenever f_1 and f_2 are increasing. NA may also refer to the vector $X = (X_1, \ldots, X_k)$ or to the underlying distribution of X. Additionally, NA may denote negative association. If $|A_1| = |A_2| = 1$ we say that X is pairwise negatively associated *(PNA). Note that the definition is the same if both f_1 and f_2 are decreasing.*

Theorem 15 (Restatement of Theorem 1) *Let μ be a probability measure on $\{0,1\}^{\mathbb{N}}$ such that for $X \sim \mu$, the ith and jth entries of X satisfy $\mathbb{E}[X_i X_j] \leq \mathbb{E}[X_i]\,\mathbb{E}[X_j]$. Let $X^{(1)}, \ldots, X^{(n)}$ be n independent copies of X and define $Y = n^{-1}\sum_{i=1}^n X^{(i)} - \mathbb{E}[X]$. Let $\tilde{Y}$ be the independent version of Y: each entry $\tilde{Y}_j$ is equal to Y_j in distribution and the $(\tilde{Y}_j)_{j \in \mathbb{N}}$ are mutually independent. Then*

$$\mathbb{E} \sup_{j \in \mathbb{N}} |Y_j| \geq \frac{1}{4} \mathbb{E} \sup_{j \in \mathbb{N}} \left|\tilde{Y}_j\right|.$$

Proof The proof consists of three parts. We first show that the condition $\mathbb{E}[X_i X_j] \leq \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ implies that X is pairwise negatively associated. Then, we show that this implies Y is also PNA. Finally, we invoke Corollary 13 for the vector $|Y|$ and we are done.

For any i, j let $p_i := \mathbb{E}[X_i]$ and $r_{ij} := \mathbb{E}[X_i X_j]$. By assumption, $r_{ij} \leq p_i p_j$. Let f, g be two real-valued non-decreasing functions, then noting that $\mathbb{P}(X_i = 1, X_j = 0) = \mathbb{E}[X_i - X_i X_j] = p_i - r_{ij}$ and similarly, $\mathbb{P}(X_i = 0, X_j = 1) = p_j - r_{ij}$, we have

$$\begin{aligned}
&\mathbb{E}[f(X_i)g(X_j)] - \mathbb{E}[f(X_i)]\,\mathbb{E}[g(X_j)] \\
&\quad = r_{ij} f(1)g(1) + (p_i - r_{ij}) f(1)g(0) + (p_j - r_{ij}) f(0)g(1) + (1 - p_i - p_j + r_{ij}) f(0)g(0) \\
&\quad - p_i p_j f(1)g(1) - p_i(1 - p_j) f(1)g(0) - (1 - p_i) p_j f(0)g(1) - (1 - p_i)(1 - p_j) f(0)g(0) \\
&\quad = (r_{ij} - p_j p_j)(f(1)g(1) - f(1)g(0) - f(0)g(1) + f(0)g(0)) \\
&\quad = (r_{ij} - p_j p_j)(f(1) - f(0))(g(1) - g(0)) \\
&\quad \leq 0
\end{aligned}$$

for all i, j. This shows that X is PNA. Since any pair (X_i, X_j) is NA, by Joag-Dev and Proschan (1983, Property P_7), the union of independent sets of NA random variables are NA, so

$$(X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(n)}, X_j^{(1)}, X_j^{(2)}, \ldots, X_j^{(n)})$$

is also NA. Now, we apply Joag-Dev and Proschan (1983, Property P_6): increasing (or decreasing) functions defined on disjoint subsets of a set of NA random variables are NA, where the increasing functions are $[Y_i]_+ := f_+(X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(n)}) := \left[n^{-1}\sum_{k=1}^n X_i^{(k)} - \mathbb{E}[X_i]\right]_+$ for all i. We conclude that $[Y]_+ = ([Y_1]_+, [Y_2]_+, \dots)$ is PNA. In the same manner, $[Y]_-$ is also PNA. Moreover, PNA obviously implies pairwise negative upper orthant dependence. Now,

$$\begin{aligned}
\mathbb{E}\sup_{i\in\mathbb{N}}|Y_i| &= \mathbb{E}\sup_{i\in\mathbb{N}}[Y_i]_+ + [Y_i]_- \\
&\geq \max\left(\mathbb{E}\sup_{i\in\mathbb{N}}[Y_i]_+, \mathbb{E}\sup_{i\in\mathbb{N}}[Y_i]_-\right) \\
&\geq \frac{1}{2}\mathbb{E}\sup_{i\in\mathbb{N}}[Y_i]_+ + \frac{1}{2}\mathbb{E}\sup_{i\in\mathbb{N}}[Y_i]_- \\
&\geq \frac{1}{4}\mathbb{E}\sup_{i\in\mathbb{N}}\left[\tilde{Y}_i\right]_+ + \frac{1}{4}\mathbb{E}\sup_{i\in\mathbb{N}}\left[\tilde{Y}_i\right]_- \\
&\geq \frac{1}{4}\mathbb{E}\sup_{i\in\mathbb{N}}\left|\tilde{Y}_i\right|,
\end{aligned}$$

where the second inequality is due to Corollary 13. ■

4. Proof of Theorem 3 (Covariance does not characterize Δ_n)

In this section, we provide an example of two distributions μ and ν on $\{0,1\}^{\mathbb{N}}$ that share the same covariance matrix but for which $\Delta_n(\mu) \to 0$ and $\Delta_n(\nu) \to \frac{1}{2}$ as $n \to \infty$.

Construction of the example. We partition $\mathbb{N}$ as follows: $\mathcal{S}_k = \{2^{(k-1)^3} < t \leq 2^{k^3}\}$ for $k \geq 1$. Let $(Z_1)_{i\geq 1} \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ and $(Y_k)_{k\geq 0} \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ be iid sequences.

Let $\gamma_k, \delta_k \in (0,1)$ be the solutions to

$$\begin{cases} (1-\delta_k)(2\gamma_k - \gamma_k^2) &= 2^{-k} \\ \gamma_k + \delta_k - \gamma_k\delta_k &= 2^{-k}. \end{cases}$$

We can check that $\gamma_k = \frac{2^{-k}-\delta_k}{1-\delta_k}$ and $\delta_k = 1 - 2^{-k-1} - \sqrt{(1-2^{-k-1})^2 - 2^{-k} + 2^{-2k}}$. In particular, $\gamma_k \sim \delta_k \sim 2^{-k-1}$ as $k \to \infty$.

- Let $(B_k)_{k\geq 1}$ be a sequence of independent random variables with $B_k \sim \text{Bernoulli}(2^{-k})$. Put

  $$X_t^\mu = (1-B_k)Z_0 + B_k Z_t, \quad t \in \mathcal{S}_k,\ k \geq 1.$$

  We define μ as the distribution of $(X_i^\mu)_{i\geq 1}$.

- Let $(C_i)_{i\geq 1}$ and $(D_k)_{k\geq 1}$ be independent sequences of independent random variables with $C_i \sim \text{Bernoulli}(\gamma_k)$ for $i \in \mathcal{S}_k$ and $D_k \sim \text{Bernoulli}(\delta_k)$. We let

  $$X_t^\nu = D_k Y_k + (1-D_k)((1-C_t)Z_0 + C_t Z_t), \quad t \in \mathcal{S}_k,\ k \geq 1.$$

  We define ν as the distribution of $(X_i^\nu)_{i\geq 1}$.

We can easily check that these distributions have the same covariance matrix.

Lemma 16 *The two distributions μ and ν constructed above satisfy*

$$\mathbb{E}_{X\sim\mu}[X_iX_j] = \mathbb{E}_{X\sim\nu}[X_iX_j], \quad i,j \in \mathbb{N}.$$

The proof is given in Appendix A.

At the high level, both distributions are constructed by block $\mathcal{S}_k$ such that within the block, the random variables are correlated up to a level $\approx 1 - 2^{-k}$. In the first distribution μ, the correlation between variables X_i for $i \in \mathcal{S}_k$ is made uniform through the decision variable B_k. Hence, to have convergence of the maximum deviation on $\mathcal{S}_k$ it suffices to handle a single decision variable B_k, which is a Bernoulli with parameter 2^{-k}. Because these parameters are summable, we can control these variables uniformly for all k sufficiently large.

In the second case for ν, this correlation is made heterogeneous, by introducing decision variables C_i for each $i \in \mathcal{S}_k$. These are independent and by taking $|\mathcal{S}_k|$ sufficiently large, one can enforce rare deviation events to happen with high probability for one of the variables $i \in \mathcal{S}_k$.

We formalize these ideas in the rest of this section. For clarity, for any $i \geq 1$, we denote by $\hat{p}_i(\mu)$ and $\hat{p}_i(\nu)$ the quantities $\hat{p}_i$ for μ and ν respectively. They share the same means p_i so we need not make the distinction here. We also denote with an exponent $U^{(1)}, U^{(2)}, \ldots$ iid samples from any random variable U.

Expected maximum deviation for μ. Fix $k \geq 1$. One has

$$\begin{aligned}\max_{i\in\mathcal{S}_k} |\hat{p}_i(\mu) - p_i| &\leq \left|\frac{1}{n}\sum_{i=1}^n Z_0^{(i)} - \frac{1}{2}\right| + \max_{i\in\mathcal{S}_k}\left|\hat{p}_i(\mu) - \frac{1}{n}\sum_{i=1}^n Z_0^{(i)}\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^n Z_0^{(i)} - \frac{1}{2}\right| + \frac{1}{n}\sum_{i=1}^n B_k^{(i)}.\end{aligned}$$

As a result,

$$\begin{aligned}\mathbb{E}\sup_{l\geq k}\max_{i\in\mathcal{S}_l} |\hat{p}_i(\mu) - p_i| &\leq \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_0^{(i)} - \frac{1}{2}\right| + \mathbb{E}\sup_{l\geq k}\frac{1}{n}\sum_{i=1}^n B_l^{(i)} \\ &\leq \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_0^{(i)} - \frac{1}{2}\right| + \sum_{l\geq k}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n B_l^{(i)}\right] \\ &= \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_0^{(i)} - \frac{1}{2}\right| + 2^{-k+1}.\end{aligned}$$

Then,

$$\begin{aligned}\Delta_n(\mu) &\leq \sum_{t=1}^{2^{(k-1)^3}} \mathbb{E}|\hat{p}_t(\mu) - p_t| + \mathbb{E}\sup_{l\geq k}\max_{i\in\mathcal{S}_l}|\hat{p}_i(\mu) - p_i| \\ &\leq (1 + 2^{(k-1)^3})\mathbb{E}|\hat{p}_1(\mu) - p_1| + 2^{-k+1}.\end{aligned}$$

In particular, this shows that $\limsup_{n\to\infty}\Delta_n(\mu) \leq 2^{-k+1}$. Because this holds for all $k \geq 1$, we obtained $\Delta_n(\mu) \to 0$ as $n \to \infty$.

Expected maximum deviation for ν. Fix $n \geq 1$. We have

$$\mathbb{P}(\exists i \in \mathcal{S}_n, \forall j \in [n], C_i^{(j)} = 1,\ Z_i^{(j)} = 1) = 1 - \left(1 - \left(\frac{\gamma_n}{2}\right)^n\right)^{|\mathcal{S}_n|} \geq 1 - \exp\left(-\left(\frac{\gamma_n}{2}\right)^n |\mathcal{S}_n|\right)$$

Denote by $\mathcal{E}_n$ this event. On this event, there exists $i \in \mathcal{S}_n$ such that

$$\hat{p}_i(\nu) = \frac{1}{n}\sum_{j=1}^n X_i^{\nu(j)} = \frac{1}{n}\sum_{i=1}^n D_n^{(j)} Y_n^{(j)} + 1 - D_n^{(j)}.$$

Then,

$$\begin{aligned}
\Delta_n(\nu) &\geq \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_n} \sup_{i \in \mathcal{S}_n} |\hat{p}_i(\nu) - p_i|\right] \\
&\geq \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_n} \left|\frac{1}{n}\sum_{j=1}^n (D_n^{(j)} Y_n^{(j)} + 1 - D_n^{(j)}) - \frac{1}{2}\right|\right] \\
&\geq \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n (D_n^{(j)} Y_n^{(j)} + 1 - D_n^{(j)}) - \frac{1}{2}\right] - \mathbb{P}(\mathcal{E}_n^c) \\
&\geq \frac{1}{2} - \frac{\delta_n}{2} - \exp\left(-\left(\frac{\gamma_n}{2}\right)^n |\mathcal{S}_n|\right).
\end{aligned}$$

We recall that $\gamma_n \sim 2^{-n-1}$ so that $\left(\frac{\gamma_n}{2}\right)^n |\mathcal{S}_n| \to \infty$ as $n \to \infty$. As a result, we obtain

$$\lim_{n \to \infty} \Delta_n(\nu) = \frac{1}{2}.$$

Acknowledgments

AK was partially supported by the Israel Science Foundation (grant No. 1602/19), an Amazon Research Award, and the Ben-Gurion University Data Science Research Center. We thank Mark Kozlenko for helpful comments on an earlier manuscript.

References

Witold Bednorz and Rafał Latała. On the boundedness of bernoulli processes. *Ann. Math.*, pages 1167–1203, November 2014.

Moïse Blanchard and Václav Voráček. Tight bounds for local glivenko-cantelli. In *Algorithmic Learning Theory, ALT 2024*, Proceedings of Machine Learning Research. PMLR, 2024.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806. PMLR, 2019. URL http://proceedings.mlr.press/v99/cherapanamjeri19b.html.

Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter L. Bartlett, and Michael I. Jordan. Optimal mean estimation without a variance. *CoRR*, abs/2011.12433, 2020. URL https://arxiv.org/abs/2011.12433.

Lorán Chollete, Victor de la Peña, and Michael Klass. The price of independence in a model with unknown dependence. *Mathematical Social Sciences*, 123:51–58, 2023. ISSN 0165-4896. doi: https://doi.org/10.1016/j.mathsocsci.2023.02.008. URL https://www.sciencedirect.com/science/article/pii/S0165489623000215.

Doron Cohen and Aryeh Kontorovich. Local Glivenko-Cantelli. In *Conference on Learning Theory*, Proceedings of Machine Learning Research, 2023a.

Doron Cohen and Aryeh Kontorovich. Open problem: log(n) factor in "local glivenko-cantelli. In Gergely Neu and Lorenzo Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 5934–5936. PMLR, 2023b. URL https://proceedings.mlr.press/v195/cohen23b.html.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695 – 2725, 2016. doi: 10.1214/16-AOS1440. URL https://doi.org/10.1214/16-AOS1440.

Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33:1830–1840, 2020.

Devdatt Dubhashi and Desh Ranjan. Balls and bins: a study in negative dependence. *Random Struct. Algorithms*, 13(2):99–124, September 1998. ISSN 1042-9832. doi: 10.1002/(SICI)1098-2418(199809)13:2⟨99::AID-RSA1⟩3.0.CO;2-M. URL http://dx.doi.org/10.1002/(SICI)1098-2418(199809)13:2<99::AID-RSA1>3.0.CO;2-M.

Samuel B Hopkins. Mean estimation with sub-gaussian rates in polynomial time. 2020.

Kumar Joag-Dev and Frank Proschan. Negative Association of Random Variables with Applications. *The Annals of Statistics*, 11(1):286 – 295, 1983. doi: 10.1214/aos/1176346079. URL https://doi.org/10.1214/aos/1176346079.

I. Kaplansky. *Set Theory and Metric Spaces*. AMS Chelsea Publishing Series. AMS Chelsea Publishing, 2001. ISBN 9780821826942. URL https://books.google.co.il/books?id=FbKhAQAAQBAJ.

Jasper C.H. Lee and Paul Valiant. Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683, 2022. doi: 10.1109/FOCS52979.2021.00071.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783 – 794, 2019a. doi: 10.1214/17-AOS1639. URL https://doi.org/10.1214/17-AOS1639.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019b. doi: 10.1007/s10208-019-09427-x. URL https://doi.org/10.1007/s10208-019-09427-x.

Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393 – 410, 2021. doi: 10.1214/20-AOS1961. URL https://doi.org/10.1214/20-AOS1961.

Michel Talagrand. *Upper and lower bounds for stochastic processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Springer, Berlin, Germany, 2014 edition, February 2014.

Thomas. Is uniform convergence faster for low-entropy distributions? Theoretical Computer Science Stack Exchange, 2018. URL https://cstheory.stackexchange.com/q/42009. URL:https://cstheory.stackexchange.com/q/42009 (version: 2018-12-10).

Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Appendix A. Second and third-order moments do not characterize $\Delta_n(\mu)$

We start this section by giving the omitted proof of Lemma 16, which states that the two distributions constructed for Theorem 3 indeed have the same second-order moments.

Proof of Lemma 16 Both sequences $(X_i^\mu)_{i\geq 1}$ and $(X_i^\nu)_{i\geq 1}$ are sequences of Bernoullis of parameter $1/2$ because they are mixtures of Z_i and Y_k that are independent $\mathrm{Bernoulli}(1/2)$. Now for any $i,j \in \mathcal{S}_k$ with $i \neq j$, one has

$$\mathbb{E}[X_i^\mu X_j^\mu] = \frac{1}{2}\mathbb{P}[B_k = 0] + \frac{1}{4}\mathbb{P}[B_k = 1] = \frac{1}{2} - \frac{1}{2^{k+2}}.$$

Also, for $i \in \mathcal{S}_k$ and $j \in \mathcal{S}_l$ with $k \neq l$, one has

$$\mathbb{E}[X_i^\mu X_j^\mu] = \frac{1}{2}\mathbb{P}[B_k = B_l = 0] + \frac{1}{4}(1 - \mathbb{P}[B_k = B_l = 0]) = \frac{1}{2} - \frac{2^{-k} + 2^{-l} - 2^{-k-l}}{4}.$$

Next, we turn to the sequence $(X_i^\nu)_{i\geq 1}$. For any $i,j \in \mathcal{S}_k$ with $i \neq j$, one has

$$\begin{aligned}\mathbb{E}[X_i^\nu X_j^\nu] &= \frac{1}{2}\mathbb{P}[D_k = 1] + \frac{1}{2}\mathbb{P}[D_k = 0]\mathbb{P}[C_i = C_j = 0] + \frac{1}{4}\mathbb{P}[D_k = 0](1 - \mathbb{P}[C_i = C_j = 0]) \\ &= \frac{1}{2} - \frac{(1-\delta_k)(2\gamma_k - \gamma_k^2)}{4} = \mathbb{E}[X_i^\mu X_j^\mu].\end{aligned}$$

Last, for $i \in \mathcal{S}_k$ and $j \in \mathcal{S}_l$ with $k \neq l$, one has

$$\begin{aligned}\mathbb{E}[X_i^\nu X_j^\nu] &= \frac{1}{2}\mathbb{P}[D_k = D_l = C_i = C_j = 0] + \frac{1}{4}(1 - \mathbb{P}[D_k = D_l = C_i = C_j = 0]) \\ &= \frac{1}{2} - \frac{1 - (1-\gamma_k)(1-\delta_k)(1-\gamma_l)(1-\delta_l)}{4} \\ &= \frac{1}{2} - \frac{1 - (1-2^{-k})(1-2^{-l})}{4} = \mathbb{E}[X_i^\mu X_j^\mu].\end{aligned}$$

As a result, if μ (resp. ν) denotes the distribution of $(X_i^\mu)_{i\geq 1}$ (resp. $(X_i^\nu)_{i\geq 1}$), they both share the same covariance matrix. ■

Theorem 3 shows that knowledge of the covariance matrix is not sufficient to characterize the behavior of $\Delta_n(\mu)$ or not. We suspect that this negative result holds more generally, that is, for any $k \geq 1$, the kth order moments are not sufficient to characterize whether $\Delta_n(\mu) \to 0$ or not. As it turns out, the distributions μ and ν constructed for Theorem 3 also share the same 3rd-order moments.

Proposition 17 (Third-order moments do not characterize Δ_n) *There exist probability measures μ and ν on $\{0,1\}^{\mathbb{N}}$ that agree on their 3rd-order moments,*

$$\mathop{\mathbb{E}}_{X\sim\mu}[X_i X_j X_k] \quad = \quad \mathop{\mathbb{E}}_{X\sim\nu}[X_i X_j X_k], \qquad i,j,k \in \mathbb{N},$$

while $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$ and $\Delta_n(\nu) \underset{n\to\infty}{\longrightarrow} \frac{1}{2}$.

Proof We simply check that the two distributions μ and ν from the proof of Theorem 3 also share the same 3rd-order moments. Let $i,j,k \in \mathbb{N}$ be three indices. When they are not all distinct, the desired equation follows from the proof of Theorem 3 because the coordinates X_i are binary. Without loss of generality, we then suppose that they are distinct. We denote by $l(i)$ the index of the block corresponding to i, that is, such that $i \in \mathcal{S}_{l(i)}$. We define similarly $l(j)$ and $l(k)$.

Case 1. We first treat the simple case when $l(i) \notin \{l(j), l(k)\}$. Then, note that we can write

$$X_i^\nu = (1 - B'_{l(i)})Z_0 + B_{l(i)'} Z'_{i,l(i)}$$

where $B'_{l(i)} := D_{l(i)} + C_i - D_{l(i)}C_i$ and $Z'_{i,l(i)} = \mathbb{1}_{D_{l(i)}=1} Y_{l(i)} + \mathbb{1}_{D_{l(i)}=0} Z_i$. Note that $B'_{l(i)} \sim B_{l(i)}$ since $\delta_{l(i)} + \gamma_i - \delta_{l(i)}\gamma_i = 2^{-l(i)}$. Also, $Z'_{i,l(i)} \sim \text{Bernoulli}(\frac{1}{2})$, hence $Z'_{i,l(i)} \sim Z_i$. Further, they are independent, but more importantly, they are independent from all the variables that would be used to define X_j and X_k, except for Z_0. As a result, all that remains to check is that

$$\mathbb{E}[Z_0 X_i^\mu X_j^\mu] = \mathbb{E}[Z_0 X_i^\nu X_j^\nu].$$

We start with the distribution μ. If $l(j) \neq l(k)$, then,

$$\begin{aligned}\mathbb{E}[Z_0 X_i^\mu X_j^\mu] &= \frac{1}{2}\mathbb{P}[B_{l(j)} = B_{l(k)} = 0] + \frac{1}{8}(1 - \mathbb{P}[B_{l(j)} = B_{l(k)} = 0]) \\ &= \frac{1}{2} - \frac{3}{8}(2^{-l(j)} + 2^{-l(k)} - 2^{-l(j)-l(k)}).\end{aligned}$$

On the other hand, if $l(j) = l(k)$, then

$$\mathbb{E}[Z_0 X_i^\mu X_j^\mu] = \frac{1}{2}\mathbb{P}[B_{l(j)} = 0] + \frac{1}{8}\mathbb{P}[B_{l(j)} = 1] = \frac{1}{2} - \frac{3}{8}2^{-l(j)}.$$

Next, for ν, if $l(j) \neq l(k)$, we have

$$\begin{aligned}\mathbb{E}[Z_0 X_i^\mu X_j^\mu] &= \frac{1}{2}\mathbb{P}[D_{l(j)} = C_j = D_{l(k)} = C_k = 0] + \frac{1}{8}(1 - \mathbb{P}[D_{l(j)} = C_j = D_{l(k)} = C_k = 0]) \\ &= \frac{1}{2} - \frac{3}{8}(1 - (1 - \gamma_{l(j)})(1 - \delta_{l(j)})(1 - \gamma_{l(k)})(1 - \delta_{l(k)})) \\ &= \frac{1}{2} - \frac{3}{8}(2^{-l(j)} + 2^{-l(k)} - 2^{-l(j)-l(k)}).\end{aligned}$$

Last, if $l(j) = l(k)$,

$$\begin{aligned}\mathbb{E}[Z_0 X_i^\nu X_j^\nu] &= \frac{1}{2}\mathbb{P}[D_{l(j)} = C_j = C_k = 0] + \frac{1}{8}\mathbb{P}[D_{l(j)} = 0, C_j = C_k = 1] \\ &\qquad + \frac{1}{4}\mathbb{E}[\mathbb{1}_{D_{l(j)}=1} + \mathbb{1}_{D_{l(j)}=0}(\mathbb{1}_{C_j=0}\mathbb{1}_{C_k=1} + \mathbb{1}_{C_j=1}\mathbb{1}_{C_k=0})] \\ &= \frac{1}{2} - \frac{4\gamma_{l(j)} + 2\delta_{l(j)} - 4\gamma_{l(j)}\delta_{l(j)} - \gamma_{l(j)}^2 + \gamma_{l(j)}^2\delta_{l(j)}}{8} \\ &= \frac{1}{2} - \frac{\gamma_{l(j)} + \delta_{l(j)} - \gamma_{l(j)}\delta_{l(j)}}{4} - \frac{(1 - \delta_{l(j))}(2\gamma_{l(j)} - \gamma_{l(j)}^2)}{8} = \frac{1}{2} - \frac{3}{8}2^{-l(j)}.\end{aligned}$$

This concludes the first case.

Case 2. By symmetry, the only remaining case is if $l(i) = l(j) = l(k)$. For simplicity, we then denote $l := l(i)$. Then,

$$\mathbb{E}[X_i^\mu X_j^\mu X_k^\mu] = \frac{1}{2}\mathbb{P}[B_l = 0] + \frac{1}{8}\mathbb{P}[B_l = 1] = \frac{1}{2} - \frac{3}{8}2^{-l}.$$

On the other hand,

$$\begin{aligned}\mathbb{E}[X_i^\nu X_j^\nu X_k^\nu] &= \frac{1}{2}\mathbb{E}[\mathbb{1}_{D_l=1} + \mathbb{1}_{D_l=0}\mathbb{1}_{C_i=0}\mathbb{1}_{C_j=0}\mathbb{1}_{C_k=0}] + \frac{3}{4}\mathbb{E}[\mathbb{1}_{D_l=0}\mathbb{1}_{C_i=1}\mathbb{1}_{C_j=0}\mathbb{1}_{C_k=0}] \\ &\quad + \frac{3}{8}\mathbb{E}[\mathbb{1}_{D_l=0}\mathbb{1}_{C_i=1}\mathbb{1}_{C_j=1}\mathbb{1}_{C_k=0}] + \frac{1}{8}\mathbb{E}[\mathbb{1}_{D_l=0}\mathbb{1}_{C_i=1}\mathbb{1}_{C_j=1}\mathbb{1}_{C_k=1}] \\ &= \frac{1}{2} - \frac{3(2\gamma_l - 2\gamma_l\delta_l - \gamma_l^2 + \gamma_l^2\delta_l)}{8} \\ &= \frac{1}{2} - \frac{3(1-\delta_l)(2\gamma_l - \gamma_l^2)}{8} = \frac{1}{2} - \frac{3}{8}2^{-l}.\end{aligned}$$

This ends the proof that all 3rd-order moments agree for μ and ν. ■

Appendix B. Bounds on the expected maximum empirical deviation with ξ-covering numbers

The previous result Theorem 3 shows that having tight characterizations of when $\Delta_n(\mu)$ converges 0, cannot be achieved by focusing solely on pair-wise correlations of the coordinates. Nevertheless, we are still able to give useful bounds on $\Delta_n(\mu)$ using such information. In this section, we focus on the metric space $(\mathbb{N}, \xi)$ where the metric ξ is defined as in Eq (6) and provide necessary and sufficient conditions for the decay of the expected maximum empirical deviation $\Delta_n(\mu)$.

We start with proving Theorem 4 which shows that having finite ξ-covering numbers is necessary.

Proof of Theorem 4 Let $\varepsilon > 0$ such that the ε-covering number of $(\mathbb{N}, \xi)$ is infinite and let $S_0 \subset \mathbb{N}$ be an infinite set such that for all $i, j \in S_0$, $\xi(i,j) = \mathbb{P}(X_i \neq X_j) \geq \varepsilon$. Further, there must exist an interval $I = [p - \frac{\eta}{2}, p + \frac{\eta}{2}]$ of length $\eta > 0$ to be fixed later, such that $S_1 := S_0 \cap \{i \in \mathbb{N} : p_i \in I\}$ is also infinite. Since for distinct $i, j \in S_1$,

$$\varepsilon \leq \xi(i,j) \leq p_i + p_j \leq 2p_i + \eta,$$

necessarily, for any $i \in S_1$, one has

$$p_i \geq \frac{\varepsilon - \eta}{2}. \tag{11}$$

By symmetry, we also have

$$1 - p_i \geq \frac{\varepsilon - \eta}{2}. \tag{12}$$

We will focus only on the indices in S_1, using

$$\Delta_n(\mu) \geq \mathbb{E} \sup_{i \in S} |\hat{p}_i - p_i|.$$

Hence, without loss of generality, we will suppose that $S_1 = \mathbb{N}$. In the rest of the proof, we will use the notations $(X_i^{(1)})_i, (X_i^{(2)})_i, \ldots \overset{\text{iid}}{\sim} \mu$ for an iid sequence of samples of μ. Similarly, we will indicate by an exponent (n) any event corresponding to the sample sequence $(X_i^{(n)})_i$. For convenience, we will also consider a sample sequence $(X_i)_i \sim \mu$.

Fix an arbitrary infinite subset $S \in \mathbb{N}$. We consider the event that $(X_i)_{i\in S}$ contains an infinite number of 0 and 1,

$$\mathcal{E}(S) := \left\{ \min\left(\sum_{i\in S} X_i, \sum_{i\in S} (1-X_i) \right) = \infty \right\}.$$

On $\mathcal{E}(S)^c$, the sequence $(X_i)_{i\in S}$ contains either a finite number of 0 or 1. We then denote by $\bar{X}(S)$ the random variable on $\{0,1\}$ equal to the (infinite) majority among the sequence $(X_i)_{i\in S}$, that is,

$$\bar{X}(S) = \mathbb{1}_{\mathcal{E}(S)^c} \cdot \mathbb{1}\left[\sum_{i\in S} (1-X_i) < \infty \right].$$

As a first step, we will show that $\mathbb{P}(\mathcal{E}(S)) \geq \varepsilon$. By Fatou's lemma, enumerating $S = \{i_1 < i_2 < \dots\}$, one has

$$\begin{aligned}
\mathbb{P}(\mathcal{E}(S)) &= \mathbb{P}\left(\limsup_{j\to\infty} \mathbb{1}[X_{i_j} \neq X_{i_{j+1}}] > 0 \right) \\
&= \mathbb{E}\left[\limsup_{j\to\infty} \mathbb{1}[X_{i_j} \neq X_{i_{j+1}}] \right] \\
&\geq \limsup_{j\to\infty} \mathbb{P}(X_{i_j} \neq X_{i_{j+1}}) \geq \varepsilon.
\end{aligned}$$

Now supposing that $\mathbb{P}(\mathcal{E}(S)) < 1$ we define $\bar{p}(S) = \mathbb{E}[\bar{X}(S) \mid \mathcal{E}(S)^c]$ (otherwise, we can set it to $\bar{p}(S) = p$ for instance), the expected value of the majority vote on $(X_i)_{i\in S}$ provided that there is consensus (only a finite number of disagreements).

Case 1. We first consider the case when there exists an infinite subset $S \in \mathbb{N}$ such that for any infinite subset $S' \subset S$, one has $\mathbb{P}(\mathcal{E}(S')) < 1$ and $\bar{p}(S) \leq p$. We now prove a lower bound on $\Delta_n(\mu)$ by showing that with good probability there is an index $i \in S$ for which $\hat{p}_i$ deviates from p_i from below.

To do so, given the iid sequence $(X_i^{(1)})_i, (X_i^{(2)})_i, \dots \overset{\text{iid}}{\sim} \mu$, let $N(S)$ be the first index n for which the event $\mathcal{E}(S)$ holds. For every $n < N(S)$, there are only a finite number of disagreements in the sequences $(X_i^{(n)})_{i\in S}$, while $(X_i^{(n)})_{i\in S}$ contains an infinite number of 0. Hence, provided that $N(S) < \infty$, the set

$$S(1) = \{ i \in S, \forall n < N(S), X_i^{(n)} = \bar{X}^{(n)}, \text{ and } X_i^{(N(S))} = 0 \}$$

is infinite. All indices in $S(1)$ share the same values for the samples $n \in [N(S)]$. We denote that value $X_{S(1)}^{(n)}$ for convenience. We can now repeat the construction process with $S(1)$ under the almost-sure event $\{N(S) < \infty\}$: we can define the geometric random variable $N(S(1))$ which is the waiting time starting from $n = N(S) + 1$ for the event $\mathcal{E}(S(1))$ to occur. On the almost sure event $\{N(S) < \infty\} \cap \{N(S(1)) < \infty\}$, this defines a new set $S(2) \subset S(1)$ for which all indices in $S(2)$ shared the same values for the samples in $n \in [N(S)+1, N(S)+N(S(1))]$, and we denote by $\bar{X}_{S(2)}^{(n)}$ these common values.

As a result of the construction, on an almost sure event $\mathcal{F}$, we obtain a sequence of geometric random variables $(N(S(k)))_{k\geq 1}$ with parameter $\mathbb{P}(\mathcal{E}(S(k))) \geq \varepsilon$, together with random sets

$(S(k))_{k\geq 1}$ that are decreasing and all infinite. Intuitively, $\mathcal{F} = \bigcap_k \{N(S(k)) < \infty\}$. For convenience, letting $S(0) = S$, we denote by $N_k = N(S(0)) + \ldots + N(S(k))$. Under $\mathcal{F}$, we have that for any $N_{k-1} < n < N_k$,

$$\mathbb{E}[\bar{X}^{(n)}_{S(k)} \mid N_l, S(l), l \leq k] = \bar{p}(S(k)) \leq p$$

On the other hand, under $\mathcal{F}$, we have $\bar{X}^{(N_k)}_{S(k)} = 0$ by construction.

Now fix $n \geq 1$. We recall that on $\mathcal{F}$, there always exists an index $i_n \in \mathbb{N}$ for which $(X_i^{(l)})_{l\in[n]}$ coincides exactly with the sequence $\bar{X}^{(1)}_{S(0)}, \ldots, \bar{X}^{(N_0)}_{S(0)}, \bar{X}^{(N_0+1)}_{S(1)}, \ldots, X^{(N_1)}_{S(1)}, \ldots$ truncated at n samples. We denote by k_n the number of finished periods before sample n, that is, the index k such that $N_k \leq n < N_{k+1}$. Combining the previous equations gives

$$\mathbb{E}\left[\sum_{l=1}^{n} X^{(l)}_{i_n}\right] \leq (n - \mathbb{E}[k_n])p.$$

We recall that k_n corresponds to the maximum number of geometric variables with parameter at least ε such that the sum is at most n. Hence, k_n is dominated by the maximum number of $\mathcal{G}(\varepsilon)$ random variables such that the sum is at most n. That is, if $(T_k)_{k\geq 1} \overset{\text{iid}}{\sim} \mathcal{G}(\varepsilon)$ and we let $\tilde{k}_n$ be the index such that $\sum_{l=1}^{k} T_l \leq n < \sum_{l=1}^{k+1} T_l$, we have $\mathbb{E}[k_n] \leq \mathbb{E}[\tilde{k}_n]$. Further, we have that

$$\begin{aligned} n < \mathbb{E}\left[\sum_{k=1}^{\tilde{k}_n+1} T_k\right] &= \mathbb{E}\left[\sum_{k\geq 1} T_k \mathbb{1}_{k\leq \tilde{k}_n+1}\right] \\ &= \sum_{k\geq 1} \frac{1}{\varepsilon} \mathbb{P}(k \leq \tilde{k}_n + 1) \\ &= \frac{1 + \mathbb{E}[\tilde{k}_n]}{\varepsilon}. \end{aligned}$$

In the second inequality, we used the Wald observation that knowing whether $k \leq \tilde{k}_n + 1$ only requires knowing $T_1, \ldots, T_{k-1}$ (this corresponds exactly to the event $T_1 + \ldots + T_{k-1} \leq n$). For small values of n, we can use the following simple bound

$$\mathbb{E}[k_n] \geq \mathbb{P}(k_n \geq 1) = \mathbb{P}(T_1 \leq n) = 1 - (1-\varepsilon)^n.$$

This implies that $\mathbb{E}[k_n] \geq \mathbb{E}[\tilde{k}_n] > (\varepsilon n - 1) \vee 1 - (1-\varepsilon)^n$. We now show that this implies $\mathbb{E}[k_n] \gtrsim n\varepsilon$. For $n \geq \frac{2}{\varepsilon}$, this already shows $\mathbb{E}[k_n] \geq \frac{\varepsilon n}{2}$. Note that the function $n \mapsto 1 - (1-\varepsilon)^n$ is concave. Given that its value is 0 for $n = 0$, this gives for $n \leq \frac{2}{\varepsilon}$,

$$1 - (1-\varepsilon)^n \geq \frac{n\varepsilon}{2}\left(1 - (1-\varepsilon)^{2/\varepsilon}\right) \geq \frac{n\varepsilon}{2}(1 - \mathrm{e}^{-2}) \geq \frac{n\varepsilon}{3}.$$

This shows that in all cases, we obtained

$$\mathbb{E}[k_n] \geq \frac{n\varepsilon}{3}.$$

In particular, recalling that $p_{i_n} \in [p - \frac{\eta}{2}, p + \frac{\eta}{2}]$, and combining the previous equations we obtain

$$\begin{aligned}\Delta_n(\mu) &\geq \mathbb{E}\left[p - \frac{\eta}{2} - \hat{p}_{i_n}\right] \\ &\geq \frac{\mathbb{E}[k_n]}{n} p - \frac{\eta}{2} \\ &\geq \frac{\varepsilon}{6}(\varepsilon - \eta) - \frac{\eta}{2}.\end{aligned}$$

In the last inequality, we used the lower bound on p from Eq (11). We now turn to the second case.

Case 2. We now suppose that for every infinite subset $S \subset \mathbb{N}$, there exists an infinite subset $S' \subset S$ such that either $\mathbb{P}(\mathcal{E}(S')) = 1$ or $\bar{p}(S') \geq p$. We now give a lower bound on $\Delta_n(\mu)$ by showing that with good probability there is an index $i \in \mathbb{N}$ for which $\hat{p}_i$ deviates from p_i by above.

To do so, we construct some decreasing random sets similarly to Case 1. To begin, there exists an infinite subset $S(0) \subset \mathbb{N}$ such that either $\mathbb{P}(\mathcal{E}(S(0))) = 1$ or $\bar{p}(S(0)) \geq p$. As in Case 1, let $N(S(0))$ be the number of samples to wait before the event $\mathcal{E}(S(0))$ occurs. As before, on the event $\{N(S(0)) < \infty\}$, the set

$$\tilde{S}(1) = \{i \in S(0), \forall n < N(S(0)), X_i^{(n)} = \bar{X}^{(n)}, \text{ and } X_i^{(N(S(0)))} = 1\}$$

is infinite. Hence, there exists a subset $S(1) \subset \tilde{S}(1)$ for which either $\mathbb{P}(\mathcal{E}(S(1))) = 1$ or $\bar{p}(S(1)) \geq p$. We can now repeat the process starting from $S(1)$. We use the same notations as in Case 1: the induction constructs under an event $\mathcal{F}$ of full probability, some decreasing infinite sets $(S(k))_{k\geq 1}$, as well as their sequence of geometric random variables $(N(S(k)))_{k\geq 1}$ with parameter at least ε. We denote $N_k = N(S(0)) + \ldots + N(S(k))$. Similarly to before, under $\mathcal{F}$, we have that for any $N_{k-1} < n < N_k$,

$$\mathbb{E}[\bar{X}_{S(k)}^{(n)} \mid N_l, S(l), l \leq k] = \bar{p}(S(k)) \geq p,$$

and for any $k \geq 1$, by construction $\bar{X}_{S(k)}^{(N_k)} = 1$. Hence, for a fixed number of samples $n \geq 1$, there exists an index i_n for which until n the values $(X_{i_n}^l)_{l\in[n]}$ coincide with that of $\bar{X}$. As before, if k_n is the index k for which $N_k \leq n < N_{k+1}$, we obtain

$$\mathbb{E}\left[\sum_{l=1}^{n} X_{i_n}^{(l)}\right] \geq np + (1-p)\mathbb{E}[k_n].$$

As a result, we obtain

$$\begin{aligned}\Delta_n(\mu) \geq \mathbb{E}[\hat{p}_{i_n} - p_{i_n}] &\geq \mathbb{E}[\hat{p}_{i_n} - p] - \frac{\eta}{2} \\ &\geq \frac{\mathbb{E}[k_n]}{n}(1-p) - \frac{\eta}{2} \\ &\geq \frac{\varepsilon}{6}(\varepsilon - \eta) - \frac{\eta}{2}.\end{aligned}$$

In the last inequality, we used Eq (12). Combining the two cases and noting that these hold for any value of $\eta > 0$ yields the desired result. ■

We next show that essentially is the tightest necessary conditions that can be obtained using only the covering number $\mathcal{N}_\xi(\cdot)$.

Proof of Proposition 5 Fix such a non-increasing function $N : (0, 1] \to \mathbb{N}$. We start by constructing a distribution μ, then we check that it has the same covering numbers as $N(\cdot)$. Last we prove the desired convergence $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$.

Constructing the distribution μ. Since we will only focus on covering numbers for $\varepsilon \in (0, \frac{1}{2}]$ anyways, without loss of generality, we suppose that $N(\varepsilon) = 1$ for $\varepsilon \in (\frac{1}{2}, 1]$. Note that because N is non-increasing, it is discontinuous on a countable (potentially finite) number of points $(\varepsilon_k)_{k\geq 1}$, which we ordered by decreasing order, that is $0 < \varepsilon_{k+1} < \varepsilon_k \leq \frac{1}{2}$ for all $k \geq 1$. For convenience, let $\varepsilon_0 = \frac{1}{2}$. Since N takes values only on integers $\mathbb{N}$, we can also define the sequence $(N_k)_{k\geq 1}$ such that N_k is the value taken by N on the interval $(\varepsilon_{k-1}, \varepsilon_k)$, with the convention $N_1 = 1$ in the specific case when $\varepsilon_0 = \varepsilon_1$.

Let $(Z_i)_{i\geq 1} \overset{iid}{\sim} \text{Bernoulli}(\frac{1}{2})$ an iid sequence of Bernoulli random variables and $U \sim \mathcal{U}([0, 1])$ be an a independent uniform random variable in $[0, 1]$. We construct a sequence $(X_i)_{i\geq 1}$ as follows for $k \geq 0$,

$$X_i := \mathbb{1}[U > 2\epsilon_k]Z_1 + \mathbb{1}[U \leq 2\epsilon_k]Z_i, \quad N_k < i \leq N_{k+1}.$$

For convenience, we denote the event $\mathcal{E}_k := \{U \leq 2\epsilon_k\}$. We can directly check that $\mathbb{P}(\mathcal{E}_k) = 2\varepsilon_k$ and in particular, $X_1 = Z_1$.

Because the events $\mathcal{E}_k$ only depend on U which is independent of Z_1. In particular, this shows that $X_i \sim \text{Bernoulli}(\frac{1}{2})$ for all $i \geq 1$. Also, for $i \geq 2$ with $N_k < i \leq N_{k+1}$ for some $k \geq 2$, this implies

$$\xi(i, 1) = \mathbb{P}(X_i \neq X_1) = \frac{\mathbb{P}(\mathcal{E}_k)}{2} = \varepsilon_k.$$

We next compute distances between any two distinct entries $i < j \geq 2$. We let $k_i, k_j \geq 1$ such that $N_{k_i} < i \leq N_{k_i+1}$ and similarly for j.

$$\xi(i, j) = \mathbb{P}(X_i \neq X_j) = \frac{\mathbb{P}(\mathcal{E}_{k_i} \cup \mathcal{E}_{k_j})}{2} = \frac{\mathbb{P}(\mathcal{E}_{k_i})}{2} = \varepsilon_{k_i}.$$

In the third equality, we used the fact that the events $(\mathcal{E}_k)_{k\geq 0}$ are decreasing.

Computing the covering numbers of μ We clearly have $\mathcal{N}_\xi(\frac{1}{2}) = 1$ because $\xi(1, i) \leq \frac{1}{2}$ for all $i \geq 1$. Next, for any fixed $\varepsilon \in (0, \frac{1}{2})$, let $k \geq 1$ such that $\varepsilon \in [\varepsilon_k, \varepsilon_{k-1})$. We first note that the random variables $[N_k]$ form an ε-cover of $(\mathbb{N}, \xi)$. Indeed, for $i > N_k$, if $N_{k_i} < i \leq N_{k_i+1}$, one has $k_i \geq k$ so that

$$\xi(1, i) = \varepsilon_{k_i} \leq \varepsilon_k \leq \varepsilon.$$

As a result, $\mathcal{N}_\xi(\varepsilon) \leq N_{k-1}$. On the other hand, for any $2 \leq i, \leq N_k$, we observe that for any $j \neq i$, $\xi(i, j) \geq \varepsilon_{k-1} > \varepsilon$. As a result, an ε-cover of $(\mathbb{N}, \xi)$ must contain all elements $\{2, \ldots, N_k\}$ which has $N_k - 1$ elements. Note that this set does not ε-cover the element 1 since $\xi(1, i) \geq \varepsilon_{k-1} > \varepsilon$ for all $i \in \{2, \ldots, N_k\}$. Hence, the ε-cover must have at least N_k elements. Together with the previous remark, we obtained

$$\mathcal{N}_\xi(\varepsilon) = N_k, \quad \varepsilon \in [\varepsilon_k, \varepsilon_{k-1}).$$

As a result, wit $E = \{\varepsilon_k, k \geq 1\}$, for any $\varepsilon \in (0, \frac{1}{2}] \setminus E$, we obtained $\mathcal{N}_\xi(\varepsilon) = N(\varepsilon)$.

Proving the convergence. We show that $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$ by checking that it satisfies the condition from Theorem 9, proved later in D. The proof of that result is completely separate so that there is no circular logic. We take $K = 1$ and for $\varepsilon > 0$, fix $k \geq 1$ such that $\varepsilon_k \leq \varepsilon/2$. We simply take one event $E_1 = \mathcal{E}_k$, and we use the centers $J = [N_k]$. For any $i > N_k$, letting $k_i \geq k$ such that $N_{k_i} < i \leq N_{k_i+1}$, we indeed have

$$\{X_i \neq X_1\} \subset \mathcal{E}_{k_i} \subset \mathcal{E}_k = E_1.$$

This ends the proof of the proposition. ■

Because of the necessary condition in Theorem 4, there is no loss of generality in assuming that for any $\varepsilon > 0$, the ε-covering number for $(\mathbb{N}, \xi)$ is finite: $\mathcal{N}_\xi(\varepsilon) < \infty$. On the other hand, Theorem 6 shows that the condition given in Eq (8), which we restate here for convenience,

$$\int_0^1 \mathcal{N}_\xi(\varepsilon)\mathrm{d}\varepsilon < \infty,$$

is a sufficient condition for $\Delta_n(\mu) \to 0$.

Proof of Theorem 6 We first note that this condition is equivalent to $\sum_{k\geq 0} 2^{-k}\mathcal{N}_\xi(2^{-k}) < \infty$. Indeed, $\mathcal{N}_\xi(\cdot)$ is non-increasing, hence

$$\sum_{k\geq 0} 2^{-k-1}\mathcal{N}_\xi(2^{-k}) \leq \int_0^1 \mathcal{N}_\xi(\varepsilon)\mathrm{d}\varepsilon \leq \sum_{k\geq 1} 2^{-k}\mathcal{N}_\xi(2^{-k}).$$

To obtain our bounds on $\Delta_n(\mu)$, we will use chaining techniques. First, let $\mathcal{S}_\xi(\varepsilon)$ be an ε-covering of $(\mathbb{N}, \xi)$ with minimal cardinality $\mathcal{N}_\xi(\varepsilon)$. For any $k \geq 1$ and $i \in \mathcal{S}_\xi(2^{-k})$, we denote by $\hat{i}$ the ξ-nearest neighbor of i within $\mathcal{S}_\xi(2^{-k+1})$. In particular, by definition of $\mathcal{S}_\xi(2^{-k+1})$ we have $\xi(i, \hat{i}) \leq 2^{-k+1}$.

Fix $\varepsilon \in (0, 1]$ and consider $(X_i)_{i\geq 1} \sim \mu$. By hypothesis, there exists $k_\varepsilon \geq \log_2 \frac{1}{\varepsilon}$ such that

$$\sum_{k\geq k_\varepsilon} 2^{-k}\mathcal{N}_\xi(2^{-k}) \leq \varepsilon.$$

Then,

$$\begin{aligned}
\mathbb{P}\left(\exists i \in \bigcup_{k\geq k_\varepsilon} \mathcal{S}_\xi(2^{-k}), X_i \neq X_{\hat{i}}\right) &\leq \mathbb{E}\left[\left|\left\{i \in \bigcup_{k\geq k_\varepsilon} \mathcal{S}_\xi(2^{-k}) : X_i \neq X_{\hat{i}}\right\}\right|\right] \\
&= \sum_{k\geq k_\varepsilon} \sum_{i\in\mathcal{S}_\xi(2^{-k})} \mathbb{P}(X_i \neq X_{\hat{i}}) \\
&\leq \sum_{k\geq k_\varepsilon} 2^{-k}\mathcal{N}_\xi(2^{-k}) \leq \varepsilon.
\end{aligned}$$

Hence, if $\mathcal{E}_\varepsilon := \{\forall i \in \bigcup_{k\geq k_\varepsilon} \mathcal{S}_\xi(2^{-k}), X_i = X_{\hat{i}}\}$, we have $\mathbb{P}(\mathcal{E}_\varepsilon) \geq 1 - \varepsilon$.

For any $i \in \mathcal{S} := \bigcup_{k\geq 0} \mathcal{S}_\xi(2^{-k})$, let $k \geq 0$ be such that $i \in \mathcal{S}_\xi(2^{-k}) \setminus \mathcal{S}_\xi(2^{-k+1})$, with the convention $\mathcal{S}_\xi(2) = \emptyset$. We can then construct the sequence $i_k = i, i_{k-1}, \ldots, i_0$ such that $i_{p-1} = \hat{i}_p$.

We denote by $i(\varepsilon)$ the first element of this list within $\mathcal{S}_\varepsilon := \bigcup_{k \leq k_\varepsilon} \mathcal{S}_\xi(2^{-k})$, that is $i(\varepsilon) = i_{k_\varepsilon \wedge k}$. Note that by the triangle inequality,

$$\xi(i, i(\varepsilon)) \leq \xi(i_k, i_{k-1}) + \ldots + \xi(i_{k_\varepsilon \wedge k+1}, i_{k_\varepsilon \wedge k}) \leq \sum_{k-1 \leq l \leq k_\varepsilon} 2^{-l} \leq 2^{-k_\varepsilon + 1} \leq 2\varepsilon.$$

Also, under $\mathcal{E}_\varepsilon$, for any $i \in \mathcal{S}$, one has $X_i = X_{i(\varepsilon)}$.

We now focus on indices in $i \notin \mathcal{S}$ and aim to prove an equivalent equation. Fix any $k \geq 0$, we denote by i_k the ξ-nearest neighbor of i within $\mathcal{S}_\xi(2^{-k})$. Because $\mathcal{S}_\varepsilon$ is finite, we can fix some element that we denote $i(\varepsilon)$ that appears infinitely often in the sequence $(i_k(\varepsilon))_{k \geq 0}$. In particular, for any $k \geq 0$ such that $i_k(\varepsilon) = i(\varepsilon)$, we have

$$\xi(i, i(\varepsilon)) \leq \xi(i, i_k) + \xi(i_k, i_k(\varepsilon)) \leq 2^{-k} + 2\varepsilon.$$

Because this holds for an infinite number of indices $k \geq 0$, this shows that $\xi(i, i(\varepsilon)) \leq 2\varepsilon$. Next, since $\mathbb{P}(X_i \neq X_{i(k)}) \leq 2^{-k}$ for $k \geq 0$, which forms a summable sequence, by the Borel-Cantelli lemma and the union bound, the following event

$$\mathcal{F} := \{\forall i \notin \mathcal{S}, \exists \hat{k}_i \geq 0, \forall k \geq \hat{k}_i, X_i = X_{i(k)}\},$$

has probability one. As a result, on $\mathcal{E}_\varepsilon \cap \mathcal{F}$, the sequence $(X_{i(k)})_{k \geq 0}$ is equal to X_i for k large enough but also contains an infinite number of times the value $X_{i(\varepsilon)}$. Hence $X_i = X_{i(\varepsilon)}$. In summary, we obtained

$$\mathcal{E}_\varepsilon \cap \mathcal{F} \subset \{\forall i \geq 1, X_i = X_{i(\varepsilon)}\},$$

and for all $i \geq 1$, $\mathbb{P}(X_i \neq X_{i(\varepsilon)}) = \xi(i, i(\varepsilon)) \leq 2\varepsilon$.

Now consider iid samples $(X_i^{(n)})_{i \geq 1} \sim \mu$ for $n \geq 1$ and denote by $\mathcal{E}_\varepsilon^{(n)}$ and $\mathcal{F}^{(n)}$ the corresponding event. In particular, $(\mathbb{1}(\mathcal{E}_\varepsilon^{(n)} \cap \mathcal{F}^{(n)}))_{n \geq 1}$ is an iid Bernoulli sequence of parameter $\mathbb{P}(\mathcal{E})$. For any $i \geq 1$,

$$\begin{aligned}
|\hat{p}_i - p_i| &\leq |\hat{p}_i - \hat{p}_{i(\varepsilon)}| + |\hat{p}_{i(\varepsilon)} - p_{i(\varepsilon)}| + |p_{i(\varepsilon)} - p_i| \\
&\leq \frac{1}{n} \sum_{m=1}^{n} \mathbb{1}((\mathcal{E}_\varepsilon^{(m)} \cap \mathcal{F}^{(m)})^c) + \sup_{j \in \mathcal{S}_\varepsilon} |\hat{p}_j - p_j| + \mathbb{P}(X_i \neq X_{i(\varepsilon)}) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}((\mathcal{E}_\varepsilon^{(m)} \cap \mathcal{F}^{(m)})^c) + \sup_{j \in \mathcal{S}_\varepsilon} |\hat{p}_j - p_j| + 2\varepsilon.
\end{aligned}$$

As a result,

$$\begin{aligned}
\Delta_n(\mu) = \mathbb{E} \sup_{i \geq 1} |\hat{p}_i - p_i| &\leq (1 - \mathbb{P}(\mathcal{E})) + \mathbb{E} \sup_{j \in \mathcal{S}_\varepsilon} |\hat{p}_j - p_j| + 2\varepsilon \\
&\leq \mathbb{E} \sup_{j \in \mathcal{S}_\varepsilon} |\hat{p}_j - p_j| + 3\varepsilon.
\end{aligned}$$

In particular, this gives $\limsup_{n \to \infty} \Delta_n(\mu) \leq 3\varepsilon$ because $\mathcal{S}_\varepsilon$ is finite. This holds for any ε, which ends the proof that $\Delta_n(\mu) \to 0$ as $n \to \infty$.

Additionally, the previous equation holds for any ε, hence

$$\Delta_n(\mu) \leq \inf_{\varepsilon \in (0,1]} 3\varepsilon + \mathbb{E} \sup_{j \in \mathcal{S}_\varepsilon} |\hat{p}[j] - p_j|.$$

The right-most term can be upper bounded using the upper bound on maximum empirical mean deviations for general distributions on $\{0,1\}^{\mathbb{N}}$ from Blanchard and Voráček (2024, Corollary 3). Precisely, we need to order the elements $j \in \mathcal{S}_\varepsilon$ by decreasing order of $p_j \wedge (1-p_j)$. The worst case upper bound is achieved when all these mean probabilities are equal to $\frac{1}{2}$. Hence, Blanchard and Voráček (2024, Corollary 3) yields

$$\mathbb{E}\sup_{j\in\mathcal{S}_\varepsilon}|\hat{p}[j]-p_j| \lesssim 1\wedge\left(\sqrt{\frac{\log(1+|\mathcal{S}_\varepsilon|)}{n}}+\frac{\log(1+|\mathcal{S}_\varepsilon|)}{n\log\left(2+\frac{2\log(1+|\mathcal{S}_\varepsilon|)}{n}\right)}\right) \asymp 1\wedge\sqrt{\frac{\log(1+|\mathcal{S}_\varepsilon|)}{n}}.$$

Putting the upper bounds together yields the following bound

$$\Delta_n(\mu) \lesssim \inf_{\varepsilon\in(0,1]}\varepsilon+\sqrt{\frac{\log(\mathcal{N}_\xi(\varepsilon)+1)}{n}} \leq \inf_{\varepsilon\in(0,1]}\varepsilon+\sqrt{\frac{\log(1+\frac{1}{\varepsilon}\int_0^1\mathcal{N}_\xi(\eta)d\eta)}{n}}$$

Recalling the notation $C_\mu := \int_0^1 \mathcal{N}_\xi(\eta)d\eta$, and using the value of $\varepsilon_{\mu,n} := 1\wedge\sqrt{\log(n(1+C_\mu))/n}$, we then obtain $\Delta_n(\mu)\leq 1$ if $\varepsilon_{\mu,n}=1$, or if $\varepsilon_n<1$,

$$\Delta_n(\mu) \lesssim \varepsilon_{\mu,n}+\sqrt{\frac{\log(1+\frac{1}{\varepsilon_{\mu,n}}C_\mu)}{n}} \asymp \sqrt{\frac{\log(n(1+C_\mu))}{n}}.$$

This ends the proof of the proposition. ■

We next show that this sufficient condition, Eq (8), is as tight as can be using the covering numbers $\mathcal{N}_\xi(\cdot)$. We recall that this cannot be a necessary condition in view of Theorem 3—instead, we show in Proposition 8 that if the covering numbers do not satisfy Eq (8), one can construct some distribution with (almost) the same covering numbers but for which the expected maximum deviation does not converge to 0.

Proof of Proposition 8 The proof has three steps, first we define the distribution μ, then we prove that its covering numbers coincide with $N(\cdot)$, then we show that $\Delta_n(\mu)\to\frac{1}{2}$.

Constructing the distribution μ. Fix the non-increasing function $N:(0,1]\to\mathbb{N}$. We use similar notations as in the proof of Proposition 5. Since we will only focus on covering numbers for $\varepsilon\in(0,\frac{1}{2}]$ anyways, without loss of generality we suppose that $N(\varepsilon)=1$ for $\varepsilon\in(\frac{1}{2},1]$. Given such a function N, we start by constructing an equidistant directed tree (with edge lengths) representing the function. Note that because N is non-increasing, it is discontinuous on a countable (potentially finite) number of points $(\varepsilon_k)_{k\geq 1}$, which we ordered by decreasing order, that is $0<\varepsilon_{k+1}<\varepsilon_k\leq\frac{1}{2}$ for all $k\geq 1$. For convenience, let $\varepsilon_0=\frac{1}{2}$. Since N takes values only on integers $\mathbb{N}$, we can also define the sequence $(N_k)_{k\geq 1}$ such that N_k is the value taken by N on the interval $(\varepsilon_{k-1},\varepsilon_k)$, with the convention $N_1=1$ in the specific case when $\varepsilon_0=\varepsilon_1$.

We construct the tree by recursion, starting for $k=0$ with only a root denoted $v(0,1)$. For context, inner nodes will be denoted $v(k,p)$ where k will correspond to level ε_k and p will correspond to the index of the node by order of construction. Now suppose that we have constructed the tree up to level $k\geq 0$ and that we have constructed a total of N_k nodes. At level $k+1$, we construct $N_{k+1}-N_k$ new nodes. If for $k\in[N_{k+1}-N_k]$, we link the node $v(k+1,N_k+l)$ to some node $v(k',p')$, the length of the edge is set to $\varepsilon_{k'}-\varepsilon_{k+1}$. Deciding of which node to link to the new nodes

at level $k+1$ is done in a specific manner to balance the construction of the overall tree. A formal construction of the tree is given in the pseudo-code Algorithm 1. Intuitively, the construction of the tree emulates the construction of a full binary tree: if we had $N_{k+1} - N_k = 1$ for all $k \geq 1$, the output tree would exactly be a binary tree that is constructed layer by layer in order. Because the jumps $N_{k+1} - N_k$ may be larger, the tree is instead the binary tree where some edges are collapsed. To keep at all times a balance in the binary tree, splits are added according to a fractal manner. For layer r, we split 2^r edges which are denoted $e(r,s)$ for $s \in \{0, 1, \ldots, 2^r - 1\}$, in the following order:

$$\begin{aligned}&\mathrm{Order}(0) := (0),\\ &\mathrm{Order}(r) := (2i, i \in \mathrm{Order}(r-1)) \cup (2i+1, i \in \mathrm{Order}(r-1)), \quad r \geq 1.\end{aligned}$$

For instance, $\mathrm{Order}(1) = (0,1)$, $\mathrm{Order}(2) = (0,2,1,3)$, and $\mathrm{Order}(3) = (0,4,2,6,1,5,3,7)$. A visualization of the trees constructed for two different covering number functions $N(\cdot)$ are given in Figure 1, one for the simpler case when $N_{k+1} - N_k = 1$ for all $k \geq 1$ and one for the general case.

Algorithm 1: Constructing the tree skeleton for the distribution in Proposition 8

Data: $(\varepsilon_k)_{k\geq 1}, (N_k)_{k\geq 1}$
Result: An equidistant skeleton tree $\mathcal{T}$
Initialize $\mathcal{T}$ as a root $v(0,1)$ at level $\frac{1}{2}$ with an exiting edge denoted $e(0,0)$
$k \leftarrow 1$, $n \leftarrow 1$ and $m \leftarrow 1$
for $r \geq 0$ **do**
  for $s \in \mathrm{Order}(r)$ **do**
    Split edge $e(r,s)$ in two at level ε_k. That is:
    if *the top end node of edge* $e(r,s)$ *is a node* $v(k',n')$ *with* $k' < k$ **then**
      Create a node $v(k, n+1)$ at level ε_k to end edge $e(r,s)$: $e(r,s)$ has length $\varepsilon_{k'} - \varepsilon_k$
      $n \leftarrow n+1$
      Create two edges $e(r+1, 2s), e(r+1, 2s+1)$ exiting from node $v(k, n+1)$
    else
      Delete edge $e(r,s)$ and create two edges $e(r+1,2s), e(r+1,2s+1)$ exiting from node $v(k',n') = v(k,n')$
    end
    $m \leftarrow m+1$
    if $m = N_{k+1}$ **then** $k \leftarrow k+1$;
  end
end

Note that at every level $\varepsilon \in (0, \frac{1}{2}] \setminus \{\varepsilon_k, k \geq 1\}$, the tree has $N(\varepsilon)$ edges. Because of the hypothesis $\int_0^{1/2} N(\varepsilon)\mathrm{d}\varepsilon = \infty$, we have $N(\varepsilon) \to \infty$ as $n \to \infty$ or equivalently $N_k \to \infty$ as $k \to \infty$ In particular, the tree output by Algorithm 1 has an infinite number of edges. We denote by $\mathcal{L}$ the set of leaves of the tree, which corresponds to sequences of nodes $(v_i = v(k_i, n_i))_{i\geq 1}$ that start from the root $v(0,1)$, follow edges of the tree, and go down in the tree, that is the sequence $(k_i)_{i\geq 1}$ is increasing. We note that because of the breadth-first search procedure to construct the tree, all leaves $l = (v_i)_{i\geq 1}$ contain an infinite number of edges—that is, no path ended after a finite number of edges. The tree naturally induces a distance ρ on leaves $\mathcal{L}$ such that two leaves

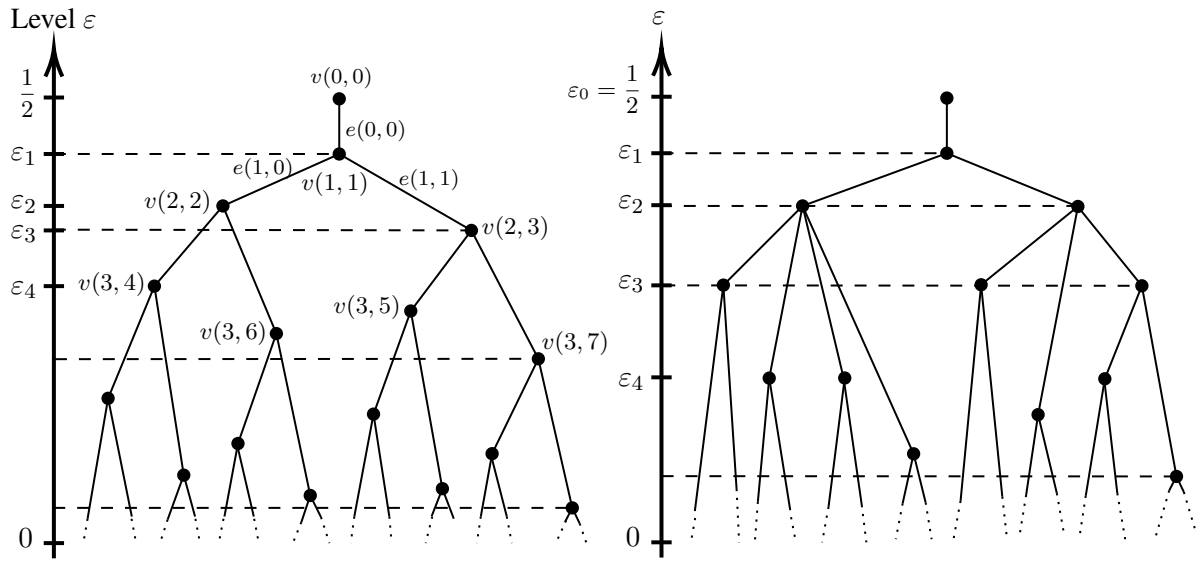

Figure 1: Two examples of skeleton trees constructed by Algorithm 1. The entries of the distribution μ for Proposition 8 are associated to leaves (or rather paths) of this infinite tree so that the distance ξ between leaves coincides with the natural tree metric (up to a constant factor 2). On the left we represent the simpler case when $N_{k+1} - N_k = 1$ for $k \geq 1$, that is, covering numbers $N(\varepsilon)$ grow by one at a time as $\varepsilon \to 0$. In this case, the constructed tree is exactly a binary tree, constructed according to the exact ordering given by $Order(l)$ for $l \geq 1$. On the right, we represent a general case when covering numbers can grow via jumps $N_{k+1} - N_k \geq 1$. In the specific example, we have $(N_i, i \leq 7) = (1, 2, 7, 10, 13, 14, 15)$. Although the tree is not formally a complete binary tree, the ordering choice balances all subtrees evenly. We represent with dashed lines, all levels ε which complete a layer of the constructed binary tree.

$l_u = (v(k_i^{(u)}, n_i^{(u)}))_{i \geq 1}$ for $u \in \{1, 2\}$ have distance

$$d(l_1, l_2) := \varepsilon_{k_i}, \quad i = \max\{j : n_j^{(1)} = n_j^{(2)}\}.$$

Our goal is to use the tree to construct a binary stochastic process $(X_l)_{l \in \mathcal{L}}$ on $\mathcal{L}$, for which the induced metric $\xi(l_1, l_2) = \mathbb{P}(X_{l_1} \neq X_{l_2})$ coincides with d. We start by constructing a distribution on the inner nodes of the tree recursively. First, let $(U_n)_{n \geq 1} \overset{\text{iid}}{\sim} \mathcal{U}([0, 1])$ be a sequence of iid uniform random variables on $[0, 1]$ and $(Z_n)_{n \geq 1} \overset{i.i.d}{\sim} \text{Bernoulli}(\frac{1}{2})$ be an independent iid sequence of Bernoulli random variables. At the root $r = v(0, 1)$, we let $Y_r = \mathbb{1}[Z_1 \geq \frac{1}{2}]$. Next, for any node $v = v(k, n)$ that has parent $v' = v(k', n')$ (they are linked by an edge and $k' < k$), we pose

$$Y_v := \mathbb{1}[U_n \geq \eta_v] Y_{v'} + \mathbb{1}[U_n < \eta_v] Z_n, \quad \eta_v := \frac{\sqrt{1 - 2\varepsilon_k} - \sqrt{1 - 2\varepsilon_{k'}}}{\sqrt{1 - 2\varepsilon_k}}.$$

We now define the binary stochastic process on $\mathcal{L}$ as follows. For any leaf $l = (v_i)_{i\geq 1}$,

$$X_l := \begin{cases} \lim_{i\to\infty} Y_{v_i} & \liminf_{i\to\infty} Y_{v_i} = \limsup_{i\to\infty} Y_{v_i}, \\ 0 & \text{otherwise.} \end{cases}$$

We can now check that on $\mathcal{L}$ introduced by the distribution of $(X_l)_{l\in\mathcal{L}}$ coincides with d. For a leaf $l = (v_i = v(k_i, n_i))_{i\geq 1}$, we first show that $\mathbb{P}(X_l \neq Y_{v_i}) = \frac{1}{2}(1 - \sqrt{1-2\varepsilon_{k_i}})$. Indeed, recalling that the sequence $(Z_n)_{n\geq 1}$ is iid Bernoulli($\frac{1}{2}$), for any $j > i$, we can write

$$Y_{v_j} = \mathbb{1}[\forall i+1 \leq p \leq j : U_{n_p} \geq \eta_{v_p}]Y_{v_i} + \mathbb{1}[\exists i+1 \leq p \leq j : U_{n_p} < \eta_{v_p}]A_j, \tag{13}$$

where $A_j \sim$ Bernoulli($\frac{1}{2}$) is a function of the variables Z_{n_p} for $i+1 \leq p \leq j$. Next, observe that since for $i \to \infty$, one has $\eta_{v_i} \sim \varepsilon_{k_{i-1}} - \varepsilon_{k_i}$, then $\sum_{i\geq 1} \eta_{n_i} < \infty$. In particular, the Borel-Cantelli lemma implies that on an almost sure event $\mathcal{E}_l$, for sufficiently large index i, one has $U_{n_i} \geq \eta_{v_i}$. Hence, on $\mathcal{E}_l$ the sequence $(Y_{v_i})_{i\geq 1}$ is either finite or converges, that is

$$\mathcal{E}_l \subset \{X_l = \lim_{i\to\infty} Y_{v_i}\}.$$

Hence, using Eq (13), we can write with $\mathcal{F}_l(i) := \{\exists j \geq i+1, U_{n_j} \geq \eta_{v_j}\}$,

$$X_l = \mathbb{1}_{\mathcal{E}_l}\left(\mathbb{1}_{\mathcal{F}_l(i)^c} Y_{v_i} + \mathbb{1}_{\mathcal{F}_l(i)} B_l(i)\right),$$

where $B_l \sim$ Bernoulli($\frac{1}{2}$) is a function of the variables Z_{n_j} for $j \geq i+1$. Also, note that

$$\mathbb{P}[\mathcal{F}_l(i)] = \mathbb{P}[\exists j \geq i+1, U_{n_j} \geq \eta_{v_j}] = 1 - \prod_{j\geq i+1}(1-\eta_{v_p}) = 1 - \sqrt{1-2\varepsilon_{k_i}},$$

where in the last equality, we used a telescoping argument. To summarize, we showed that X_l coincides with Y_{v_i} except on an (independent) event of probability $1 - \sqrt{1-2\varepsilon_{n_i}}$ on which it is an independent Bernoulli $B_l(i)$. We are now ready to compute the $\mathbb{P}(X_{l_1} \neq X_{l_2})$ for two leaves $l_1, l_2 \in \mathcal{L}$. Let $v = v(k_i, n_i)$ be the first node for which the paths $l_1 = (v_i^{(1)})_{i\geq 1}$ and $l_2 = (v_i^{(2)})_{i\geq 1}$ differ, that is

$$i = \max\{j \geq 1 : v_j^{(1)} = v_j^{(2)}\}.$$

Then, using the previous characterization of X_{l_1} and X_{l_2}, we obtain

$$\begin{aligned} \xi(l_1, l_2) = \mathbb{P}(X_{l_1} \neq X_{l_2}) &= \frac{1}{2}\mathbb{P}\left(\mathcal{F}_{l_1}(i) \cup \mathcal{F}_{l_2}(i)\right) \\ &= \frac{1}{2}\left(2(1-\sqrt{1-2\varepsilon_{k_i}}) - (1-\sqrt{1-2\varepsilon_{k_i}})^2\right) \\ &= \varepsilon_{k_i} = d(l_1, l_2). \end{aligned}$$

This ends the proof that ξ can directly be computed as the tree distance (up to a factor 2). As defined currently, the space of leaves $\mathcal{L}$ can potentially be uncountable. Because we need to construct a distribution on $\{0,1\}^{\mathbb{N}}$, we restrict ourselves to a countable subset of leaves $\mathcal{Q}$, one at most for each inner node. Precisely, to any node $v = v(k,n)$ we associate the leaf $l(v)$ which arrives at $v(k,n)$ and from there always selects the left-most edge (first added in the FIFO pile) at any intersection. We then pose $\mathcal{Q} = \{l(v), \text{ nodes } v\}$. We recall that there are countably-many nodes, hence $\mathcal{Q}$ is countable. The distribution μ is defined as the distribution of $(X_l)_{l\in\mathcal{Q}}$. We note that because $\mathcal{Q}$ is now countable, the event $\mathcal{E} := \bigcap_{l\in\mathcal{Q}} \mathcal{E}_l$ has full probability. Hence with probability one,

$$\forall l = (v_i)_{i\geq 1} \in \mathcal{Q}, X_l = \lim_{i\to\infty} Y_{v_i}.$$

Computing the covering numbers of ξ. Let $E = \{\varepsilon_k, k \geq 1\}$ and fix $\varepsilon \in (0, \frac{1}{2}] \setminus E$. By construction of the skeleton tree, at level ε there are exactly $N(\varepsilon)$ edges $f_1, \ldots, f_{N(\varepsilon)})$. For each of these edges say f_p for $p \in [N(\varepsilon)]$ if its end nodes are u_p, v_p with u_p being the parent of v_p (that is, has lower index n and number k as well), we now show that $\{l(v_p), p \in [N(\varepsilon)\}$ is an ε-covering of $(\mathcal{Q}, \xi)$. Indeed, for $l = (w_i)_{i\geq 1} \in \mathcal{Q}$, one of the nodes on the corresponding leaf path must belong to $\{v_p, p \in [N(\varepsilon)\}$. Hence, for some $i \geq 1$ and $p \in [N(\varepsilon)]$, we have $w_i = v_p := v(k_p, n_p)$. Then, we directly have

$$\xi(l, l(v_p)) = d(l, l(v_p)) \leq \varepsilon_{k_p} < \varepsilon.$$

Hence, we obtained $\mathcal{N}_\xi(\varepsilon) \leq N(\varepsilon)$.

We now turn to the lower bound. We will show that to ε-cover the set of leaves $\{l(v_p), p \in [N(\varepsilon)]\}$ one needs $N(\varepsilon)$ elements. Suppose that this is not the case, then we have a leaf $l(v)$ such that $\xi(l(v), l(v_p)), \xi(l(v), l(v_q)) \leq \varepsilon$. In particular, $l(v)$ and $l(v_p)$ share the same path until length ε, hence the path of $l(v)$ contains edge f_p. By symmetry, this shows it also contains $f_{p'}$ which is impossible because they are at same level (and paths only go "down"). This ends the proof that $\mathcal{N}_\xi(\varepsilon) \geq N(\varepsilon)$.

In summary, we have that for any $\varepsilon \in (0, \frac{1}{2}] \setminus E$, $\mathcal{N}_\xi(\varepsilon) = N(\varepsilon)$. Additionally, the same arguments show that for any $k \geq 1$, one has $\mathcal{N}_\xi(\varepsilon_k) = N_{k-1}$: we again look at level ε_k of the tree. If this cuts edges, we proceed similarly as above. However, it will also be the case that at this level are nodes $v = v(k, n)$. These are then also included to construct a set of N_{k-1} nodes $v_1, \ldots, v_{N_{k-1}}$ at level ε_k or below. The same proof shows that they ε_k-cover the space $\mathcal{Q}$, and that $\{l(v_p), p \in [N_{k-1}]\}$ requires at least N_{k-1} elements to be ε_k-covered. This shows in particular that $\mathcal{N}_\xi(\cdot)$ is right-continuous.

Estimating $\Delta_n(\mu)$. In this last step, we show that $\Delta_n(\mu) \to \frac{1}{2}$. First, note that by construction and from the above estimates, for any $l \in \mathcal{Q}$, we have $X_l \sim \text{Bernoulli}(\frac{1}{2})$ so that $p_l = \mathbb{E}[X_l] = \frac{1}{2}$.

We recall that the construction of the skeleton tree emulates binary tree that is constructed layer by layer r. Consider the state of the tree at the very beginning of the construction of the rth layer for some fixed $r \geq 1$. At this point, there are 2^r edges $e(r, s)$ for $s \in \{0, 1, \ldots, 2^r - 1\}$ and we can consider the corresponding subtrees at each of these edges $e(r, s)$, which correspond to the set of nodes and edges descendants from $e(r, s)$ (in the case when an edge e was removed and replaced by two new edges, these new edges are also considered descendants of e), which we will denote $\mathcal{T}(r, s)$. The main interest of the fractal order for the construction of the tree is that starting from the rth layer and for all next layers, we are adding a single edge to $\mathcal{T}(r, s)$ in the order of $s \in \text{Order})$ then the process is repeated indefinitely. As a result, the subtrees $\mathcal{T}(r, s)$ for $s \in \{0, \ldots, 2^r - 1\}$ are always filled equally, up to at most one edge. The property that the trees are filled evenly is crucial for the proof.

In the rest of the proof, we denote by $\text{parent}(v)$ the parent node of any node v. For any fixed $s \in \{0, \ldots, 2^r - 1\}$, we define

$$A(r, s) := \sum_{v=v(k,n)\in\mathcal{T}(r,s)} \varepsilon_{k(\text{parent}(v))} - \varepsilon_k,$$

where the parent node of v is written as $\text{parent}(v) = v(k(\text{parent}(v)), n')$. Our goal is to show that the above quantity is infinite. Let k_r be the value of the level k at the beginning of the construction of the rth layer in Algorithm 1. Without loss of generality, we will assume that $\{\varepsilon_k, k \geq 1\} \cap \{2^{-t}, t \geq 1\} = \emptyset$. If that is not the case, we can replace all terms 2^{-t} with some terms $c2^{-t}$ for some constant

$c \in [\frac{1}{2}, 1]$ since there will exist such a constant c for which $\{\varepsilon_k, k \geq 1\} \cap \{c2^{-t}, t \geq 1\} = \emptyset$. For any $t \geq t_r := \lceil \log_2 1/\varepsilon_{k_r} \rceil$, we let $M(r, s; t)$ be the number of edges within $\mathcal{T}(r, s)$ at level 2^{-t}. Then, we can check that

$$A(r, s) \geq \sum_{t \geq t_r} M(r, s; t)(2^{-t} - 2^{-t-1}) = \frac{1}{2} \sum_{t \geq t_r} 2^{-t} M(r, s; t). \tag{14}$$

Now we recall that the number of edges in the complete tree $\mathcal{T}$ at level 2^{-t} is precisely $\mathcal{N}_\xi(2^{-t}) = N(2^{-t})$, where we used the result from the previous steps on covering numbers of ξ. As a result, we have that for $t \geq t_r$,

$$\sum_{s'=0}^{2^r-1} M(r, s'; t) = N(2^{-t}).$$

Now from the above discussion on the evenness of the tree construction, the number of edges at any level $\varepsilon \leq \varepsilon_{k_r}$ for the subtrees $\mathcal{T}(r, s)$ can differ at most by one. Hence, we obtain,

$$2^r (M(r, s; t) + 1) \geq \sum_{s'=0}^{2^r-1} M(r, s'; t) = N(2^{-t}).$$

Plugging this into Eq (14) yields

$$A(r, s) \geq \frac{1}{2^{r+1}} \sum_{t \geq t_r} 2^{-t} N(2^{-t}) - 2^{-t_r} = \infty.$$

The last inequality use the hypothesis $\int_0^{1/2} N(\varepsilon) \mathrm{d}\varepsilon = \infty$ and the fact that N is non-increasing, so that this condition is equivalent to $\sum_{t \geq 1} 2^{-t} N(2^{-t}) = \infty$. As a result, this shows that $A(r, s) = \infty$. Now let $v_{r,s}$ be the top end node of edge $e(r, s)$, which is intuitively the "root" of $\mathcal{T}(r, s)$. We obtained

$$\begin{aligned}
\mathbb{P}(\exists v \in \mathcal{T}(r, s), Y_v \neq Y_{v_{r,s}}) &\geq \mathbb{P}(\exists v \in \mathcal{T}(r, s), Y_v \neq Y_{\mathrm{parent}(v)}) \\
&= 1 - \prod_{v \in \mathcal{T}(r,s)} (1 - \mathbb{P}(Y_v \neq Y_{\mathrm{parent}(v)})) \\
&= 1 - \prod_{v \in \mathcal{T}(r,s)} \left(1 - \frac{\eta_v}{2}\right) \geq 1 - \exp\left(-\frac{1}{2} \sum_{v \in \mathcal{T}(r,s)} \eta_v\right).
\end{aligned}$$

Now as the index n of $v = v(k, n)$ grows to infinity, we have $\eta_v \sim \varepsilon_{k(\mathrm{parent}(v))} - \varepsilon_k$ because $\varepsilon_{k(\mathrm{parent}(v))}, \varepsilon_k \to 0$. The fact that $A(r, s) = \infty$ then shows that $\sum_{v \in \mathcal{T}(r,s)} \eta_v = \infty$. Hence, we showed that for any $r \geq 1$ and $s \in \{0, \ldots, 2^r - 1\}$,

$$\mathbb{P}(\exists v \in \mathcal{T}(r, s), Y_v \neq Y_{v_{r,s}}) = 1.$$

We denote by $\mathcal{G}(r, s)$ the above event. Hence $\mathcal{G} = \bigcap_{r \geq 1} \bigcap_{0 \leq s \leq 2^r - 1} \mathcal{G}(r, s)$ has probability one. The main property of $\mathcal{G}$ is that on this almost sure event, for any node v of the tree, there exists a descendant node v' that disagrees in the sense $Y_v \neq Y_{v'}$.

We are now ready to show that $\Delta_n(\mu)$ does not decay to 0 as $n \to \infty$. Fix $n \geq 1$ and $\delta > 0$. We will indicate that we consider the ith iid sample of a certain random variable (or event) V with an

exponent as in $V^{(i)}$. We construct a sequence of nodes $(\hat{v}_i)_{i\geq 0}$ recursively. We let $\hat{v}_0 := v_\delta$ where $v_\delta = v(k_\delta, n_\delta)$ is an arbitrary node for which ε_{k_δ} is sufficiently small such that

$$\frac{n}{2}\left(1-\sqrt{1-2\varepsilon_{k_\delta}}\right) < \delta. \tag{15}$$

Next, for $i = 1$, we define

$$\hat{v}_i = \begin{cases} v_\delta & \text{if } \mathcal{G}^{(i)} \text{ is not satisfied,} \\ \hat{v} & \text{otherwise, and } \hat{v} = \arg\min\{n : v = v(k,n) \text{ is a descendant of } \hat{v}_{i-1} \text{ s.t. } Y_v = 1\}. \end{cases}$$

On the almost sure event $\bigcap_{i\geq 1}\mathcal{G}^{(i)}$, this constructs a sequence of nodes descendants from each other and such that we have

$$Y^{(i)}_{\hat{v}_i} = 1, \quad i \geq 1.$$

Our candidates for variables whose empirical mean deviates highly from the mean $\frac{1}{2}$ will be the leaves $l(\hat{v}_i)$ for $i \geq 1$. Importantly, having constructed $\hat{v}_{i-1}$, the construction of $\hat{v}_i = v(\hat{k}_i, \hat{n}_i)$ can be done completely independently from all the variables $U^{(i)}_n, Z^{(i)}_n$ where $n > \hat{n}_i$. This is because we can simply generate the variables $Y^{(i)}_v$ for nodes v with index $n \in \{\hat{n}_{i-1}, \hat{n}_{i-1}+1, \ldots\}$ and stop whenever the conditions $Y^{(i)}_v = 1$ and v is a descendant of $\hat{v}_{i-1}$ are met.

We now reason conditionally on $\bigcap_{i\geq 1}\mathcal{G}^{(i)}$ and $(\hat{v}_j)_{j\leq i}$. Note that up to this conditioning, the variables $(\hat{v}_j)_{j>i}$ only depend on the iid samples of the distribution with index $j > i$. In particular, for any $j \geq i$, this shows that all variables use to define $X_{l(\hat{v}_j)}$ starting from $Y_{\hat{v}_i}$ are still all distributed according to their distribution without conditioning. Precisely, write $\hat{v}_i = v(\hat{k}_i, \hat{n}_i)$. Conditioned on $\bigcap_{i\geq 1}\mathcal{G}^{(i)}$ and $(\hat{v}_j)_{j\geq 1}$, all variables $U^{(i)}_n$ and $Z^{(i)}_n$ for $n > \hat{n}_i$ and $i \geq 1$ are still all independent and distributed as $\mathcal{U}([0,1])$ and $\mathrm{Bernoulli}(\frac{1}{2})$ respectively. In particular, for a fixed $n \geq 1$, conditionally on $\bigcap_{i\geq 1}\mathcal{G}^{(i)}$ and $(\hat{v}_j)_{j\geq 1}$, we have

$$\left(\mathbb{1}\left[X^{(i)}_{l(\hat{v}_n)} \neq 1\right]\right)_{i\in[n]} = \left(\mathbb{1}\left[X^{(i)}_{l(\hat{v}_n)} \neq Y^{(i)}_{\hat{v}_i}\right]\right)_{i\in[n]} \sim \bigotimes_{i\in[n]} \mathrm{Bernoulli}\left(\frac{1}{2}\left(1-\sqrt{1-2\varepsilon_{\hat{k}_i}}\right)\right).$$

As a result,

$$\begin{aligned}\mathbb{P}\left[\exists i \in [n], X^{(i)}_{l(\hat{v}_n)} \neq 1 \mid \bigcap_{i\geq 1}\mathcal{G}^{(i)}, (\hat{v}_j)_{j\geq 1}\right] &\leq \sum_{i\in[n]} \frac{1}{2}\left(1-\sqrt{1-2\varepsilon_{\hat{k}_i}}\right) \\ &\leq \frac{n}{2}\left(1-\sqrt{1-2\varepsilon_{k_\delta}}\right) < \delta.\end{aligned}$$

In the last inequality, we used Eq (15). Hence,

$$\Delta_n(\mu) \geq \mathbb{E}\left[\left|\hat{p}_{l(\hat{v}_n)} - \frac{1}{2}\right| \mid \bigcap_{i\geq 1}\mathcal{G}^{(i)}\right] \geq \frac{1}{2}\mathbb{P}\left[\forall i \in [n], X^{(i)}_{l(\hat{v}_n)} = 1 \mid \bigcap_{i\geq 1}\mathcal{G}^{(i)}\right] \geq \frac{1-\delta}{2}.$$

This holds for any $\delta > 0$. Thus, we showed that $\Delta_n(\mu) \geq \frac{1}{2}$. Also, we clearly have $\|\hat{p} - \frac{1}{2}\|_\infty \leq \frac{1}{2}$ since the empirical means lie in $[0,1]$. This shows that

$$\Delta_n(\mu) = \frac{1}{2}, \quad n \geq 1,$$

and ends the proof of the result. ■

The proof of the previous result introduces a tree structure for the entries of μ. In order to have as much deviations as possible (so that $\Delta_n(\mu) = \frac{1}{2}$ for $n \geq 1$, this tree was constructed using a "wide" full binary tree (see Figure 1). As a remark, we can compare this to the tree generated by the distribution from Proposition 5. In that result, the goal is instead to construct distributions with large covering numbers, but that still have $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$. As a result, the corresponding tree is as "thin" as possible, as represented in Figure 2.

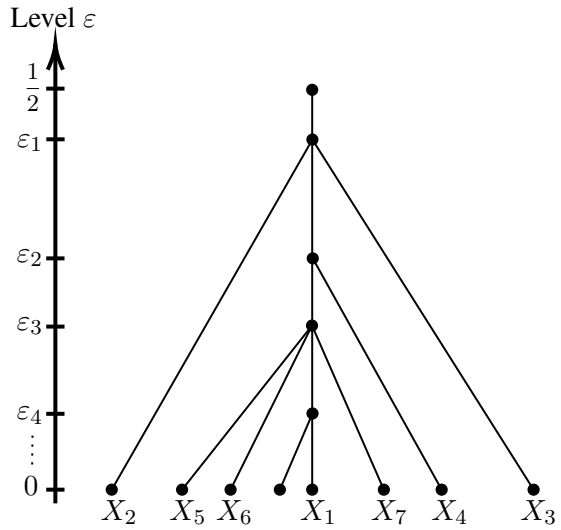

Figure 2: Tree corresponding to the distribution μ constructed in Proposition 5. As in Figure 1, the leaves of the tree represent the entries of the distribution, and the distance ξ coincides exactly with the natural tree metric (up to a constant factor 2). In this example, the covering numbers are $(N_k, k \leq 5) = (1, 3, 4, 7, 8)$.

Appendix C. Upper bound on $\Delta_n(\mu)$ via subgaussian differences

For $i, j \in \mathbb{N}$, let us define $r_{ij} = \mathbb{E}[X_i X_j]$. We claim the following bound on the moment generating function of the difference of correlated Bernoullis.

Lemma 18

$$\mathbb{E}\exp\left\{t[(X_i - p_i) - (X_j - p_j)\right\} \quad \leq \quad \exp\left(\frac{t^2}{\log\frac{2}{p_i+p_j-2r_{ij}}}\right), \qquad t \geq 0.$$

Proof Consider the following functions.

$$f(x) \quad := \quad \log\left((p_1 - p_{12})\,\mathrm{e}^{(-p_1+p_2+1)x} + (p_2 - p_{12})\,\mathrm{e}^{(-p_1+p_2-1)x} + (-p_1 - p_2 + 2p_{12} + 1)\,\mathrm{e}^{(p_2-p_1)x}\right),$$

$$g(x) \quad := \quad \frac{x^2}{\log\left(\frac{2}{p_1+p_2-2p_{12}}\right)}.$$

We would like to show $f(x) \le g(x)$ for all $p_1, p_2 \in [0,1]$, $p_{12} \in [0 \vee p_1 + p_2 - 1, p_1 \wedge p_2]$ and $x \in \mathbb{R}$. Note that, by symmetry, it is enough to consider $x \ge 0$.
Re-parametrizing $p_1 = \frac{1}{2}(a - b + 2p_{12})$ and $p_2 = \frac{1}{2}(a + b + 2p_{12})$, we have

$$\begin{aligned} f(x) &:= \log\left(\frac{1}{2}\mathrm{e}^{(b-1)x}\left(a\left(\mathrm{e}^x - 1\right)^2 - b\mathrm{e}^{2x} + b + 2\mathrm{e}^x\right)\right), \\ g(x) &:= \frac{x^2}{\log\left(\frac{2}{a}\right)}. \end{aligned}$$

Note that $g(0) - f(0) = 0$. If we show that $\frac{\partial}{\partial x}(g - f) = 0$ for $1 \ge a \ge |b|, x \ge 0$, we are done. First, we multiply the first derivative by the non-negative $\log\left(\frac{2}{a}\right)\left(aw^2 - b(w+1)^2 + b + 2(w+1)\right)$, and then change variables $x := \log(w+1)$.

$$\begin{aligned} H &:= \left(\frac{\partial}{\partial x}(g - f)\right)\log\left(\frac{2}{a}\right)\left(aw^2 - b(w+1)^2 + b + 2(w+1)\right) \\ &= 2\log(w+1)\left(w^2(a-b) - 2(b-1)w + 2\right) + w\log\left(\frac{2}{a}\right)(b(b(w+2)+w) - a(bw + w + 2)), \end{aligned}$$

whose non-negativity we must verify for $w \ge 0$.
We now consider two cases, first $0 \le w < 3$ and then $w \ge 3$.

Case 1: $0 \le w < 3$**.** It can be shown that the coefficient of $\log(1+w)$ is positive, because it is a quadratic polynomial in w with positive coefficients, so $\log(1+w)$ can be replaced with something smaller (in that range), such as $\frac{w}{\mathrm{e}}$. After doing that, we have

$$H \ge 2\mathrm{e}\left(b^2 - a\right)\log\left(\frac{2}{a}\right) + 2w^2(a-b) + w\left(-\mathrm{e}(b+1)(a-b)\log\left(\frac{2}{a}\right) - 4b + 4\right) + 4 =: I,$$

which is non-negative for $w = 0$, so we can show that the derivative (multiplied by a non-negative),

$$\left(\frac{\partial}{\partial w}I\right)\frac{\mathrm{e}}{w} = 4w(a-b) + \mathrm{e}(b+1)(b-a)\log\left(\frac{2}{a}\right) - 4b + 4,$$

is non-negative. The above expression is increasing in w, so the worst case is $w = 0$, in which case we have

$$4w(a-b) + \mathrm{e}(b+1)(b-a)\log\left(\frac{2}{a}\right) - 4b + 4 = -\mathrm{e}(b+1)(a-b)\log\left(\frac{2}{a}\right) - 4b + 4.$$

The right-hand side is a quadratic polynomial in b with the coefficient of b^2 being positive, therefore it is convex in b, so every tangent lies below the expression. We use the tangent at the point where b has the value a, and we get the following linear expression in b.

$$\begin{aligned} &-\mathrm{e}a^2\log\left(\frac{2}{a}\right) + b\left(\mathrm{e}(a+1)\log\left(\frac{2}{a}\right) - 4\right) - \mathrm{e}a\log\left(\frac{2}{a}\right) + 4 \\ &\quad = b\left(\mathrm{e}(a+1)\log\left(\frac{2}{a}\right) - 4\right) + \mathrm{e}(-a)(a+1)\log\left(\frac{2}{a}\right) + 4. \end{aligned}$$

We now consider the two endpoints of the above line, $b = a$ and $b = -a$, and show that both are positive. For $b = a$ the expression is just $4(1 - a)$ and therefore non-negative. At the second endpoint, $b = -a$, the expression has value of

$$b\left(\mathrm{e}(a+1)\log\left(\frac{2}{a}\right)-4\right)+\mathrm{e}(-a)(a+1)\log\left(\frac{2}{a}\right)+4=2\left(2-\mathrm{e}a\log\left(\frac{2}{a}\right)\right)(1+a).$$

The term $(1 + a)$ is positive, so we are left with $\left(2 - \mathrm{e}a \log\left(\frac{2}{a}\right)\right)$, which is convex since the second derivative is $\frac{\mathrm{e}}{a}$ and has a minimum at $a = \frac{2}{\mathrm{e}}$ with the value 0. The $0 \le w < 3$ part is done, we now move on to the second part, $w \ge 3$.

Case 2: $w \ge 3$**.** Recall that we have to show $H \ge 0$ for $a \ge |b|, x \ge 0$. H is convex in b, because

$$\frac{\partial^2}{\partial b^2}H=2w(w+2)\log\left(\frac{2}{a}\right)\ge 0,$$

so we can lower bound H by any of its tangent lines. The tangent line of H where $b = a$ is

$$\begin{aligned}T(w,a,b):=&w\log\left(\frac{2}{a}\right)\left(-\left(a^2(w+2)\right)+a(b(w+4)-w-2)+bw\right)\\&+2\log(w+1)\left(w^2(a-b)-2(b-1)w+2\right).\end{aligned}$$

Since we have a linear expression in b, we can check the endpoints of b, a and $-a$, and be done. Starting with $b = a$, we have

$$T(w,a,a)=2(a-1)aw\log\left(\frac{2}{a}\right)+4(-aw+w+1)\log(w+1).$$

Observe that $a\log(2/a)$ is concave and has a maximum of $\frac{2}{\mathrm{e}}$. Therefore, we can replace $a\log(2/a)$ with $\frac{2}{\mathrm{e}}$, divide everything by 4 and get

$$R(a):=a\left(\frac{w}{\mathrm{e}}-w\log(w+1)\right)-\frac{w}{\mathrm{e}}+w\log(w+1)+\log(w+1).$$

We now analyze two cases: $\frac{w}{\mathrm{e}} - w\log(w+1) \le 0$ and $\frac{w}{\mathrm{e}} - w\log(w+1) > 0$. If $\frac{w}{\mathrm{e}} - w\log(w+1) \le 0$ then the worst a is $a = 1$, for which, we have, $R(1) = \log(w + 1)$, which is non-negative. If $\frac{w}{\mathrm{e}} - w\log(w+1) > 0$ then the worst a is $a = 0$, for that, we have

$$R(0)=(w+1)\log(w+1)-\frac{w}{\mathrm{e}}$$

Which is non-negative since it is 0 at $w = 0$ and the first derivative, $\frac{\partial}{\partial w}R = (w+1)\log(w+1) - \frac{w}{\mathrm{e}}$, is positive. We now have to turn to the other endpoint of b , where it is $-a$. In that case, we get

$$T(w,a,-a)=4(w+1)(aw+1)\log(w+1)-2aw(a(w+3)+w+1)\log\left(\frac{2}{a}\right). \tag{16}$$

We now consider two cases, $a \le 1/10$ and $a > 1/10$. For $a > 1/10$, we do as in the other endpoint of b and bound $a\log(\frac{2}{a})$ by $2/\mathrm{e}$, divide by 4, and get

$$T(w,a,-a)/4\le a\left(w(w+1)\log(w+1)-\frac{w(w+3)}{\mathrm{e}}\right)+\frac{(w+1)(\mathrm{e}\log(w+1)-w)}{\mathrm{e}}.$$

Split to two cases, $w(w+1)\log(w+1) - \frac{w(w+3)}{\mathrm{e}}$ is non-negative or negative. If it is negative, then the worst where $a = 1$, in which case the above expression becomes

$$\left(w(w+1)\log(w+1) - \frac{w(w+3)}{\mathrm{e}}\right) + \frac{(w+1)(\mathrm{e}\log(w+1) - w)}{\mathrm{e}}$$
$$= (w+1)^2\log(w+1) - \frac{2w(w+2)}{\mathrm{e}}.$$

The above is non-negative because $w \geq 3$ it is positive for $w = 3$ and its derivative, $\frac{(w+1)(2\mathrm{e}\log(w+1)+\mathrm{e}-4)}{\mathrm{e}}$, is also positive for $w \geq 3$.
If $w(w+1)\log(w+1) - \frac{w(w+3)}{\mathrm{e}}$ is positive, then the worse is $a = \frac{1}{10}$. Plugging it into eq. (16) results in

$$T(w, 1/10, -1/10) = \mathrm{e}\left(w^2 + 11w + 10\right)\log(w+1) - w(11w+13),$$

which is again positive since it is positive for $w = 3$ and the first derivative is positive. Finally, we are left with the $a \leq 1/10$ case. Lastly, we need to prove

$$T(w, a, -a) = 4(w+1)(aw+1)\log(w+1) - 2aw(a(w+3) + w + 1)\log\left(\frac{2}{a}\right) \geq 0$$

for $0 \leq a \leq 1/10$ and $w \geq 3$. We do this first by showing that (1) $T(3, a, -a) > 0$ and (2) $\frac{\partial}{\partial w}T(w, a, -a)\big|_{w=3} > 0$ for the appropriate range. Then, proving (3) $\frac{\partial^2}{\partial w^2}T(w, a, -a) \geq 0$ for the appropriate range completes the proof. First, taking care of $T(3, a, -a)$, we have

$$T(3, a, -a) = 3\left(-5a^2 + (-7a-5)a - 3a\right)\log\left(\frac{2}{a}\right) + 2(-6(-a-1) + 18a + 2)\log(4)$$
$$= 4\left(4(a\log(64) + \log(4)) - 3a(3a+2)\log\left(\frac{2}{a}\right)\right),$$

which has the following positive second derivative,

$$\frac{\partial^2}{\partial a^2}T(3, a, -a) = 4\left(-3a\left(\frac{3a+2}{a^2} - \frac{6}{a}\right) - 6\left(3\log\left(\frac{2}{a}\right) - \frac{3a+2}{a}\right)\right)$$
$$= 12\left(\frac{2}{a} - 6\log\left(\frac{2}{a}\right) + 9\right).$$

This second derivative is positive because it is decreasing by the fact $\frac{\partial^3}{\partial a^3}T(3, a, -a) = \frac{24(3a-1)}{a^2} < 0$ and the minimum of $\frac{\partial^2}{\partial a^2}T(3, a, -a)$ at $a \in [0, 1/10]$,

$$\frac{\partial^2}{\partial a^2}T(3, a, -a)\bigg|_{a=1/10} = 12(29 - 6\log(20)) \approx 132.307,$$

is positive. Knowing that $\frac{\partial^2}{\partial a^2}T(3, a, -a)$ is convex, we lower bound it by its tangent line at $a = 0.095$,

$$a\left(\frac{1371}{50} + 16\log(64) - \frac{771}{25}\log\left(\frac{400}{19}\right)\right) + 16\log(4) + \frac{57\left(57\log\left(\frac{400}{19}\right) - 457\right)}{10000} \approx 20.566 - 0.008a.$$

This tangent line is decreasing and is positive at $a \in [0, 1/10]$. Therefore, $\frac{\partial^2}{\partial a^2} T(3, a, -a)$ is positive at $a \in [0, 1/10]$.

Next, we analyze

$$\frac{\partial}{\partial w} T(w, a, -a)\bigg|_{w=3} = 2\left(2a(3 + 7\log(4)) - a(9a + 7)\log\left(\frac{2}{a}\right) + 2 + \log(16)\right),$$

which has a positive second derivative, $\frac{2(18a-7)}{a^2}$, and thus we can use a tangent line at point $a = 0.025$ as follows. We have,

$$\begin{aligned}\frac{\partial}{\partial w} T(w, a, -a)\bigg|_{w=3} &\geq \frac{1}{800}(2911 + 6436\log(2) + 9\log(5)) - \frac{1}{20}a(-529 + 72\log(2) + 298\log(5)) \\ &\approx 9.23323 - 0.0259547a,\end{aligned}$$

which is positive for $a \in [0, 1/10]$.

Third and last, we show that $\frac{\partial^2}{\partial w^2} T(w, a, -a) \geq 0$. We have that

$$\begin{aligned}\frac{\partial^2}{\partial w^2} T(w, a, -a) &= \frac{12aw - 4(a+1)a(w+1)\log\left(\frac{2}{a}\right) + 8a(w+1)\log(w+1) + 8a + 4}{w+1} \\ &=: \frac{U(w)}{w+1},\end{aligned}$$

thus it is sufficient to prove $U(w) > 0$. This is done by finding the critical point, proving that it is the minimum point, and then showing that U is positive at that point. To locate the minimum point, we solve

$$U'(w) = 8a\log(w+1) + 20a - 4(a+1)a\log\left(\frac{2}{a}\right) = 0$$

for w and get

$$w_0 = \mathrm{e}^{\frac{1}{2}\left(a\log\left(\frac{2}{a}\right) + \log\left(\frac{2}{a}\right) - 5\right)} - 1.$$

This is indeed the minimum point because

$$U''(w) = \frac{8a}{w+1} > 0$$

for our ranges of w, a (unless $a = 0$ but then $U(w) > 0$ immediately). To show that $U(w_0) > 0$, we first compute

$$\begin{aligned}U(w_0) =& 8a + 12a\left(\mathrm{e}^{\frac{1}{2}\left(a\log\left(\frac{2}{a}\right) + \log\left(\frac{2}{a}\right) - 5\right)} - 1\right) - 4(a+1)a\mathrm{e}^{\frac{1}{2}\left(a\log\left(\frac{2}{a}\right) + \log\left(\frac{2}{a}\right) - 5\right)}\log\left(\frac{2}{a}\right) \\ &+ 8a\mathrm{e}^{\frac{1}{2}\left(a\log\left(\frac{2}{a}\right) + \log\left(\frac{2}{a}\right) - 5\right)}\log\left(\mathrm{e}^{\frac{1}{2}\left(a\log\left(\frac{2}{a}\right) + \log\left(\frac{2}{a}\right) - 5\right)}\right) + 4 \\ =& \frac{3\,2^{\frac{a+5}{2}}\left(\frac{1}{a}\right)^{\frac{a-1}{2}} - 4\mathrm{e}^{5/2}a - 2^{\frac{a+5}{2}}(a+1)\left(\frac{1}{a}\right)^{\frac{a-1}{2}}\log\left(\frac{2}{a}\right)}{\mathrm{e}^{5/2}} \\ &+ \frac{\frac{1}{2}2^{\frac{a+7}{2}}\left(\frac{1}{a}\right)^{\frac{a-1}{2}}\left((a+1)\log\left(\frac{2}{a}\right) - 5\right) + 4\mathrm{e}^{5/2}}{\mathrm{e}^{5/2}} \\ =& -\frac{2^{\frac{a+7}{2}}\left(\frac{1}{a}\right)^{\frac{a-1}{2}}}{\mathrm{e}^{5/2}} - 4a + 4,\end{aligned}$$

and additionally

$$\frac{\partial}{\partial a}U(w_0) = -\frac{2^{\frac{a+7}{2}-1}\left(\frac{1}{a}\right)^{\frac{a-1}{2}}\log(2)}{\mathrm{e}^{5/2}} - \frac{2^{\frac{a+7}{2}}\left(\frac{1}{a}\right)^{\frac{a-1}{2}}\left(\frac{1}{2}\log\left(\frac{1}{a}\right) - \frac{a-1}{2a}\right)}{\mathrm{e}^{5/2}} - 4$$
$$= -\frac{4\left(2^{\frac{a+1}{2}}\left(\frac{1}{a}\right)^{\frac{a+1}{2}}(a(\log(2)-1)+1) + 2^{\frac{a+1}{2}}\left(\frac{1}{a}\right)^{\frac{a-1}{2}}\log\left(\frac{1}{a}\right) + \mathrm{e}^{5/2}\right)}{\mathrm{e}^{5/2}},$$

and note that the latter expression is negative for $a \in [0, 1/10]$. Thus, $U(w_0)$ is decreasing in a and at the minimum point has a value of

$$U(w_0)|_{a=1/10} = \frac{18}{5} - \frac{8\sqrt[10]{2}}{5^{9/20}\mathrm{e}^{5/2}} \approx 3.25887 > 0.$$

The proof is complete. ■

Lemma 19 *The function $\rho : \mathbb{N}^2 \to \mathbb{R}_+$ defined in (7) satisfies the metric axioms.*

Proof Let $f(x) := \frac{2}{\sqrt{3}} \wedge \sqrt{\frac{2}{\log\frac{2}{x}}}$ where $f(0) = 0$ such that we have $\rho = f \circ \xi$. It is known that non-negative, non-decreasing, concave functions with $f^{-1}(0) = \{0\}$ are metric-preserving (see, e.g., Kaplansky, 2001, p. 70). It is easy to see that f is satisfies $f^{-1}(0) = \{0\}$, and is non-decreasing. To see that f is concave, observe that

$$\frac{\partial^2}{\partial x^2}\sqrt{\frac{2}{\log\left(\frac{2}{x}\right)}} = \sqrt{2}\left(\frac{3\left(\frac{1}{\log\left(\frac{2}{x}\right)}\right)^{5/2}}{4x^2} - \frac{\left(\frac{1}{\log\left(\frac{2}{x}\right)}\right)^{3/2}}{2x^2}\right)$$
$$= \frac{\left(3 - 2\log\left(\frac{2}{x}\right)\right)\left(\frac{1}{\log\left(\frac{2}{x}\right)}\right)^{5/2}}{2\sqrt{2}x^2}$$

is negative for $0 < x < \frac{2}{e^{3/2}}$, which is where $\sqrt{\frac{2}{\log\left(\frac{2}{x}\right)}} \leq \frac{2}{\sqrt{3}}$. Since the minimum of concave functions is concave, we are done. ■

Recall the notation $\mathcal{N}_\rho(\varepsilon)$ for the ε-covering number of $(\mathbb{N}, \rho)$. Because $(X_i)_{i\geq 1}$ is a subgaussian process on $(\mathbb{N}, \rho)$, Dudley's theorem directly gives upper bounds on $\Delta_n(\mu)$.

Proof of Proposition 7. Theorem 18 shows that the vector $\hat{p} - p$ is sub-Gaussian with respect to the metric $\frac{\rho}{\sqrt{n}}$. As a result, Dudley's theorem (Van Handel (2014, Corollary 5.25)) shows that

$$\Delta_n(\mu) = \mathbb{E}\sup_{i\in\mathbb{N}}|\hat{p}_i - p_i| \leq 24\int_0^\infty \sqrt{\log\mathcal{N}_\rho(\varepsilon\sqrt{n})}\mathrm{d}\varepsilon = \frac{24}{\sqrt{n}}\int_0^\infty \sqrt{\log\mathcal{N}_\rho(\varepsilon)}\mathrm{d}\varepsilon.$$

In the last equality, we noted that for any $\varepsilon \geq 1$, one has $\mathcal{N}_\xi(\varepsilon) = 1$. This ends the proof. ■

As a remark, we can check that the proposed condition $\int_0^1 \sqrt{\mathcal{N}_\rho(\varepsilon)}\mathrm{d}\varepsilon < \infty$ in Proposition 7 is stronger than the sufficient condition $\int_0^1 \mathcal{N}_\xi(\varepsilon)\mathrm{d}\varepsilon < \infty$ from Theorem 6. Indeed, suppose that one

has $\int_0^1 \sqrt{\mathcal{N}_\rho(\varepsilon)}\mathrm{d}\varepsilon < \infty$, then in particular, $\varepsilon\sqrt{\log \mathcal{N}_\rho(\varepsilon)} \underset{\varepsilon\to 0^+}{\longrightarrow} 0$. This implies that for any $c > 0$, we have

$$e^{-\frac{c}{\varepsilon^2}}\mathcal{N}_\rho(\varepsilon) \underset{\varepsilon\to 0^+}{\longrightarrow} 0.$$

Also, for $\varepsilon \in (0, \frac{2}{\sqrt{3}}]$, we have

$$\mathcal{N}_\rho(\varepsilon) = \mathcal{N}_\xi\left(2e^{-\frac{2}{\varepsilon^2}}\right).$$

As a result, this shows that for any $c > 0$, we have

$$\varepsilon^c \mathcal{N}_\xi(\varepsilon) \underset{\varepsilon\to 0}{\longrightarrow} 0.$$

The above bound for any $c < 1$ already shows that $\int_0^1 \mathcal{N}_\xi(\varepsilon)\mathrm{d}\varepsilon < \infty$.

Appendix D. On the exact conditions for the convergence $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$

In this section, we provide sufficient conditions for the convergence of the expected maximum deviation Δ_n and identify some key challenges for the general characterization.

We start by proving Theorem 9. This gives a sufficient condition for the decay of Δ_n to 0 that is a significantly weaker condition than the condition Eq (8) from Theorem 6. To the best of our knowledge, we are not aware of any distribution that would not satisfy it, but would still exhibit the convergent behavior for Δ_n. For the sake of exposition, we recall the condition from Theorem 9 for distributions μ on $\{0,1\}^{\mathbb{N}}$:

Sufficient Condition (SC) *The metric space $(\mathbb{N}, \xi)$ is totally bounded and there exists $K \geq 1$ such that for any $\varepsilon > 0$, there exist events $(E_k)_{k\in\mathbb{N}}$ and a finite set $J \subset \mathbb{N}$ with*

- $\mathbb{P}(E_k) \leq \varepsilon, \ \forall k \in \mathbb{N}$,
- $\sup_{k\in\mathbb{N}} \dfrac{\log(k+1)}{\log \frac{1}{\mathbb{P}(E_k)}} < \infty$,
- $\forall i \in \mathbb{N}, \exists j \in J, \exists \mathcal{K} \subset \mathbb{N}$, *such that* $|\mathcal{K}| \leq K$ *and* $\{X_i \neq X_j\} \subset \bigcup_{k\in\mathcal{K}} E_k$.

Then $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$.

Proof of Theorem 9 Fix $\varepsilon > 0$ and consider the events $(E_k)_{k\in\mathbb{N}}$ as provided by the condition. Fix $n \geq 1$. Mirroring the notation for $\hat{p}$, we define $\hat{q}$ (resp. $\hat{u}$) as the empirical probability vector for the variable $(\mathbb{1}[E_k])_{k\in\mathbb{N}}$ (resp. $(X_j)_{j\in J}$). That is, if we denote by an exponent (n) different samples from these random variables, we pose

$$\hat{q}_k = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[E_k^{(i)}], \qquad \hat{u}_j = \frac{1}{n}\sum_{i=1}^n X_j^{(i)}.$$

From Cohen and Kontorovich (2023b); Blanchard and Voráček (2024), we know that since

$$T(E) := \sup_{k\in\mathbb{N}} \frac{\log(k+1)}{\log \frac{1}{\mathbb{P}(E_k)}} < \infty,$$

letting $q_k = \mathbb{P}(E_k) \leq \varepsilon$, one has

$$\mathbb{E}\|\hat{q} - q\|_\infty \underset{n\to\infty}{\longrightarrow} 0. \tag{17}$$

Next, for any $i \in \mathbb{N}$, let $\mathcal{K}_i \subset \mathbb{N}$ and $j_i \in [J]$ be the set of indices indices such that $\{X_i \neq Z_{j_i}\} \subset \bigcup_{k\in\mathcal{K}_i} E_k$. Then, with $u_j = \mathbb{E}[Z_j]$ for $j \in [J]$, we have

$$\begin{aligned}
|\hat{p}_i - p_i| &\leq |\hat{p}_i - \hat{u}_{j_i}| + |\hat{u}_{j_i} - u_{j_i}| + |u_{j_i} - p_i| \\
&\leq \frac{1}{n}\sum_{l=1}^{n} \mathbb{1}[X_i^{(l)} \neq X_{j_i}^{(l)}] + |\hat{u}_{j_i} - u_{j_i}| + \mathbb{P}(X_i \neq X_{j_i}) \\
&\leq \frac{1}{n}\sum_{l=1}^{n}\sum_{k\in\mathcal{K}_i} \mathbb{1}[E_k^{(l)}] + \|\hat{u} - u\|_\infty + \sum_{k\in\mathcal{K}_i} \mathbb{P}(E_k) \\
&\leq \sum_{k\in\mathcal{K}_i} (\hat{q}_k + \varepsilon) + \|\hat{u} - u\|_\infty.
\end{aligned}$$

Next, for any $k \in \mathcal{K}_i$,

$$\hat{q}_k \leq q_k + |\hat{q}_k - q_k| \leq \mathbb{P}(E_k) + \|\hat{q} - q\|_\infty \leq \varepsilon + \|\hat{q} - q\|_\infty.$$

Putting the two previous inequalities together yields

$$\|\hat{p} - p\|_\infty \leq (2\varepsilon + \|\hat{q} - q\|_\infty)K + \|\hat{u} - u\|_\infty.$$

Because J is finite, $\mathbb{E}\|\hat{u} - u\|_\infty \to 0$. Together with Eq (17), this gives

$$\limsup_{n\to\infty} \Delta_n(\mu) \leq 2K\varepsilon.$$

This holds for any $\varepsilon > 0$, hence we obtained the desired result $\Delta_n(\mu) \to 0$ as $n \to \infty$. ■

An inspection of the proof shows that one does not need the random variables $(X_j)_{j\in J}$ used as "centers" to belong to the set of entries $\{X_i, i \geq 1\}$. In fact, the proof holds if we put no restriction on these centers. This yields the following result.

Corollary 20 *Let μ be a distribution on $\{0,1\}^{\mathbb{N}}$ such that $(\mathbb{N}, \xi)$ is totally bounded. Suppose that there exists $K \geq 1$ such that for any $\varepsilon > 0$, there exist events $(E_k)_{k\in\mathbb{N}}$, and a finite sequence of random variables $(Z_j)_{j\in[J]}$ (defined on the same probability space as μ) with*

- $\mathbb{P}(E_k) \leq \varepsilon, \ \forall k \in \mathbb{N}$,
- $\sup_{k\in\mathbb{N}} \dfrac{\log(k+1)}{\log \frac{1}{\mathbb{P}(E_k)}} < \infty$,
- $\forall i \in \mathbb{N}, \exists j \in J, \exists \mathcal{K} \subset \mathbb{N}$, *such that* $|\mathcal{K}| \leq K$ *and* $\{X_i \neq Z_j\} \subset \bigcup_{k\in\mathcal{K}} E_k$.

Then $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$.

While this condition seems more general than the condition (SC), they turn out to be equivalent.

Proposition 21 *The condition (SC) is equivalent to the condition from Corollary 20.*

Proof It suffices to show that if μ satisfies the condition from Corollary 20, then it also satisfies (SC). Fix such a distribution. We will use all the notations of the condition and we now aim to find adequate parameters to satisfy (SC). We will use $\tilde{K} := K + 1$. Fix $\varepsilon > 0$. Because μ satisfies the condition in Corollary 20 for $\tilde{\epsilon} = \frac{\epsilon}{K}$, there exists events $(E_k)_{k\in\mathbb{N}}$ and random variables $(Z_j)_{j\in[J]}$ satisfying the conditions for $\tilde{\epsilon}$. We also fix for $i \geq 1$, an element $j_i \in J$ and set $\mathcal{K}_i \subset \mathbb{N}$ such that

$$\{X_i \neq Z_{j_i}\} \subset \bigcup_{k\in\mathcal{K}_i} E_k. \tag{18}$$

Fix $j \in [J]$. First suppose that

$$\mathbb{P}(Z_j \neq X_i) > \epsilon, \quad i \geq 1.$$

Then, we can check that $j_i \neq j$, because

$$\mathbb{P}(X_i \neq Z_{j_i}) \leq \mathbb{P}\left(\bigcup_{k\in\mathcal{K}_i} E_k\right) \leq \sum_{k\in\mathcal{K}_i} \mathbb{P}(E_k) \leq K\tilde{\varepsilon} = \varepsilon.$$

As a result, the variable Z_j is simply not needed and we can delete it from the set of centers $(Z_j)_{j\in[J]}$. We can therefore suppose without loss of generality that for all $j \in [J]$, there is some $i(j) \geq 1$, for which

$$\mathbb{P}(X_{i(j)} \neq Z_j) \leq \varepsilon.$$

We then define the event $F_j := \{X_{i(j)} \neq Z_j\}$ for all $j \in [J]$ and add all these to the sequence of covering events $(E_k)_{k\geq 1}$ by defining

$$\tilde{E}_j := \begin{cases} F_j & j \leq J \\ E_{j-J} & j > J. \end{cases}$$

The first condition for (SC) is satisfied by construction of the events F_j for $j \in [J]$ because $\mathbb{P}(F_j) = \mathbb{P}(X_{i(j)} \neq Z_j) \leq \varepsilon$. Next, we only added a finite number of events to the sequence, hence the second property is still valid. Last, for $i \geq 1$, because Eq (18) holds, we have

$$\{X_i \neq X_{i(j_i)}\} \subset \{X_i \neq Z_{j_i}\} \cup \{X_{i(j_i)} \neq Z_{j_i}\} \subset F_j \cup \bigcup_{k\in\mathcal{K}_i} E_k = \bigcup_{k\in\{j_i\}\cup\{k+J, k\in\mathcal{K}_i\}} \tilde{E}_k$$

This ends the proof that μ satisfies condition (SC), which gives the desired result. ■

The proposed condition (SC) essentially asks that "bad events" $\{X_i \neq X_j\}$ can be adequately covered by some sequence of events $(E_k)_{k\in\mathbb{N}}$. As discussed in Section 2, this significantly generalizes the condition $\int_0^1 \mathcal{N}_\xi(\varepsilon)\mathrm{d}\varepsilon < \infty$ along two directions.

D.1. Generalization (i)

We showed in Theorem 3 and Proposition 17 that 2nd and 3rd order moment information on the distribution μ is not enough to have a necessary and sufficient characterization. The condition (SC) instead covers deviations via events in the probability space of μ directly, which allows for correlations with an arbitrarily large number of coordinates.

For instance, we can check how the condition (SC) distinguishes between the two distributions μ and ν from Theorem 3. For ν, because the variables C_t are independent even within each block $t \in \mathcal{S}_k$ for some fixed $k \geq 1$, there is no convenient choice of covering events E_k. On the other hand, for μ, one can directly choose the bad events $E_k := \{B_k = 1\}$:

$$\{X_t^\mu \neq Z_0\} \subset \{B_k = 1\} = E_k, \quad t \in \mathcal{S}_k, k \geq 1.$$

We can therefore cover the deviations of all entries X_t^μ for $t \in \mathcal{S}_k$ using a single event E_k with small probability $\mathbb{P}(E_k) = 2^{-k}$. However, for any $t \neq t' \in \mathcal{S}_k$, one has

$$\xi(t, t') = \mathbb{P}(X_t^\mu \neq X_{t'}^\mu) = \frac{\mathbb{P}(E_k)}{2}.$$

Hence, contrary to (SC), the covering number approach severely suffers from the size of the block $|\mathcal{S}_k|$ (so would any approach that looks at a fixed number of entries at once).

D.2. Generalization (ii)

The condition (SC) allows to cover the bad event $\{X_i \neq Z_j\}$ potentially with several events E_k (at most K), which departs from standard coverings for which one aims to directly cover the probability $\mathbb{P}(X_i \neq X_j)$. The alternative condition would be written as follows.

Tentative Condition 1 (TC1) *The metric space $(\mathbb{N}, \xi)$ is totally bounded and for any $\varepsilon > 0$, there exist events $(E_k)_{k\in\mathbb{N}}$ and a finite set $J \subset \mathbb{N}$ with*

- $\mathbb{P}(E_k) \leq \varepsilon, \ \forall k \in \mathbb{N}$,
- $\sup_{k\in\mathbb{N}} \dfrac{\log(k+1)}{\log \frac{1}{\mathbb{P}(E_k)}} < \infty$,
- $\forall i \in \mathbb{N}, \exists j \in J, \exists k \in \mathbb{N}, \{X_i \neq X_j\} \subset E_k$.

While this is still sufficient by Theorem 9, we can show that it is not necessary.

Proposition 22 ((TC1) is not necessary) *There exists a probability measure μ on $\{0,1\}^{\mathbb{N}}$ that does not satisfy condition (TC1) but $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$.*

Proof Let $(Y_k)_k \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ and independent random variables A_k such that $A_k \sim \text{Bernoulli}(1/\sqrt{k})$. Put

$$X_k = (1 - A_k)Y_0 + A_k Y_k, \quad k \geq 1.$$

We then define the distributions μ such that $(X_k)_{k\geq 1} \sim \mu$.

We first show that $\Delta_n(\mu) \to 0$ as $n \to \infty$ by checking that it satisfies the condition from Corollary 20. Intuitively, the random variables become closer and closer to Y_0, hence we can choose

$Z_1 := Y_0$. Fix $\varepsilon > 0$ and let $k_\varepsilon \geq 1/\varepsilon^2$. We can then pose $J = 1+k_\varepsilon$ and $Z_j = X_{j-1}$ for $2 \leq j \leq J$. For the other variables, we can simply pose $E_k := \{X_{k+k_\varepsilon} \neq Y_0\}$ for $k \geq 1$. For the covering sets, we can simply pose $K_k = \{1\}$ for $k \leq k_\varepsilon$ and $K_k = \{k - k_\varepsilon\}$ for $k > k_\varepsilon$. We can check that these parameters satisfy the condition from Corollary 20 in a straightforward manner. For $k \geq 1$, $\mathbb{P}(E_k) = \mathbb{P}(X_{k+k_\varepsilon} \neq Y_0) = \frac{1}{\sqrt{k+k_\varepsilon}} \leq \varepsilon$. Next, because these probabilities decay as $\frac{1}{\sqrt{k_\varepsilon+k}}$, the second condition in (SC) is satisfied. Last, for $i \leq k_\varepsilon$ we have $\{X_i \neq Z_{i+1}\} = \{X_i \neq X_i\} = \emptyset \subset E_1$, and for $i > k_\varepsilon$, $\{X_i \neq Z_1\} = \{X_i \neq Y_0\} = E_{i-k_\varepsilon}$. This ends the proof that the condition from Corollary 20 is satisfied and as a result,

$$\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0.$$

Next, suppose by contradiction that (TC1) is satisfied. We use this property for $\varepsilon = \frac{1}{4}$, using the same notations as in the condition. For $i \geq 1$, we also denote $j_i \in J$ and $k_i \in \mathbb{N}$ elements such that $\{X_i \neq X_{j_i}\} \subset E_{k_i}$. Because J is finite, we denote $j_{max} = \max\{j, j \in J\}$. Then, for any $i \geq 4$, we have

$$\mathbb{P}(X_i \neq X_{j_i}) = \frac{1}{2}\left(\frac{1}{\sqrt{i}} + \frac{1}{\sqrt{j_i}} - \frac{1}{\sqrt{ij_i}}\right) \geq \frac{1}{2}\left(\frac{1}{\sqrt{j_i}} - \frac{1}{\sqrt{4j_i}}\right) = \frac{1}{4\sqrt{j_i}} \geq \frac{1}{4\sqrt{j_{max}}}.$$

Next, the second condition of (TC1) implies in particular that there exists k_{max} such that

$$\mathbb{P}(E_k) < \frac{1}{4K\sqrt{j_{max}}}, \quad k \geq k_{max}.$$

Now recall that for $i \geq 4$, one has $\mathbb{P}(X_i \neq X_{j_i}) \leq \mathbb{P}(E_{k_i})$. Combining the two last equations shows that for any $i \geq 4$, we have $k_i < k_{max}$. Recalling that j_i can only take $|J|$ values, this implies that there is some couple $(j, k) \in J \times [k_{max} - 1]$ for which the set

$$\mathcal{S}(j,k) := \{4 \leq i \leq (4k_{max}|J|)^2 + 3, (j_i, k_i) = (j, k)\},$$

has at least $16k_{max}|J|$ elements. Next, note that

$$\bigcup_{i\in\mathcal{S}(j,k)} \{X_i \neq X_j\} \subset E_k.$$

Hence, taking the probabilities yields

$$\begin{aligned}
1 - \mathbb{P}(E_k) \leq \mathbb{P}\left(\bigcap_{i\in\mathcal{S}(j,k)} \{X_i = X_j\}\right) &\leq \mathbb{P}(X_j \neq Y_0) + \mathbb{P}\left(\bigcap_{i\in\mathcal{S}(j,k)} \{X_i = Y_0\}\right) \\
&\leq \frac{1}{2\sqrt{j}} + \prod_{i\in\mathcal{S}(j,k)}\left(1 - \frac{1}{2\sqrt{i}}\right) \\
&\leq \frac{1}{2} + \exp\left(-\sum_{i\in\mathcal{S}(j,k)} \frac{1}{2\sqrt{i}}\right).
\end{aligned}$$

Now we compute

$$\sum_{i\in\mathcal{S}(j,k)} \frac{1}{2\sqrt{i}} \geq \frac{16k_{max}|J|}{2\sqrt{(4k_{max}|J|)^2 + 3}} \geq \frac{3}{2}.$$

Together with the previous equation, this implies

$$\mathbb{P}(E_k) \geq \frac{1}{2} - e^{-3/2} > \frac{1}{4} = \varepsilon.$$

This contradicts the first property of condition (TC1), which proves that μ does not satisfy this condition and ends the proof. ■

In the previous example, one of the main reasons why the condition (TC1) is not satisfied is that there is no adequate finite choice of "centers" $(X_j)_{j\geq 1}$—the random variable Y_0 is missing from the sequence $\{X_j, j \geq 1\}$, while it would be a natural candidate to be used as the center. A possible tentative to fix this issue would be to allow the centers to be general random variables, in the spirit of the condition proposed in Corollary 20. This yields the following condition.

Tentative Condition 2 (TC2) *The metric space $(\mathbb{N}, \xi)$ is totally bounded and for any $\varepsilon > 0$, there exist events $(E_k)_{k\in\mathbb{N}}$ and a finite set of random variables $(Z_j)_{j\in[J]}$ (defined on the same probability space as μ) with*

- $\mathbb{P}(E_k) \leq \varepsilon, \ \forall k \in \mathbb{N}$,
- $\sup_{k\in\mathbb{N}} \dfrac{\log(k+1)}{\log \frac{1}{\mathbb{P}(E_k)}} < \infty$,
- $\forall i \in \mathbb{N}, \exists j \in [J], \exists k \in \mathbb{N}, \{X_i \neq Z_j\} \subset E_k$.

By Corollary 20 and Theorem 9, this is still a sufficient condition, which fortunately also encompasses the example provided in the previous result, Proposition 22. However, even with this fix, being able to cover bad events $\{X_i \neq Z_j\}$ with multiple events E_k is still necessary.

Proposition 23 ((TC2) is not necessary) *There exists a probability measure μ on $\{0,1\}^{\mathbb{N}}$ that does not satisfy condition (TC2) but $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$.*

Proof We partition $\mathbb{N}$ into $\mathbb{N} = \bigcup_{l\geq 1} I_l$, where $I_l = \{2^{l-1} \leq i < 2^l\}$ for $l \geq 1$. We consider binary random variables Y_i, A_i, B_i for $i \geq 1$, together independent and such that $Y_i \sim \text{Bernoulli}(1/2)$, and $A_i, B_i \sim \text{Bernoulli}(1/\sqrt{i})$, for all $i \geq 1$. For $l \geq 1$ and any $i \in I_l$, put

$$X_i = (1 - A_l)(1 - B_i)Y_0 + (A_l + B_i - A_l B_i)Y_i.$$

We then define μ as the distribution of $(X_i)_{i\geq 1}$. We first show that $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$ by proving that μ satisfies the sufficient condition from Corollary 20. Here, we use $K = 2$. Fix $\varepsilon > 0$ and let $i_0 = \lceil \frac{1}{\varepsilon^2} \rceil$ and $i_\varepsilon = 2^{i_0}$. We then define the events $(E_k)_{k\geq 1}$ as the sequence $(\{A_{i_0} = 1\}, \{B_{i_0} = 1\}, \{A_{i_0+1} = 1\}, \{B_{i_0+1} = 1\}, \ldots)$. Because of the polynomial decay of $\mathbb{P}(A_i = 1) = \mathbb{P}(B_i = 1) = \frac{1}{\sqrt{i}}$, we can check easily that the events $(E_k)_{k\geq 1}$ satisfy the first two conditions from Corollary 20. Last, we consider $J = i_\varepsilon + 1$ centers Y_0 and $(X_i)_{i\leq i_\varepsilon}$. The third condition from Corollary 20 is trivially satisfied for $i \leq i_\varepsilon$, since $\{X_i \neq X_i\} = \emptyset$. And for $i > i_\varepsilon$, letting $l \geq 1$ such that $i \in I_l$, since $i_\varepsilon = 2^{i_0}$, we have $l \geq i_0$. In particular, the events $\{A_l = 1\}$ and $\{B_i = 1\}$ belong to the sequence $(E_k)_{k\geq 1}$. We can conclude by noting that

$$\{X_i \neq Y_0\} \subset \{A_l = 1\} \cup \{B_i = 1\}.$$

This ends the proof that $\Delta_n(\mu) \underset{n\to\infty}{\longrightarrow} 0$.

We now show that μ does not satisfy (TC2). We suppose by contradiction that it does and use the property for $\varepsilon = \frac{1}{2}$. We use the notations of the condition and for any $i \geq 1$, we denote by $j_i \in [J]$ and $k_i \in \mathbb{N}$ elements such that $\{X_i \neq Z_{j_i}\} \subset E_{k_i}$. Because of the second property, there exists $C > 0$ such that

$$\mathbb{P}(E_k) \leq \frac{1}{k^{1/C}}, \quad l \geq 1. \tag{19}$$

Next, we recall that the sequence $(j_i)_{i\geq 1}$ only takes values in $[J]$. As a result, for any $l \geq |J| + 1$ there exists some index $j(l)$ such that

$$|\{i \in I_l : j_i = j(l)\}| \geq \frac{|I_l|}{|J|} = \frac{2^{l-1}}{|J|}.$$

We denote this set $\mathcal{A}(l) = \{i \in I_l : j_i = j(l)\}$. Suppose for now that for some $i \in \mathcal{A}(l)$, we have

$$\mathbb{P}(X_i \neq Z_{j(l)}) \leq \frac{3}{8\sqrt{l}}.$$

Then, for any $i' \in \mathcal{A}(l) \setminus \{i\}$, one has

$$\mathbb{P}(X_{i'} \neq Z_{j(l)}) \geq \mathbb{P}(X_{i'} \neq X_i) - \mathbb{P}(X_i \neq Z_{j(l)}) \geq \frac{3}{4}\mathbb{P}(A_l = 1) - \frac{3}{8\sqrt{l}} = \frac{3}{8\sqrt{l}}.$$

As a result, in all cases, there is a set $\mathcal{B}(l) \subset \mathcal{A}(l)$ of cardinality $|\mathcal{B}(l)| = |\mathcal{A}(l)| - 1 \geq 2^{l-1}/|J| - 1$ and for which

$$\mathbb{P}(E_{k_i}) \geq \mathbb{P}(X_i \neq Z_{j_i}) = \mathbb{P}(X_i \neq Z_{j(l)}) \geq \frac{3}{8\sqrt{l}}.$$

By Eq (19), this implies that for all $i \in \mathcal{B}(l)$, one has $k_i \leq (8l)^{C/2}$. As a result, there exists $k(l)$ for which

$$|\{i \in I_l : (j_i, k_i) = (j(l), k(l))\}| \geq \frac{|\mathcal{B}(l)|}{(8l)^{C/2}} \geq \frac{2^{l-1} - |J|}{(8l)^{C/2}|J|}.$$

We denote this set by $\mathcal{C}(l) := \{i \in I_l : (j_i, k_i) = (j(l), k(l))\}$. In particular, we obtained that for $l \geq (|J| + 2) \wedge \log_2(8(8l)^{C/2}|J|)$,

$$|\mathcal{C}(l)| \geq \frac{2^l}{4(8l)^{C/2}|J|} \geq 2.$$

We now use similar arguments to that of Proposition 22. Fix some element $i(l) \in \mathcal{C}(l)$. We have

$$\begin{aligned}
1 - \mathbb{P}(E_{k(l)}) &\leq \mathbb{P}\left(\bigcap_{i\in\mathcal{C}(l)} \{X_i = Z_{j(l)}\}\right) \\
&\leq \mathbb{P}(X_{i(l)} \neq Y_0) + \mathbb{P}\left(\bigcap_{i\in\mathcal{C}(l)} \{X_i = Y_0\}\right) \\
&= \frac{1}{2}\left(\frac{1}{\sqrt{i(l)}} + \frac{1}{\sqrt{l}} - \frac{1}{\sqrt{li(l)}}\right) + \prod_{i\in\mathcal{C}(i)}\left(1 - \frac{1}{2\sqrt{i}} - \frac{1}{2\sqrt{l}} + \frac{1}{2\sqrt{il}}\right) \\
&\leq \frac{1}{2}\left(\frac{1}{2^{(l-1)/2}} + \frac{1}{\sqrt{l}}\right) + \exp\left(-\frac{|\mathcal{C}(l)|}{2\sqrt{l}}\right).
\end{aligned}$$

For l sufficiently large, this gives $1 - \mathbb{P}(E_{k(l)}) \leq \frac{1}{4}$, which contradicts the hypothesis $\mathbb{P}(E_k) \leq \varepsilon = \frac{1}{2}$ for all $k \geq 1$. Hence μ does not satisfy (TC2), which ends the proof. ■