Proceedings of Machine Learning Research vol 237:1–42, 2024 35th International Conference on Algorithmic Learning Theory

Tight Bounds for Local Glivenko-Cantelli

Moïse Blanchard MOISEB@MIT.EDU
Massachusetts Institute of Technology

Václav Voráček VACLAV.VORACEK@UNI-TUEBINGEN.DE
University of Tübingen - Tübingen AI center

Editors: Claire Vernade and Daniel Hsu

Abstract

This paper addresses the statistical problem of estimating the infinite-norm deviation from the empirical mean to the distribution mean for high-dimensional distributions on $\{0,1\}^d$, potentially with $d=\infty$. Unlike traditional bounds as in the classical Glivenko-Cantelli theorem, we explore the instance-dependent convergence behavior. For product distributions, we provide the exact non-asymptotic behavior of the expected maximum deviation, revealing various regimes of decay. In particular, these tight bounds demonstrate the necessity of a previously proposed factor for an upper bound, answering a corresponding COLT 2023 open problem (Cohen and Kontorovich, 2022, 2023). We also consider general distributions on $\{0,1\}^d$ and provide the tightest possible bounds for the maximum deviation of the empirical mean given only the mean statistic. Along the way, we prove a localized version of the Dvoretzky–Kiefer–Wolfowitz inequality. Additionally, we present some results for two other cases, one where the deviation is measured in some q-norm, and the other where the distribution is supported on a continuous domain $[0,1]^d$, and also provide some high-probability bounds for the maximum deviation in the independent Bernoulli case.

1. Introduction

We consider the fundamental statistical problem of estimating the maximal empirical mean deviation for multiple independent Bernoulli random variables. Precisely, for a potentially infinite sequence $\boldsymbol{p}$ of parameters $p(j)\in[0,1]$ for $j\geq 1$, we consider the product distribution μ such that the coordinates of $X\sim\mu$ are independent Bernoulli random variables with parameters given by $\boldsymbol{p}$; that is $\mathbb{E}[\mu]=\boldsymbol{p}$ (we refer to the textbook Kallenberg (1997) for measure-theoretic concerns). Given n i.i.d. samples $X_1,\ldots,X_n$ of μ, we aim to understand the maximum deviation of the empirical mean $\hat{\boldsymbol{p}}_n=\frac{1}{n}\sum_{i=1}^n X_i$ to the mean $\boldsymbol{p}$. We mainly focus on its expectation

$$\Delta_n(\boldsymbol{p}):=\mathbb{E}\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_\infty=\mathbb{E}\sup_j|\hat{p}_n(j)-p(j)|,$$

Understanding the convergence of the empirical mean of i.i.d. sequences and studying mean estimators are foundational problems in statistical analysis. A substantial body of literature has explored convergence rates for mean estimation problems in fixed dimensions d, under diverse distributional assumptions (Catoni, 2012; Devroye et al., 2016; Lugosi and Mendelson, 2019b,a; Cherapanamjeri et al., 2019; Diakonikolas et al., 2020; Hopkins, 2020; Lugosi and Mendelson, 2021; Cherapanamjeri et al., 2022; Lee and Valiant, 2022). We note that one uses different estimators of the expectation based on the different use cases. For instance, if one seeks for an estimator such that the sample complexity of $\mathbb{P}(|\mu-\hat{\mu}_n|\geq\varepsilon)\leq\gamma$ is minimized, then sample mean is usually not the right choice.

On the other hand, the classical Glivenko-Cantelli theorem provides distribution-free convergence bounds for the empirical mean, quantified by Dvoretzky–Kiefer–Wolfowitz inequality: $\Delta_n(\boldsymbol{p}) \lesssim \sqrt{\ln(d+1)/n}$ for d-dimensional distributions μ. While this rate is optimal up to constants without further assumptions on p—it is attained when $\boldsymbol{p}$ is a d-dimensional constant vector $(c, c, \ldots, c)$ for some constant $c > 0$—this worst-case bound may not capture the correct behavior of $\Delta_n(\boldsymbol{p})$ for specific instances of $\boldsymbol{p}$. In particular, this bound is overly pessimistic when the coordinates of $\boldsymbol{p}$ decay to 0 sufficiently fast. As a simple example, in the infinite-dimensional case when $p(j) = 1/j$ for $j \geq 1$, $\Delta_n(\boldsymbol{p})$ converges to 0 as the number of samples n grows, while the Glivenko-Cantelli theorem does not provide a useful bound. Instead, we are interested in the instance-dependent convergence behavior, which allows us to provide dimension-free results; that is, results without an explicit dependency on dimension, and have potentially an infinite vector $\boldsymbol{p}$ of non-zero probabilities.

This problem was first posed and studied by Thomas (2018); Cohen and Kontorovich (2022). By symmetry, without loss of generality, we will assume that $p(j) \in [0, \frac{1}{2}]$ for every $j \geq 1$, and that the probabilities $p(1), p(2), \ldots$ are sorted in descending order that is $p(j) \geq p(j+1)$ for $j \geq 1$. Following the notation of Cohen and Kontorovich (2022), we denote by $[0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ these sequences. Having introduced the following functionals,

$$S(\boldsymbol{p}) = \sup_{j \in \mathbb{N}} p(j) \ln(j+1) \quad \text{and} \quad T(\boldsymbol{p}) = \sup_{j \in \mathbb{N}} \frac{\ln(j+1)}{\ln(1/p(j))},$$

they showed that $\Delta_n(\boldsymbol{p})$ converges to 0 if and only if $T(\boldsymbol{p}) < \infty$. Further, they characterized the asymptotic behavior of $\Delta_n(\boldsymbol{p})$ for sequences for which $T(\boldsymbol{p}) < \infty$ and showed that it decays as $\sqrt{S(\boldsymbol{p})/n}$ for $n \to \infty$, which corresponds to a sub-Gaussian decay regime for binomials, thoroughly studied in the literature (Kearns and Saul, 2013; Berend and Kontorovich, 2013a; Buldygin and Moskvichova, 2013). The same $\sqrt{S(\boldsymbol{p})/n}$ behavior also typically arises in the literature on minimax testing and goodness-of-fit problems with Gaussian, multinomial, or Poisson models Valiant and Valiant (2017); Balakrishnan and Wasserman (2019); Chhor and Carpentier (2020, 2021); Chhor et al. (2022). In terms of non-asymptotic results, they provide the following upper bound for a universal constant $c > 0$,

$$\Delta_n(\boldsymbol{p}) \leq c\left(\sqrt{\frac{S(\boldsymbol{p})}{n}} + \frac{T(\boldsymbol{p})\ln(n)}{n}\right), \quad n \geq e^3. \tag{1}$$

and conjectured that the $\ln(n)$ factor is superfluous in an open problem presented at COLT 2023 (Cohen and Kontorovich, 2023, 2022). In this work, we completely characterize the non-asymptotic behavior of $\Delta_n(\boldsymbol{p})$. In particular, we show that the $\ln(n)$ factor in Eq (1) is necessary when considering only the functionals $S(\boldsymbol{p})$ and $T(\boldsymbol{p})$. Our characterization unveils different regimes of decay for $\Delta_n(\boldsymbol{p})$, ranging from a somewhat Poissonian sub-gamma regime (Boucheron et al., 2013) to the asymptotic $\sqrt{S(\boldsymbol{p})/n}$ sub-Gaussian regime.

Notation We use the following notations for maxima and minima $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$ respectively. We write $f(n) \gtrsim g(n)$ (respectively $f(n) \asymp g(n)$) when there exists a universal constant $c > 0$ (respectively exist universal constants $c, C > 0$) such that $f(n) \geq cg(n)$ (respectively $cg(n) \leq f(n) \leq Cg(n)$) for every integer $n \geq 1$. The positive part of x is denoted as $[x]_+ := \max(0, x)$. Sequences are typed in bold (for example $\boldsymbol{p}$).

Outline of the paper We state our main results in Section 2. We then give an overview of the proof for the characterization of the expected maximum empirical deviation for product distributions in Section 3 and compare to the literature in Subsection 3.4. We next consider general distributions on $\{0,1\}^d$ in Section 4. Last, we discuss in Section 5 the implications of our results for COLT 2023 open problem and conclude in Section 6. Full proofs are given in the appendix.

2. Main results and discussion

In this section, we outline the main results. In particular, in Subsection 2.1 we outline the results for the case of independent Bernoulli random variables for ∞ norm deviations. Next, in Subsection 2.2 we show the results for the case of dependent Bernoulli random variables. As a stepping stone, we derive a variance dependent version of Dvoretzky-Kiefer-Wolfowitz inequality. In Subsection 2.3 we present the results for the case where the continuous distributions is supported on $[0,1]^d$ and not just on $\{0,1\}^d$. In Subsection 2.4 we provide the treatment for the case where we have independent Bernoulli random variables, but we measure the deviation in $q-$norm instead of ∞ norm. Finally, in Subsection 2.5 we present some high-probability bounds for the independent Bernoulli case.

2.1. Non-asymptotic bounds for independent Bernoulli random variables

We start with a brief overview of the results, ignoring the corner cases. It turns out that the crucial aspect is to determine the behavior of Δ for "step-like" sequences of probabilities $\mathbf{step}_{J,q}$ such that $\mathbf{step}_{J,q}(i) = q$ for all $i \le J$ and $\mathbf{step}_{J,q}(i) = 0$ otherwise. On the one hand, we will demonstrate that for a given sequence $\boldsymbol{p}$ it holds that $\Delta(\boldsymbol{p}) \gtrsim \Delta(\mathbf{step}_{i,p(i)})$. The reason is that $\boldsymbol{p} \ge \mathbf{step}_{i,p(i)}$ element-wise; and therefore the random variables following $\boldsymbol{p}$ have heavier tails. On the other hand, it also holds that $\Delta(\boldsymbol{p}) \lesssim \sup_{i\ge 1} \Delta(\mathbf{step}_{i,p(i)})$, which we will prove through tail-summation. We refer to Figure 1 for the illustration of this approach.

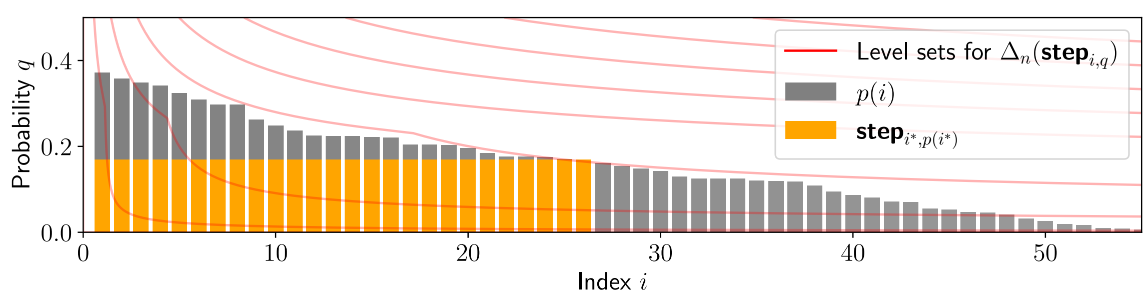

Figure 1: Illustration of the reduction from general probability profiles $\boldsymbol{p}$ to step functions $\mathbf{step}_{i,q}$. Red curves represent level sets of the expected maximum deviations for step functions $(i,q) \mapsto \Delta_n(\mathbf{step}_{i,q})$. The expected maximum deviation $\Delta_n(\boldsymbol{p})$ for general probabilities $\boldsymbol{p}$ is dominated by the maximum deviation of a step function of the form $\mathbf{step}_{i,p(i)}$, attained for $i^\star$.

Next, we compute the value of $\Delta(\mathbf{step}_{J,q})$ which exhibits three regimes. We state our main characterization in terms of a functional $\phi_{J,q}(n) \asymp \Delta_n(\mathbf{step}_{J,q})$. Formally, $\phi_{J,q}(n)$ is defined for all $n, J \ge 1$ and $q \in [0, \frac{1}{2}]$ via

$$\phi_{J,q}(n) := \begin{cases} 1 & n \leq \frac{\ln(J+1)}{\ln \frac{1}{q}}, \\ \frac{\ln(J+1)}{n \ln \frac{\ln(J+1)}{nq}} & \frac{\ln(J+1)}{\ln \frac{1}{q}} \leq n \leq \frac{\ln(J+1)}{eq}, \\ \sqrt{\frac{q \ln(J+1)}{n}} & n \geq \frac{\ln(J+1)}{eq}. \end{cases} \tag{2}$$

By convention, when $q = 0$, we pose $\phi_{J,q}(n) = 0$ for all $n, J \geq 1$. We now give some interpretation.

- First, a constant regime when $n \leq T(\mathbf{step}_{J,q})$, which was to be expected from the following bound from Cohen and Kontorovich (2022),

$$\Delta_n(\mathbf{step}_{J,q}) \geq 1 \wedge \frac{T(\mathbf{step}_{J,q})}{n}.$$

- The second regime interpolates between a behavior $T(\mathbf{step}_{J,q})/n$ when $n \leq \ln(J+1)/q^a$ for some arbitrary (but fixed) exponent $a < 1$; and a decay of the form $\ln(J+1)/n$ towards the end of the regime, when $n \sim \ln(J+1)/q$.

- Last, the third regime in which $\Delta_n(\mathbf{step}_{J,q}) \asymp \sqrt{S(\mathbf{step}_{J,q})/n}$ specifies when the asymptotic bound from Cohen and Kontorovich (2022) is tight.

The complete characterization of $\Delta_n(\boldsymbol{p})$ additionally exhibits a separate behavior for the small probability regime. The main result now can be written as follows.

Theorem 1 *Let $n \geq 1$ and $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$.*

- *If for all $j \geq 1$, one has $p(j) \leq \frac{1}{2nj}$, then $\Delta_n(\boldsymbol{p}) \asymp \frac{1}{n} \wedge \sum_{j \geq 1} p(j)$.*

- *Otherwise,*

$$\Delta_n(\boldsymbol{p}) \asymp \sup_{j \geq 1} \phi_{j,p(j)}(n) \asymp 1 \wedge \sup_{j \geq 1} \left(\sqrt{\frac{p(j)\ln(j+1)}{n}} \vee \frac{\ln(j+1)}{n \ln\left(2 + \frac{\ln(j+1)}{np(j)}\right)} \right).$$

In the second case, our bounds exhibit the asymptotic sub-gaussian term $\sqrt{S(\boldsymbol{p})/n}$, with a sub-gamma extra term that interpolates between the regime $n \leq T(\boldsymbol{p})$ for which $\Delta_n(\boldsymbol{p}) = \Theta(1)$ and the regime when the sub-gaussian term dominates. As a comparison to the bound Eq (1) written in terms of the functionals $S(\boldsymbol{p})$ and $T(\boldsymbol{p})$, in this intermediate regime, the expected maximum deviation lies between $T(\boldsymbol{p})/n$ and $T(\boldsymbol{p}) \ln n/n$. We refer to the end of Section 3 for a complete discussion on the implications of this result.

As a consequence of the characterization, we answer the open problem (Cohen and Kontorovich, 2023) by the negative. We show that if one only seeks bounds of $\Delta_n(\boldsymbol{p})$ in terms of the sub-Gaussian term $\sqrt{S(\boldsymbol{p})/n}$, and the functional $T(\boldsymbol{p})$, there are sequence instances for which the $\ln(n)$ term from Eq (1) is necessary. A constructive proof can be found in Section 5.

Theorem 2 *Suppose that there exists a constant $C \geq 1$ and $n_0 \geq 1$, and a function $\psi : \mathbb{N} \to \mathbb{R}$ such that the inequality*

$$\Delta_n(\boldsymbol{p}) \leq C\sqrt{\frac{S(\boldsymbol{p})}{n}} + \frac{T(\boldsymbol{p})}{n}\psi(n)$$

holds for all $n \geq n_0$ and $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ (product measures), then for an integer n_1 and a constant $c > 0$ depending only on C,

$$\psi(n) \geq c \ln n, \quad n \geq n_0 \vee n_1.$$

2.2. Non-asymptotic bounds for correlated Bernoulli random variables

The previous results focused on the particular case of product measures μ on $\{0,1\}^{\mathbb{N}}$, i.e., such that all coordinates of $X \sim \mu$ are mutually independent. In that case, the mean $\boldsymbol{p} = \mathbb{E}_\mu[X]$ completely characterizes the distribution, which in turn allows having the precise descriptions of the decay rate of $\Delta_n(\boldsymbol{p})$ from Theorem 1. Similarly, one can consider the considerably more general case of arbitrary distributions μ on $\{0,1\}^{\mathbb{N}}$ when coordinates may be correlated. As before, we study the expected maximum deviation $\Delta_n(\mu) = \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty$, where $\hat{\boldsymbol{p}}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for i.i.d. samples $X_i \sim \mu$. Our upper bounds from Theorem 1 extend directly to the general case; however, these may not be tight in general.

Corollary 3 *Let μ be a distribution on $\{0,1\}^{\mathbb{N}}$ with mean $\boldsymbol{p} = \mathbb{E}_{X\sim\mu}[X]$. Without loss of generality, suppose that $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$.*

- *If for all $j \geq 1$, one has $p(j) \leq \frac{1}{2nj}$, then $p(1) \lesssim \Delta_n(\mu) \lesssim \frac{1}{n} \wedge \sum_{j\geq 1} p(j)$.*
- *Otherwise,*

$$p(1) \wedge \sqrt{\frac{p(1)}{n}} \lesssim \Delta_n(\mu) \lesssim 1 \wedge \sup_{j\geq 1}\left(\sqrt{\frac{p(j)\ln(j+1)}{n}} \vee \frac{\ln(j+1)}{n\ln\left(2+\frac{\ln(j+1)}{np(j)}\right)}\right).$$

We emphasize that the gap between the upper and lower bounds from Theorem 3 can be large in general, but we show in Section 4 that these are the tightest bounds achievable if one only uses the mean statistic $\boldsymbol{p}$ to describe the distribution μ. As an extreme example, if $p(1) = p(i)$ for all $i \geq 1$, we can consider the perfectly-correlated case when μ is such that for $X \sim \mu$, almost surely $X(j) = X(1)$ for all $j \geq 1$. In that case, understanding $\Delta_n(\mu)$ reduces to computing the deviation from the mean for a single binomial $Y \sim \mathcal{B}(n,q)$ where $q = p(1) \in [0, \frac{1}{2}]$. It is well known that in this case, $\mathbb{E}|Y - nq| \asymp nq \wedge \sqrt{nq}$ (e.g. Berend and Kontorovich (2013b)), which corresponds to the lower bounds provided in Corollary 3.

En route to proving the tightness of Corollary 3 for bounds involving only the mean statistic $\boldsymbol{p}$, we prove a localized version of the classical Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Massart, 1990) which is of independent interest. Given n i.i.d. samples $X_1, \ldots, X_n$, from a real-valued random variable X, let $F : x \mapsto \mathbb{P}(X \leq x)$ be the cumulative distribution function (CDF) of X, and let $F_n : x \mapsto \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq x)$ be the empirical CDF. The standard DKW theorem shows that the deviations of $F_n(x)$ can be bounded uniformly in x.

Theorem 4 (DKW theorem (Massart, 1990)) *Let $X_1, \ldots, X_n$ be i.i.d. samples and denote by F (respectively F_n) the true CDF (respectively empirical CDF). Then, for any $t \geq 0$,*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \frac{t}{\sqrt{n}}\right) \leq 2e^{-2t^2}.$$

We aim to bound the deviation of the CDF on a smaller interval $[x_0, x_1]$ instead of the full domain $\mathbb{R}$. Indeed, when the maximum variance of $F(x)$ for $x \in [x_0, x_1]$ is small, one would expect to have stronger empirical deviation bounds than those provided by the vanilla Theorem 4. We note that Maillard (2021) provides an exact formula for the localized deviation of the CDF. This can be computed numerically with the formula, but an analytical simple upper bound will be more convenient for our purposes. In the following result, we show that one can achieve essentially the same DKW tail bounds uniformly on the interval $[x_0, x_1]$ as those for the single random variable $F_n(x)$ for $x \in [x_0, x_1]$ that has maximum variance. The proof is deferred to Appendix B.

Theorem 5 *Let $X_1, \ldots, X_n$ be i.i.d. samples and denote by F (respectively F_n) the true CDF (respectively empirical CDF). Then, for any $x_0 \leq x_1 \in \mathbb{R} \cup \{\pm\infty\}$ and $t \geq 0$, if $V = \max_{x\in[x_0,x_1]} F(x)(1-F(x))$ (with the convention $F(-\infty) = 0$ and $F(+\infty) = 1$), we have*

$$\mathbb{P}\left(\sup_{x\in[x_0,x_1]} |F_n(x) - F(x)| > t\sqrt{\frac{V}{n}}\right) \leq c_1 e^{-c_2 \min(t^2, t\sqrt{nV})},$$

for some universal constants $c_1, c_2 > 0$.

A similar result recently appeared in Bartl and Mendelson (2023) which gives a variance-dependent DKW inequality. They show that for some absolute constants $c, c' > 0$ and any $t \geq c\sqrt{\ln \ln n}$,

$$\mathbb{P}\left(\exists x \in I_t \text{ s.t. } |F_n(x) - F(x)| > t\sqrt{\frac{F(x)(1-F(x))}{n}}\right) \leq 2e^{-c't^2},$$

where $I_t = \{x \mid t \leq \sqrt{nF(x)(1-F(x))}\}$ is precisely the set of points falling in the sub-Gaussian regime in our Theorem 5. Note that the width of their confidence-band depends on the variance of the empirical CDF at that point. This is in a contrast with our result, where having confidence-band of uniform width allowed us to derive a bound valid for all $t \geq 0$.

2.3. Non-asymptotic bounds for general distributions on $[0, 1]$

The results so far focused on the case when the distributions are supported on $\{0,1\}^{\mathbb{N}}$. However, some of the results can be generalized for the case of $[0,1]^{\mathbb{N}}$, as detailed below.

Corollary 6 *Let μ be a distribution on $[0,1]^{\mathbb{N}}$. Let $\sigma^2(i) = Var_{X\sim\mu}(X_i)$ for $i \geq 1$ be the variance of coordinate i. Without loss of generality, suppose that $\boldsymbol{\sigma}^2$ is decreasing.*

- *If for all $j \geq 1$, one has $\sigma^2(j) \leq \frac{1}{2nj}$, then*

$$\Delta_n(\mu) \lesssim \frac{1}{n} \wedge \sqrt{\frac{\sum_{j\geq1} \sigma^2(j)}{n}}.$$

- *Otherwise,*

$$\Delta_n(\mu) \lesssim 1 \wedge \sup_{j\geq1} \left(\sqrt{\frac{\sigma^2(j)\ln(j+1)}{n}} \vee \frac{\ln(j+1)}{n\ln\left(2 + \frac{\ln(j+1)}{n\sigma^2(j)}\right)}\right).$$

The proof is given in Appendix C. Given that the random variables are supported on $[0,1]$, we can use the inequality $\sigma^2(i) \leq p(i)$ for all $i \geq 1$ to obtain similar (but weaker) bounds as in Corollary 6 but replacing the variances $\sigma^2(i)$ by the means $p(i)$. As for the case of distributions on $\{0,1\}^{\mathbb{N}}$, the upper bounds from the previous result are not tight in general, however, these are the tightest bounds achievable if one only uses the variance statistic $\boldsymbol{\sigma}^2$. In particular, the case of independent Bernoulli random variables characterized in Theorem 1 always achieves the upper bound except in the regime when $\sum_{j\geq 1}\sigma^2(j) \leq \frac{1}{2n}$. In that case, we can show that the upper bound is attained not by random variables supported on $\{0,1\}$, but on $\{0, \sqrt{2n\sum_{j\geq 1}\sigma^2(j)}\}$. We refer to Appendix C for further details.

2.4. Expected empirical deviations in ℓ^q norms

While the infinite norm deviation $\Delta_n(\mu) = \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty$ is the main focus of this paper, a natural question is whether we can obtain similar results for general ℓ^q-norm expected deviations for $q \geq 1$. We have the following characterization for the decay of the expected ℓ^q deviation.

Proposition 7 *Let $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$. Then, $\lim_{n\to\infty} \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q = 0$ if and only if $\|\boldsymbol{p}\|_1 < \infty$. Moreover, if $\|\boldsymbol{p}\|_1 = \infty$, then $\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q = \infty$ $(a.s.)$.*

The proof is given in Appendix D. The analysis of the convergence of $\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q$ is quite different from the ℓ^∞ case since for instance the quantity $\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q^q$ can be computed directly as a sum of expectations. In particular, one can obtain bounds on the expected ℓ^q deviation $\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q$ using the following Jensen inequalities,

$$\left(\sum_{j\geq 1}(\mathbb{E}|\hat{p}_n(j) - p(j)|)^q\right)^{1/q} \leq \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \leq (\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q^q)^{1/q}.$$

These bounds give the correct asymptotic convergence rate of the expected ℓ^q deviation when $q \geq 2$ up to a factor $\Theta(\sqrt{q})$.

Proposition 8 *Let $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ such that $\|\boldsymbol{p}\|_1 < \infty$, and $q \geq 2$. Then,*

$$1 \lesssim \liminf_{n\to\infty} \sqrt{\frac{n}{\|\boldsymbol{p}\|_{q/2}}}\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \leq \limsup_{n\to\infty} \sqrt{\frac{n}{\|\boldsymbol{p}\|_{q/2}}}\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \lesssim \sqrt{q}.$$

Hence the convergence in this case is of the order $\sqrt{\frac{\|\boldsymbol{p}\|_{q/2}}{n}}$. The proof of this result as well as non-asymptotic bounds can be found in Appendix D.

2.5. High probability bounds for independent Bernoulli

While we focused on bounding the expectation of the maximal deviation, we also provide some high probability concentration bounds. From the bounded differences inequality (Boucheron et al., 2013, Thm. 6.2) – also known as McDiarmid's inequality – we can directly have for $\gamma \in (0,1)$,

$$\mathbb{P}\left(\left|\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty - \Delta_n(\boldsymbol{p})\right| \geq \sqrt{\frac{\ln\frac{2}{\gamma}}{2n}}\right) \leq \gamma.$$

Notably, this bound is often pessimistic and can be significantly tightened. To write the high-probability bounds concisely, we extend the definition of the quantities $\phi_{J,q}(n)$ to all reals $J > 0$. For $J \geq 1$, we extend the definition with the same formula in Eq (2). For $J \in (0,1)$, we pose

$$\phi_{J,q}(n) := e^{-1/J}\sqrt{\frac{q}{n}}.$$

We are now ready to state the high-probability bounds.

Proposition 9 *Let $\gamma \in (0, \frac{1}{2})$ and $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ such that there exists $j \geq 1$ with $p(j) \geq \frac{\gamma}{2nj}$. Then, for some universal constants $a_1, a_2 > 0$,*

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \geq a_1 \sup_{j\geq 1} \phi_{\frac{j}{\gamma}, p(j)}(n)\right) \leq \gamma.$$

Also,

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \leq a_2 \sup_{j\geq 1} \phi_{\frac{j}{\ln 1/\gamma}, p(j)}(n)\right) \leq \gamma.$$

Let $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ such that $p(j) \geq \frac{\gamma}{2nj}$ for all $j \geq 1$. Then,

$$\mathbb{P}\left(\sup_{j\geq 1} \hat{p}_n(j) \geq \frac{2}{n}\right) \leq \gamma \quad \textit{and} \quad \mathbb{P}\left(\sup_{j\geq 1} \hat{p}_n(j) = \frac{1}{n}\right) \asymp 1 \wedge n \sum_{j\geq 1} p(j).$$

The proof of this result and further high-probability bounds showing that these are tight in most cases can be found in Appendix E.

3. Expected maximum empirical mean deviation for product distributions

In this section, we give the main steps for the proof of our main characterization in Theorem 1. For the sake of conciseness, we only present sketches of the proofs here, all formal proofs of this section are given in Appendix A.

3.1. Preliminaries and general strategy

We first recall some basic tail inequalities for binomials. In the following, $D(q \parallel p) = q\ln(\frac{q}{p}) + (1-q)\ln(\frac{1-q}{1-p})$ is the KL-divergence between Bernoulli distributions with parameters $p, q \in [0,1]$. We start with the classical Chernoff bound (Boucheron et al., 2013).

Lemma 10 (Chernoff bound) *For any $0 \leq p \leq q \leq 1$, letting $Y \sim \mathcal{B}(n,p)$, we have*

$$\mathbb{P}\left(\frac{Y}{n} \geq q\right) \leq e^{-nD(q\|p)}.$$

We will also use the following anti-concentration bound from Zhang and Zhou (2020, Theorem 9).

Lemma 11 (Zhang and Zhou (2020)) *There exist constants $0 < c_0 < \frac{1}{4}$ and $C \geq 1$, such that for any $0 < p \leq q < 1$ satisfying $\frac{1}{n} \leq q \leq \frac{1+p}{2}$, letting $Y \sim \mathcal{B}(n,p)$, we have*

$$\mathbb{P}\left(\frac{Y}{n} \geq q\right) \geq c_0 e^{-CnD(q\|p)}.$$

As a first observation, defining $\Delta_n^+(\boldsymbol{p}) = \mathbb{E}\sup_{j\geq 1}[\hat{p}_n(j) - p(j)]_+$ and $\Delta_n^-(\boldsymbol{p}) = \mathbb{E}\sup_{j\geq 1}[p(j) - \hat{p}_n(j)]_+$, we have the following decomposition,

$$\frac{1}{2}(\Delta_n^+(\boldsymbol{p}) + \Delta_n^-(\boldsymbol{p})) \leq \Delta_n^+(\boldsymbol{p}) \vee \Delta_n^-(\boldsymbol{p}) \leq \Delta_n(\boldsymbol{p}) \leq \Delta_n^+(\boldsymbol{p}) + \Delta_n^-(\boldsymbol{p}).$$

We will show in the rest of this paper that the leading term is $\Delta_n^+(\boldsymbol{p})$, which we now focus on. The main intuition is that Bernoulli random variables $\mathcal{B}(p)$ with $p \leq 1/2$ have heavier right tails than left tails. To give estimates for $\Delta_n^+(\boldsymbol{p})$ for general values of the sequence $\boldsymbol{p}$, we first start with a reduction to the case when the profile of $\boldsymbol{p}$ is "step-like". Consider such a vector $\mathbf{step}_{J,q}$ with $p(i) = q$ for all $i \in [J]$ and $p(i) = 0$ for $i > J$. Then,

$$\mathbb{P}\left(\max_{i\leq J}\{\hat{p}(i) - p(i)\} \geq \varepsilon\right) = 1 - (1 - \mathbb{P}(\hat{p}(1) - p(1) \geq \varepsilon))^J.$$

Intuitively, this probability is approximately $1 - \exp(J\mathbb{P}(\hat{p}(1) - p(1) \geq \varepsilon))$. If ε is the expected maximal deviation, then one would expect this probability above to be bounded away from both 0 and 1 by some absolute constants. This motivates the definition of the following quantity $\varepsilon_{J,q}(n)$ for any $J \geq 1$ and $q \in (0, 1/2]$, where $c_0 \in (0, \frac{1}{4})$ is the same constant as in Lemma 11,

$$\varepsilon_{J,q}(n) = \inf\left\{\varepsilon \geq 0 : \mathbb{P}_{Y\sim\mathcal{B}(n,q)}\left(\frac{Y}{n} \geq q + \varepsilon\right) \leq \frac{c_0}{2J}\right\},$$

In particular, note that $q + \varepsilon_{J,q}(n) \in \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}$ and that

$$\mathbb{P}_{Y\sim\mathcal{B}(n,q)}\left(\frac{Y}{n} > q + \varepsilon_{J,q}(n)\right) \leq \frac{c_0}{2J} < \mathbb{P}_{Y\sim\mathcal{B}(n,q)}\left(\frac{Y}{n} \geq q + \varepsilon_{J,q}(n)\right).$$

Our goal is to give a characterization of $\Delta_n^+(\boldsymbol{p})$ using these coefficients.

3.2. Step-like sequences describe the behavior of general sequences

It turns out that not only $\Delta_n^+(\mathbf{step}_{J,q}) \asymp \varepsilon_{J,q}$, but we even have $\Delta_n^+(\boldsymbol{p}) \asymp \sup_{i\geq 1}\varepsilon_{i,p(i)}$ for most vectors $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ as shown in the following result.

Proposition 12 *Let $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$. Suppose that there exists $i \geq 1$ such that $\varepsilon_{i,p(i)}(n) \geq 0$. Then, there exist universal constants $c, C > 0$ such that for all $n \geq 2$,*

$$c \cdot \sup_{i\geq 1}\varepsilon_{i,p(i)}(n) \leq \Delta_n^+(\boldsymbol{p}) \leq C \cdot \sup_{i\geq 1}\varepsilon_{i,p(i)}(n).$$

Further, the upper bound holds for any general distribution μ on $\{0,1\}^{\mathbb{N}}$, that is, with $\boldsymbol{p} = \mathbb{E}_{X\sim\mu}[X]$,

$$\mathbb{E}_{X_i \overset{i.i.d.}{\sim} \mu}\sup_{i\in\mathbb{N}}[\hat{p}_n(i) - p(i)]_+ \leq C \cdot \sup_{i\geq 1}\varepsilon_{i,p(i)}(n).$$

Sketch of proof. We start with the lower bound $\Delta_n^+(\boldsymbol{p}) \gtrsim \sup_{i\geq 1} \varepsilon_{i,p(i)}$. For any index $i \geq 1$ it holds that $\boldsymbol{p} \geq \mathbf{step}_{i,p(i)}$ element-wise and thus we can prove that we have $\Delta_n^+(\boldsymbol{p}) \gtrsim \Delta_n^+\left(\mathbf{step}_{i,p(i)}\right)$ since the tails at the individual coordinates are heavier for $\boldsymbol{p}$. Now pick index J such that $\varepsilon := \varepsilon_{J,p(J)} \geq \frac{1}{2}\sup_{i\geq 1} \varepsilon_{i,p(i)}$, and let $q = p(J)$; then for independent $Y_i \sim \mathcal{B}(n,q)$:

$$\mathbb{P}\left(\max_{i\leq J}\left\{\frac{Y_i}{n}\right\} \geq q+\varepsilon\right) = 1-\left(1-\mathbb{P}\left(\frac{Y_i}{n}\geq q+\varepsilon\right)\right)^J \geq 1-\left(1-\frac{c_0}{2J}\right)^J \geq 1-e^{-c_0/2}.$$

We finish by applying Markov's inequality:

$$\mathbb{E}\left[\max_{i\leq J}\left|\frac{Y_i}{n}-q\right|\right] \geq \mathbb{E}\left[\max_{i\leq J}\left\{\frac{Y_i}{n}\right\}-q\right] \geq \varepsilon\mathbb{P}\left(\max_{i\leq J}\left\{\frac{Y_i}{n}\right\}\geq q+\varepsilon\right) \asymp \varepsilon.$$

That is, $\Delta_n(\boldsymbol{p}) \gtrsim \varepsilon_{J,q} \asymp \sup_{i\geq 1}\varepsilon_{i,p(i)}$.

We next turn to the upper bound $\Delta_n^+(\boldsymbol{p}) \lesssim \sup_{i\geq 1}\varepsilon_{i,p(i)}$. We use a pair of tight concentration and anti-concentration inequalities to estimate tail probabilities of binomial random variables and then we upper bound the expectation by tail-summation. Let $\varepsilon = \sup_{i\geq 1}\varepsilon_{i,p(i)}(n)$. Thus, for every index i it holds that

$$\mathbb{P}\left(\hat{p}(i) \geq p(i)+\varepsilon\right) \lesssim \frac{1}{2i}.$$

On the other hand, the anti-concentration inequality from Lemma 11 for $C \geq 1$ gives

$$\mathbb{P}\left(\hat{p}(i)\geq p(i)+\varepsilon\right) \geq c_0 e^{-CnD(p(i)+\varepsilon\|p(i))}.$$

Combining both estimates results in

$$D(p(i)+\varepsilon \parallel p(i)) \gtrsim \frac{\ln(2i)}{Cn}.$$

By a convexity argument on the KL-divergence we obtain $D(p(i)+kC\varepsilon \parallel p(i)) \gtrsim k\ln(2i)/n$ for $k\geq 1$. Together with the standard Chernoff bound (Lemma 10)

$$\mathbb{P}\left(\hat{p}(i)\geq p(i)+kC\varepsilon\right) \leq e^{-nD(p(i)+kC\varepsilon\|p(i))} \leq \frac{1}{(2i)^k}.$$

Then, by the union bound, for any $k\geq 2$,

$$\mathbb{P}\left(\sup_{i\geq 1}\left\{\hat{p}(i)-p(i)\right\}\geq kC\varepsilon\right) \leq \sum_{i\geq 1}\frac{1}{(2i)^k} \leq \frac{1}{2^{k-1}}.$$

Finally, summing the tails yields the desired bound $\Delta_n^+(\boldsymbol{p}) = \mathbb{E}\sup_{i\geq 1}\left[\hat{p}(i)-p(i)\right]_+ \lesssim \varepsilon$. Note that in the proof of this upper bound, we only needed the union bound. Hence, the upper bound also applies to general non-product distributions μ on $\{0,1\}^{\mathbb{N}}$. ■

3.3. Estimating the expected deviation for step-like sequences

We next give estimates on the quantities $\varepsilon_{J,q}$.

Proposition 13 *There exists a universal constants $C_1 \geq 1$ and $c_2 > 0$ such that for all $n, J \geq 1$ and $q \in (0, 1/2]$,*

$$\varepsilon_{J,q}(n) \leq C_1 \cdot \phi_{J,q}(n).$$

Further, if $1-(1-q)^n > \frac{c_0}{2J}$ (e.g. for $q \geq \frac{c_0}{nJ}$), one has

$$\varepsilon_{J,q}(n) \geq \left(\frac{1}{n} - q\right) \vee c_2 \cdot \phi_{J,q}(n).$$

On the other hand, if $1-(1-q)^n \leq \frac{c_0}{2J}$, we have $\varepsilon_{J,q}(n) = -q$.

Sketch of proof. Informally, the concentration and anti-concentration inequalities from Lemma 10 and Lemma 11 show that in most cases one has

$$\ln\left(\frac{c_0}{2J}\right) \approx \ln \mathbb{P}\left(\frac{Y}{n} \geq q + \varepsilon_{J,q}(n)\right) \asymp -nD(q + \varepsilon_{J,q}(n) \parallel q).$$

These estimates are not tight in the "Poissonian" regime when $q \lesssim \frac{1}{nJ}$ which has to be treated separately. Otherwise, $\varepsilon_{J,q}(n)$ is essentially a solution to $D(q+\varepsilon \parallel q) \asymp \frac{\ln(J+1)}{n}$ in ε. The KL divergence shows two major regimes: either $D(q+\varepsilon \parallel q) \asymp \varepsilon \ln \frac{\varepsilon}{q}$ or $D(q+\varepsilon \parallel q) \asymp \frac{\varepsilon^2}{q}$. As a remark, these two regimes for the KL divergence are equivalent to the two standard regimes for Bennett's inequality (see Boucheron et al. (2013, Subsection 2.7) for a more detailed overview of this inequality, or Lemma 16 for a precise statement). These two asymptotic behaviors translate into the second and third regimes in the definition of $\phi_{J,q}(n)$ respectively.

Case 1: $\frac{\ln(J+1)}{n} \asymp D(q+\varepsilon_{J,q}(n) \parallel q) \asymp \varepsilon_{J,q}(n) \ln \frac{\varepsilon_{J,q}(n)}{q}$. In this case we obtain $\varepsilon_{J,q}(n) \asymp \frac{\ln(J+1)}{n \ln\left(\frac{\ln(J+1)}{nq}\right)}$.

Case 2: $\frac{\ln(J+1)}{n} \asymp D(q+\varepsilon_{J,q}(n) \parallel q) \leq \frac{\varepsilon_{J,q}(n)^2}{q}$. This corresponds to a sub-Gaussian regime and we obtain $\varepsilon_{J,q}(n) \asymp \sqrt{\frac{q\ln(J+1)}{n}}$. ∎

We are now ready to complete the proof of Theorem 1.

Sketch of proof of Theorem 1 When some entries of p are sufficiently large so that one can use the estimates from Proposition 13 we combine it with Proposition 12. This shows that $\Delta_n^+(\boldsymbol{p}) \asymp \sup_{i\geq 1} \phi_{i,p(i)}(n)$ whenever there exists $j \geq 1$ with $p(j) \geq \frac{c_0}{nj}$.

We treat separately the remaining case when for all $j \geq 1$, one has $p(j) \leq \frac{c_0}{nj}$. This corresponds to a Poissonian regime and it suffices to characterize the probability that one of the coordinates $\hat{p}_n(i)$ for $i \geq 1$ is non-zero. In this case, we obtain $\Delta_n^+(\boldsymbol{p}) \asymp \frac{1}{n} \wedge \sum_{j\geq 1} p(j)$. Proving that the leading term in $\Delta_n(\boldsymbol{p}) \asymp \Delta_n^-(\boldsymbol{p}) + \Delta_n^+(\boldsymbol{p})$ is indeed $\Delta_n^+(\boldsymbol{p})$ ends the proof of our main characterization in Theorem 1. ∎

3.4. Discussion and comparison with bounds from the literature.

We first give some intuition on the decay of $\Delta_n(\boldsymbol{p})$ given in Theorem 1. The first case when for all $j \geq 1$, one has $p(j) \leq 1/(2nj)$ corresponds to rare events scenarios such that with high probability, $\hat{p}(j) \leq 1/n$ for all $j \geq 1$. This is characterized by the term $1/n$ from the bound

$\Delta_n(\boldsymbol{p}) \asymp 1/n \wedge \sum_{j\geq 1} p(j)$. The second term characterizes the probability of the event when $\sup_{j\geq 1} \hat{p}_n(j) \geq \frac{1}{n}$. In this low-probability regime, the probabilities of success can be summed: with probability $\asymp 1 \wedge \sum_{j\geq 1} np(j)$, at least one of the binomials $n\hat{p}_n(j)$ is nonzero.

We next turn to the second case when there exists $j \geq 1$ for which $p(j) \geq 1/(2nj)$. In this case, it is useful to compare our bounds using the functionals $S(\boldsymbol{p})$ and $T(\boldsymbol{p})$ from the literature. In particular, as a direct consequence of the characterization, we can recover the lower bound

$$\Delta_n(\boldsymbol{p}) \gtrsim 1 \wedge \frac{T(\boldsymbol{p})}{n}$$

from Cohen and Kontorovich (2022). By definition of $\phi_{j,p(j)}$, for $n \leq T(\boldsymbol{p}) = \sup_{j\geq 1} \frac{\ln(j+1)}{\ln 1/p(j)}$, we have $\Delta_n(\boldsymbol{p}) \asymp \sup_{j\geq 1} \phi_{j,p(j)}(n) = 1$. When $n > T(\boldsymbol{p})$, the functions $\phi_{j,p(j)}(n)$ fall in either of the last two regimes (see Eq (2)), hence

$$\begin{aligned}
\Delta_n(\boldsymbol{p}) &\asymp \sqrt{\frac{S(\boldsymbol{p})}{n}} \vee \frac{1}{n} \sup_{j\geq 1} \frac{\ln(j+1)}{\ln\left(2 + \frac{\ln(j+1)}{np(j)}\right)} \\
&\gtrsim \sqrt{\frac{S(\boldsymbol{p})}{n}} \vee \frac{1}{n} \sup_{j\geq 1} \frac{\ln(j+1)}{\ln \frac{\ln 1/p(j)}{p(j)}} \asymp \sqrt{\frac{S(\boldsymbol{p})}{n}} \vee \frac{T(\boldsymbol{p})}{n}.
\end{aligned}$$

Together with the previous case, this shows that

$$\Delta_n(\boldsymbol{p}) \gtrsim 1 \wedge \left(\sqrt{\frac{S(\boldsymbol{p})}{n}} \vee \frac{T(\boldsymbol{p})}{n}\right).$$

As suggested by the derivation, this lower bound is tight for n in the neighborhood of $T(\boldsymbol{p})$. For instance consider a step-like parameter $\mathbf{step}_{J,q}$ with $q \geq \frac{1}{2nJ}$. Fix a constant $0 \leq a < 1$, then for any $T(\boldsymbol{p}) = \frac{\ln(J+1)}{\ln 1/q} \leq n \leq \frac{\ln(J+1)}{q^a}$, Theorem 1 implies

$$\frac{T(\boldsymbol{p})}{n} \lesssim \phi_{J,q}(n) \asymp \Delta_n(\boldsymbol{p}) \lesssim \frac{\ln(J+1)}{n \ln 1/q^{1-a}} = \frac{1}{1-a} \frac{T(\boldsymbol{p})}{n}.$$

In terms of upper bounds, we recover the bound Eq (1). To do so, we give an upper bound of the functions $\phi_{J,q}(n)$ for $q \geq \frac{1}{2nJ}$ in the second regime from Eq (2) for which $\frac{\ln(J+1)}{\ln 1/q} \leq n \leq \frac{\ln(J+1)}{eq}$. First, note that the previous equation shows that for $n \leq \frac{\ln(J+1)}{\sqrt{q}}$, one has $\Delta_n(\boldsymbol{p}) \lesssim \frac{T(\boldsymbol{p})}{n}$. Therefore it remains to consider the case when $\frac{\ln(J+1)}{\sqrt{q}} \leq n \leq \frac{\ln(J+1)}{eq}$. Note that $\ln \frac{1}{q} \leq 2 \ln \frac{n}{\ln(J+1)} \leq 2 \ln n$. Hence, in that regime,

$$\phi_{J,q}(n) \leq \frac{\ln(J+1)}{n} \lesssim \frac{\ln(J+1)}{n \ln \frac{1}{q}} \ln n.$$

Together with Theorem 1, this implies

$$\Delta_n(\boldsymbol{p}) \asymp \sup_{j\geq 1} \phi_{j,p(j)}(n) \asymp \sup_{j\geq 1, p(j) \geq \frac{1}{2nj}} \phi_{j,p(j)}(n) \lesssim 1 \wedge \left(\sqrt{\frac{S(\boldsymbol{p})}{n}} + \frac{T(\boldsymbol{p})}{n} \ln n\right).$$

Here, we used the fact that terms $\phi_{j,p(j)}(n)$ for which $p(j) \leq \frac{1}{2nj}$ are not dominant. This is formally shown in Appendix A in the proof of Proposition 20. Note that the estimates are tight for n in the neighborhood of the beginning of the sub-Gaussian regime when the term $\sqrt{\frac{S(\boldsymbol{p})}{n}}$ dominates.

As a summary of this discussion, assuming that there exists $j \geq 1$ for which $p(j) \geq \frac{1}{2nj}$, the decay of $\Delta_n(\boldsymbol{p})$ shows three main regimes:

- $\Delta_n(\boldsymbol{p}) \asymp 1$ when $n \leq T(\boldsymbol{p})$,
- a somewhat sub-exponential regime when the decay of $\Delta_n(\boldsymbol{p})$ interpolates between $\frac{T(\boldsymbol{p})}{n}$ towards the start, and $\frac{T(\boldsymbol{p}) \ln n}{n}$ towards the end of this regime,
- the asymptotic sub-Gaussian regime $\Delta_n(\boldsymbol{p}) \asymp \sqrt{\frac{S(\boldsymbol{p})}{n}}$.

4. Expected maximum deviation for arbitrarily correlated distributions

In this section, we prove our estimate on the expected maximum empirical deviation for correlated distributions μ on $\{0,1\}^{\mathbb{N}}$ from Corollary 3. The latter requires the localized version of the classical Dvoretzky-Kiefer-Wolfowitz theorem given in Theorem 5. For our purposes, we only need the result on intervals $(-\infty, x_0]$.

Corollary 14 *Let $X_1, \ldots, X_n$ be i.i.d. samples and denote by F (respectively F_n) the true CDF (respectively empirical CDF). Then, for any $x_0 \in \mathbb{R}$ and $t \geq 0$,*

$$\mathbb{P}\left(\sup_{x \leq x_0} |F_n(x) - F(x)| > t\sqrt{\frac{F(x_0)}{n}}\right) \leq c_1 e^{-c_2 \min(t^2, t\sqrt{nF(x_0)})},$$

for some universal constants $c_1, c_2 > 0$.

The proof of both Theorem 5 and Corollary 14 are given in Appendix B.

We start by proving Corollary 3 that gives estimates of $\Delta_n(\mu)$ for general distributions μ on $\{0,1\}^{\mathbb{N}}$.

Proof of Corollary 3 All the upper bounds derived in the proof of Theorem 1 either used the union bound or Markov's inequality. We point in particular to Theorem 12 which gives the main upper bound whenever there exists $i \geq 0$ for which $\varepsilon_{i,p(i)}(n) \geq 0$. As a result, these still hold in the case of general distributions μ on $\{0,1\}^d$ with mean $p = \mathbb{E}_{X \sim \mu}[X]$. We also provide some simple lower bounds which correspond from only considering the deviation from the first coordinate.

$$\Delta_n(\mu) = \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \geq \mathbb{E}|\hat{p}_n(1) - p(1)| \asymp \frac{np(1) \wedge \sqrt{np(1)}}{n} = p(1) \wedge \sqrt{\frac{p(1)}{n}}$$

The last estimate is classical and can be found for instance in Berend and Kontorovich (2013b). ■

While the bounds provided in Corollary 3 may not be tight, the following result shows that if one only has access to the mean statistic $p = \mathbb{E}_{X \sim \mu}[X]$, these are tight.

Proposition 15 *Let* $p \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$. *There exists a distribution* μ *on* $\{0,1\}^{\mathbb{N}}$ *with* $\mathbb{E}_{X\sim\mu}[X] = p$ *such that for all* $n \geq 1$,

$$\Delta_n(\mu) \asymp p(1) \wedge \sqrt{\frac{p(1)}{n}}.$$

Proof We start by constructing the corresponding distribution for $X \sim \mu$. We use a standard coupling which allows having samples X with non-increasing coordinates. Precisely, let $U \sim \mathcal{U}([0,1])$ be a uniform random variable. We let $X(j) = \mathbb{1}[U \leq p(j)]$. We now show that this distribution satisfies $\Delta_n(\mu) \asymp p(1) \wedge \sqrt{\frac{p(1)}{n}}$. With n i.i.d. samples $X_i \sim \mu$, we define $Y = \sum_{i=1}^n X_i$ and $\hat{p}_n = \frac{Y}{n}$. Because of discretization issues, we distinguish several cases.

We start by proving the upper bound $\Delta_n(\mu) \lesssim p(1)$. Suppose that $\hat{p}_n(1) \geq 2p(1)$. Then, we have for any $j \geq 2$, $|\hat{p}_n(j) - p(j)| \leq p(1) + |\hat{p}_n(1) - p(1)| \leq 2|\hat{p}_n(1) - p(1)|$. Hence,

$$\Delta_n(\boldsymbol{p}) \leq 2p(1) + 2\mathbb{E}|\hat{p}_n(1) - p(1)|.$$

Because $\mathbb{E}|\hat{p}_n(1) - p(1)| \asymp p(1) \wedge \sqrt{\frac{p(1)}{n}}$, we obtain the desired bound $\Delta_n(\mu) \lesssim p(1)$.

We now turn to the upper bound $\Delta_n(\mu) \lesssim \sqrt{\frac{p(1)}{n}}$. Without loss of generality, we can therefore suppose that $p(1) \geq \frac{1}{n}$. Note that if $F_n(\cdot)$ is the empirical cumulative distribution function obtained from the i.i.d. uniform samples $U_1, \ldots, U_n$ used to define the variables $X_1, \ldots, X_n$,

$$\Delta_n(\mu) = \mathbb{E} \sup_{j \geq 1} |F_n(p(j)) - p(j)| \leq \mathbb{E} \sup_{u \leq p(1)} |F_n(u) - u|.$$

We then obtain bounds on the right-hand side using the localized version of the DKW theorem from Corollary 14. Recalling that $p(1) \geq \frac{1}{n}$, this result implies that for $t > 0$,

$$\mathbb{P}\left(\sup_{u \in [0,1]} |F_n(u) - u| > t\sqrt{\frac{p(1)}{n}} \right) \leq c_1 e^{-c_2 t}$$

for universal constants $c_1, c_2 > 0$. As a result, we obtain

$$\Delta_n(\mu) \leq \mathbb{E} \sup_{u \in [0, p(1)]} |F_n(u) - u| \lesssim \sqrt{\frac{p(1)}{n}}.$$

This ends the proof of the proposition. ■

5. On the open problem from Cohen and Kontorovich (2022)

As a consequence of our characterization, we can answer the COLT open problem posed by Cohen and Kontorovich (2023), showing that the $\ln n$ factor the bound Eq (1) is necessary.

Theorem 2 *Suppose that there exists a constant* $C \geq 1$ *and* $n_0 \geq 1$, *and a function* $\psi : \mathbb{N} \to \mathbb{R}$ *such that the inequality*

$$\Delta_n(\boldsymbol{p}) \leq C\sqrt{\frac{S(\boldsymbol{p})}{n}} + \frac{T(\boldsymbol{p})}{n}\psi(n)$$

holds for all $n \geq n_0$ *and* $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ *(product measures), then for an integer* n_1 *and a constant* $c > 0$ *depending only on* C,

$$\psi(n) \geq c \ln n, \quad n \geq n_0 \vee n_1.$$

Proof Consider the sequence

$$\begin{cases} p(j) = \frac{1}{K\sqrt{n}} := p, & \text{if } \ln(j+1) \le K\sqrt{n} \\ p(j) = 0, & \text{otherwise} \end{cases}$$

for a parameter $2 \le K \le \sqrt{n}$ to define later. We denote by J be the largest integer such that $p(J) = p$, in particular, we still have $\ln(J+1) \le K\sqrt{n}$. Then,

$$p(J) = \frac{1}{K\sqrt{n}} \ge \frac{\ln(J+1)}{K^2 n} \ge \frac{1}{2K^2 n}.$$

Now note that $\ln(J+2) > K\sqrt{n} \ge \sqrt{n}$, hence $K \le \sqrt{J}$ for any $n \ge n_1$ for some constant $n_1 \ge 1$. Together with the previous equation, we have $p(J) \ge \frac{1}{2Jn}$ so from Theorem 1 it follows

$$\begin{aligned} \Delta_n(\boldsymbol{p}) &\asymp 1 \wedge \sup_{j \ge 1} \left(\frac{\ln(j+1)}{n \ln\left(2 + \frac{\ln(j+1)}{np}\right)} + \sqrt{\frac{p(j)\ln(j+1)}{n}} \right) \\ &\asymp \frac{K}{\sqrt{n}\ln K} + \frac{1}{\sqrt{n}}. \end{aligned}$$

Hence, for $K \gtrsim C^2$ which can be achieved for n sufficiently large (depending on C),we obtain $\frac{1}{2}\Delta_n(\boldsymbol{p}) \ge C\sqrt{\frac{S(\boldsymbol{p})}{n}}$. On the other hand, note that

$$T(\boldsymbol{p}) = \frac{\ln(J+1)}{\ln \frac{1}{p}} \le \frac{2K\sqrt{n}}{\ln n}.$$

Therefore, we obtain $\psi(n) \gtrsim \frac{\ln n}{\ln K} \asymp \frac{\ln n}{\ln C}$. ■

Although the previous proof used the general characterization of $\Delta_n(\boldsymbol{p})$, Theorem 2 can be proved with elementary arguments. We provide below a simple proof to obtain the same lower bound on $\Delta_n(\boldsymbol{p})$.

Proof (elementary) of Theorem 2 We use the same example for which $p(j) = \frac{1}{K\sqrt{n}} := p$ if $\ln(j+1) \le K\sqrt{n}$, and $p(j) = 0$ otherwise. The parameter $4 \le K \le \sqrt{n}$ will be fixed later. We denote by J the largest integer such that $p(J) = p$. Let $l = \lfloor \frac{K\sqrt{n}}{4\ln K} \rfloor$. For any $j \le J$,

$$\begin{aligned} \mathbb{P}\left(\hat{p}_n(j) \ge \frac{l}{n}\right) &\ge \binom{n}{l} p^l (1-p)^{n-l} \\ &\ge \left(\frac{n}{l}\right)^l p^l (1-p)^n \\ &\ge \exp\left(-l \ln \frac{l}{np} - 2np\right) \\ &\ge \exp\left(-\frac{K\sqrt{n}}{4\ln K} \ln \frac{K^2}{4\ln K} - \frac{2\sqrt{n}}{K}\right) \\ &\ge \exp\left(-\frac{1}{2}K\sqrt{n} - \frac{2\sqrt{n}}{K}\right) \ge e^{-\frac{3}{4}K\sqrt{n}}. \end{aligned}$$

In the third inequality, we used the fact that $\ln(1-x) \geq -2x$ for $x \in [0, \frac{1}{2}]$. Now note that $\ln(J+2) > K\sqrt{n}$. Therefore,

$$\begin{aligned}
\mathbb{P}\left[\max_{j\in[J]} \hat{p}_n(j) \geq \frac{l}{n}\right] &\geq 1 - \left(1 - \mathbb{P}\left[\hat{p}_n(1) \geq \frac{l}{n}\right]\right)^J \\
&\geq 1 - \exp\left(-J \cdot \mathbb{P}\left[\hat{p}_n(1) \geq \frac{l}{n}\right]\right) \\
&\geq 1 - \exp\left(-(e^{K\sqrt{n}} - 2)e^{-\frac{3}{4}K\sqrt{n}}\right) \geq 1 - e^{2/e^2} \geq \frac{1}{2}.
\end{aligned}$$

Hence, for n sufficiently large

$$\Delta_n(\boldsymbol{p}) \geq \frac{1}{2}\left(\frac{l}{n} - p\right) \geq \frac{K}{10\sqrt{n}\ln K}.$$

This provides the same lower bound for $\Delta_n(\boldsymbol{p})$ as in the previous proof up to constants, and the proof is identical from that point. ■

6. Conclusion and future work

In this paper, we have derived the exact characterization (up to a constant factor) of the infinite-norm deviation of the empirical mean of the distribution supported on $\{0,1\}^d$ from the true mean in the case of product distributions. For the case of general (non-product) distributions, we have derived a lower and upper bound on the deviation when we only have access to the mean statistics, and provided distributions corresponding to these bounds. Along the way, we proved a localized version of Dvoretzky–Kiefer–Wolfowitz inequality. We extended the results to the cases where the deviation is measured in a general $q-$norm and provided characterization of the convergence, and both finite and assymptotic bounds on the convergence. Additionally, we considered the case where the random variables were supported on $[0,1]^d$ instad of $\{0,1\}^d$ and we have derived a lower and upper bound on the deviation when we only have access to the mean (or variance) statistics, and provided distributions corresponding to these bounds.

An interesting direction for future work would be to consider the case of dependent coordinates with information about how they are dependent; e.g., when we know the covariance matrix. Exact non-asymptotic bounds would be of particular interest, but to the best of our knowledge even characterizing the asymptotic behavior of the maximum deviation for arbitrary distributions is an open question.

Acknowledgments

The authors would like to mention that concerning the COLT 2023 open problem Cohen and Kontorovich (2023) of removing the factor $\ln n$ in Eq (1), three independent examples were first (almost) simultaneously given by Jaouad Mourtada and both authors of this paper, showing that a factor $\ln n/\ln\ln n$ is at least necessary in general. These examples were then improved in Theorem 2 of this paper to show that the full $\ln n$ factor is necessary. The authors would like to express their sincere gratitude to Aryeh Kontorovich for carefully updating a timeline of that progress at https://mathoverflow.net/questions/447472 and for invaluable discussions on

Local Glivenko-Cantelli bounds; and to Jaouad Mourtada for useful discussions and advice on the exposition of these examples. VV was supported by the DFG Cluster of Excellence "Machine Learning – New Perspectives for Science", EXC 2064/1, project number 390727645 and is thankful for the support of Open Philanthropy.

References

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials. *The Annals of Statistics*, 47(4):1893–1927, 2019.

Daniel Bartl and Shahar Mendelson. On a variance dependent Dvoretzky-Kiefer-Wolfowitz inequality, 2023.

Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18:1–7, 2013a.

Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013b.

S Boucheron, G Lugosi, and P Massart. Concentration inequalities: A nonasymptotic theory of independence. univ. press, 2013.

V Buldygin and K Moskvichova. The sub-gaussian norm of a binary random variable. *Theory of probability and mathematical statistics*, 86:33–49, 2013.

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 *Conference on Learning Theory*, pages 786–806. PMLR, 2019.

Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter Bartlett, and Michael Jordan. Optimal mean estimation without a variance. In *Conference on Learning Theory*, pages 356–357. PMLR, 2022.

Julien Chhor and Alexandra Carpentier. Sharp local minimax rates for goodness-of-fit testing in multivariate binomial and poisson families and in multinomials. *arXiv preprint arXiv:2012.13766*, 2020.

Julien Chhor and Alexandra Carpentier. Goodness-of-fit testing for h\" older-continuous densities: Sharp local minimax rates. *arXiv preprint arXiv:2109.04346*, 2021.

Julien Chhor, Rajarshi Mukherjee, and Subhabrata Sen. Sparse signal detection in heteroscedastic gaussian sequence models: Sharp minimax rates. *arXiv preprint arXiv:2211.08580*, 2022.

Doron Cohen and Aryeh Kontorovich. Local Glivenko-Cantelli. *arXiv preprint arXiv:2209.04054*, 2022.

Doron Cohen and Aryeh Kontorovich. Open problem: log(n) factor in "Local Glivenko-Cantelli. *COLT*, 2023.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-gaussian mean estimators. 2016.

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

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

Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

Michael Kearns and Lawrence Saul. Large deviation methods for approximate probabilistic inference. *arXiv preprint arXiv:1301.7392*, 2013.

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. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019a.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. 2019b.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. 2021.

Odalric-Ambrym Maillard. Local Dvoretzky–Kiefer–Wolfowitz confidence bands. *Mathematical Methods of Statistics*, 30(1-2):16–46, 2021.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

Thomas. Is uniform convergence faster for low-entropy distributions? in theoretical computer science stack exchange. 2018. URL https://cstheory.stackexchange.com/questions/42009.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

Anru R Zhang and Yuchen Zhou. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314, 2020.

Appendix A. Proofs of Section 3

We start by giving a proof of Proposition 12 that allows for reducing the problem of characterizing the expected maximum empirical mean deviation for general probability vectors $p \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ to step-like vectors $\mathbf{step}_{J,q}$ constant equal to q until coordinate J then zero afterwards.

Proof of Proposition 12 We start with the lower bound. Here, we will mainly show that Bernoulli random variables with higher mean (which is smaller than 0.5) cannot have much lighter tails compared to the lower mean ones; thus, when we would have that $\Delta_n(\boldsymbol{p}) \geq \Delta_n(\mathbf{step}_{i,p(i)})$ as $\boldsymbol{p} \geq \mathbf{step}_{i,p(i)}$ coordinate-wise.

Fix $i \geq 1$ such that $\varepsilon_i = \varepsilon_{i,p(i)}(n) \geq \frac{1}{2}\sup_{j\geq 1}\varepsilon_{j,p(j)}(n)$. Let $j \leq i$. By construction, one has $p(j) \geq p(i)$. If $p(j) \leq \frac{\varepsilon_i}{2}$, one has

$$\mathbb{P}\left(\hat{p}_n(j) \geq p(j) + \frac{\varepsilon_i}{2}\right) \geq \mathbb{P}(\hat{p}_n(j) \geq \varepsilon_i) \geq \mathbb{P}(\hat{p}_n(i) \geq p(i) + \varepsilon_i) > \frac{c_0}{2i}. \tag{3}$$

We now suppose that $p(j) > \frac{\varepsilon_i}{2}$. Next, because $p(i) + \varepsilon_i \in \frac{1}{n}\mathbb{Z}$ and $\varepsilon_i > 0$, we obtain $p(j) + \varepsilon_i \geq p(i) + \varepsilon_i \geq \frac{1}{n}$.

We first treat the case when $p(j) > \frac{1}{n}$. First, the Chernoff bound from Lemma 10 shows that

$$\frac{c_0}{2i} < \mathbb{P}(\hat{p}_n(i) \geq p(i) + \varepsilon_i) \leq e^{-nD(p(i)+\varepsilon_i \| p(i))}.$$

Hence, $D(p(i) + \varepsilon_i \parallel p(i)) \leq \frac{1}{n}\ln\frac{2i}{c_0}$. Now note that the function $p \in [0, \frac{1-\varepsilon_i}{2}] \mapsto D(p + \varepsilon_i \parallel p)$ is non-increasing (and convex), so that if $p(j) \leq \frac{1-\varepsilon_i}{2}$, one has

$$D(p(j) + \varepsilon_i \parallel p(j)) \leq D(p(i) + \varepsilon_i \parallel p(i)).$$

On the other hand, if $p(j) > \frac{1-\varepsilon_i}{2}$, then

$$D\left(p(j) + \frac{\varepsilon_i}{2} \parallel p(j)\right) \leq D\left(\frac{1+\varepsilon_i}{2} \parallel p(j)\right) \leq D\left(\frac{1+\varepsilon_i}{2} \parallel \frac{1-\varepsilon_i}{2}\right) \leq D(p(i) + \varepsilon_i \parallel p(i)),$$

where in the last inequality, we also used the fact that $p \in [0, \frac{1-\varepsilon_i}{2}] \mapsto D(p + \varepsilon_i \parallel p)$ is non-increasing. In both cases, using the convexity of the KL-divergence in the first argument, we obtain

$$D\left(p(j) + \frac{\varepsilon_i}{2C} \parallel p(j)\right) \leq \frac{1}{C}D\left(p(j) + \frac{\varepsilon_i}{2} \parallel p(j)\right) \leq \frac{1}{nC}\ln\frac{2i}{c_0}.$$

Now because $p(j) \geq \frac{1}{n}$, we can use Lemma 11 (without loss of generality we can suppose $C \geq 2$ so that $p(j) + \frac{\varepsilon_i}{2C} \leq p(j) + \frac{1}{4} \leq \frac{1+p(j)}{2}$ which ensures that we can apply Lemma 11) which gives

$$\mathbb{P}\left(\hat{p}_n(j) \geq p(j) + \frac{\varepsilon_i}{2C}\right) \geq \frac{c_0^2}{2i}. \tag{4}$$

It remains to consider the case when $p(j) \leq \frac{1}{n}$. Recall that we have $p(j)+\varepsilon_i \geq \frac{1}{n}$ and $p(j) \geq \frac{\varepsilon_i}{2}$. As a result, $\frac{1}{3n} \leq p(j) \leq \frac{1}{n}$, and $\varepsilon_i \leq \frac{2}{n}$. Now observe that for $n \geq 2$,

$$\begin{aligned}
\mathbb{P}\left(\hat{p}_n(j) \geq p(j) + \frac{1}{n}\right) &\geq \mathbb{P}\left(\hat{p}_n(j) \geq \frac{2}{n}\right) \\
&\geq \binom{n}{2} p(j)^2(1-p(j))^{n-2} \\
&\geq \frac{n^2}{4}\frac{1}{(3n)^2}\left(1 - \frac{1}{n}\right)^n \geq \frac{1}{144}.
\end{aligned}$$

In particular, this shows that $\Delta_n^+(\boldsymbol{p}) \geq \frac{1}{144n} \geq \frac{\varepsilon_i}{288}$. This shows that the lower bound is directly achieved whenever there exists such an index j. Otherwise, the previous cases in Eq (3) and Eq (4) showed that for all $j \leq i$,

$$\mathbb{P}\left(\hat{p}_n(j) \geq p(j) + \frac{\varepsilon_i}{2C}\right) \geq \frac{c_0^2}{2i}.$$

Then,

$$\mathbb{P}\left(\sup_j [\hat{p}_n(j) - p(j)] \geq \frac{\varepsilon_i}{2C}\right) \geq 1 - \left(1 - \frac{c_0^2}{2i}\right)^i \geq 1 - e^{-c_0^2/2} > 0.$$

In particular, this shows that

$$\Delta_n^+(\boldsymbol{p}) \geq \frac{1 - e^{-c_0^2/2}}{2C} \varepsilon_i.$$

This gives the desired lower bound $\Delta_n^+(\boldsymbol{p}) \geq c \cdot \sup_{i \geq 1} \varepsilon_{i,p(i)}(n)$, for some universal constant $c > 0$.

We now turn to the upper bound. Here, we show that the probability that the deviation at position i exceeds $C\varepsilon$ by a factor of k is at most $(2i)^{-k}$. Thus, decaying very quickly in both k and i. We union bound this probability over the coordinates and sum up the tails (over k) to show $\Delta_n^+(\boldsymbol{p}) \lesssim \varepsilon$.

For convenience, define $\varepsilon = \sup_{j \geq 1} \varepsilon_{j,p(j)}(n)$, and let $\tilde{\varepsilon} = (\varepsilon \wedge \frac{1}{4}) \vee \frac{1}{n} \geq \frac{\varepsilon}{4}$. As a result, for any $p \in (0, \frac{1}{2}]$, one has $p + \tilde{\varepsilon} \leq \frac{1+p}{2}$. Fix $i \geq 1$. We can then apply Lemma 11 since $\tilde{\varepsilon} \geq \frac{1}{n}$, and use the continuity of the KL-divergence to obtain

$$\mathbb{P}(\hat{p}_n(i) - p(i) > \tilde{\varepsilon}) \geq c_0 e^{-CnD(p(i)+\tilde{\varepsilon} \| p(i))}.$$

On the other hand,

$$\mathbb{P}(\hat{p}_n(i) - p(i) > \tilde{\varepsilon}) \leq \mathbb{P}(\hat{p}_n(i) - p(i) > \varepsilon_i) \leq \frac{c_0}{2i}.$$

Combining the two equations gives

$$D(p(i) + \tilde{\varepsilon} \parallel p(i)) \geq \frac{\ln(2i)}{nC}.$$

Because the KL-divergence is convex in the first argument, for any $k \geq 1$ (with $C \geq 1$), we have

$$D(p(i) + kC\tilde{\varepsilon} \parallel p(i)) \geq kCD(p(i) + \tilde{\varepsilon} \parallel p(i)) \geq \frac{k \ln(2i)}{n}.$$

Now using the Chernoff bound from Lemma 10,

$$\mathbb{P}\left(\hat{p}_n(i) \geq p(i) + kC\tilde{\varepsilon}\right) \leq e^{-nD(p(i)+kC\tilde{\varepsilon} \| p(i))} \leq \frac{1}{(2i)^k}.$$

Using the union-bound yields for any $k \geq 2$,

$$\mathbb{P}\left(\sup_{i \geq 1} [\hat{p}_n(i) - p(i)] \geq kC\tilde{\varepsilon}\right) \leq \frac{1}{2^k} \sum_{i \geq 1} \frac{1}{i^k} \leq \frac{1}{2^{k-1}}. \tag{5}$$

In particular, we obtain

$$\Delta_n^+(\boldsymbol{p}) = \mathbb{E}\left[\sup_{i \geq 1} [\hat{p}_n(i) - p(i)]\right] \leq 2C\tilde{\varepsilon} + C\tilde{\varepsilon} \sum_{k \geq 2} \mathbb{P}\left(\sup_{i \geq 1} [\hat{p}_n(i) - p(i)] \geq kC\tilde{\varepsilon}\right) \leq 3C\tilde{\varepsilon}.$$

This already gives the desired upper bound whenever say $\varepsilon \geq \frac{1}{4n}$, since this implies $\tilde{\varepsilon} \leq 4\varepsilon$. We now consider the case when $\varepsilon < \frac{1}{4n}$. As discussed above, there exists $i \geq 1$ such that $p(i) + \varepsilon \geq p(i) + \varepsilon_i \geq \frac{1}{n}$. In particular, $p(i) \geq \frac{3}{4n}$. Now note that for $n \geq 2$,

$$\mathbb{P}\left(\hat{p}_n(i) \geq p(i) + \frac{1}{n}\right) = \mathbb{P}\left(\hat{p}_n(i) \geq \frac{2}{n}\right) = 1 - \left(1 - \frac{1}{n}\right)^n - \left(1 - \frac{1}{n}\right)^{n-1} \geq 1 - \frac{1}{e} - \frac{1}{2} > \frac{c_0}{2i}.$$

As a result, we should have $\varepsilon_i \geq \frac{1}{n}$, which is contradictory. This shows that the upper bound holds in all considered cases, which ends the proof of the claim for product distributions.

The upper bound directly holds for general distributions on $\{0,1\}^{\mathbb{N}}$ because it only used the union bound to analyze the effect between coordinates. ■

The next step is to characterize the quantities $\varepsilon_{J,q}(n)$. Before doing so, we state some simple bounds on the KL-divergence.

Lemma 16 *Let $0 \leq q, \varepsilon \leq \frac{1}{4}$ and suppose $\varepsilon \geq 8q$. Recall that $h(u) = (1+u)\ln(1+u) - u$. Then,*

$$\frac{\varepsilon}{2} \ln \frac{\varepsilon}{q} \leq q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right) \leq D(q + \varepsilon \parallel q) \leq 2\varepsilon \ln \frac{\varepsilon}{q}.$$

Also, for any $0 \leq q, \varepsilon \leq 1$ with $q + \varepsilon \leq \frac{1}{2}$,

$$\frac{\varepsilon^2}{2(q+\varepsilon)} \leq q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right) \leq D(q + \varepsilon \parallel q) \leq \frac{\varepsilon^2}{q}.$$

Proof First we show that $q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right) \leq D(q + \varepsilon \parallel q)$:

$$\begin{aligned} D(q + \varepsilon \parallel q) &= (q+\varepsilon)\ln\left(\frac{q+\varepsilon}{q}\right) + (1-q-\varepsilon)\ln\left(\frac{1-q-\varepsilon}{1-q}\right) \\ &= q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right) + \varepsilon + (1-q-\varepsilon)\ln\left(\frac{1-q-\varepsilon}{1-q}\right) \\ &\geq q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right), \end{aligned}$$

because $\varepsilon + (1-q-\varepsilon)\ln\left(\frac{1-q-\varepsilon}{1-q}\right) = 0$ for $\varepsilon = 0$ and is increasing in ε since $\frac{\partial}{\partial \varepsilon}\left(\varepsilon + (1-q-\varepsilon)\ln\left(\frac{1-q-\varepsilon}{1-q}\right)\right) = -\ln\left(\frac{1-q-\varepsilon}{1-q}\right) \geq 0$.

We have for $\varepsilon \geq 8q$

$$q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right) \geq \varepsilon \ln \frac{\varepsilon}{q} - \varepsilon \geq \frac{\varepsilon}{2}\ln\frac{\varepsilon}{q} + \varepsilon\left(\frac{1}{2}\ln\frac{\varepsilon}{q} - 1\right) \geq \frac{\varepsilon}{2}\ln\frac{\varepsilon}{q},$$

since $\ln \frac{\varepsilon}{q} \geq \ln 8 \geq 2$. On the other hand,

$$D(q + \varepsilon \parallel q) \leq (q+\varepsilon)\ln\frac{q+\varepsilon}{q} \leq \frac{9}{8}\varepsilon \ln \frac{2\varepsilon}{q} \leq 2\varepsilon \ln \frac{\varepsilon}{q}.$$

We now turn to the second bound when $\varepsilon \leq q$. Letting $f(\varepsilon) = D(q+\varepsilon \parallel q)$ and $g(\varepsilon) = qh(\frac{\varepsilon}{q}))$, we have $f''(\varepsilon) = \frac{1}{q+\varepsilon} + \frac{1}{1-q-\varepsilon}$ and $g''(\varepsilon) = \frac{1}{q+\varepsilon})$. As a result, for any $x \in [0, \varepsilon]$,

$$\frac{1}{q+\varepsilon} \leq g''(x) \leq f''(x) \leq \frac{2}{q+x} \leq \frac{2}{q},$$

An application of Taylor's expansion theorem ends the proof. ■

Lemma 17 *Let* $0 \leq \varepsilon, q \leq \frac{1}{2}$. *Then,*

$$q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right) \geq D\left(q + \frac{\varepsilon}{2} \parallel q\right).$$

Proof Let $f(\varepsilon) = q\,\mathrm{h}\left(\frac{\varepsilon}{q}\right) - D\left(q + \frac{\varepsilon}{2} \parallel q\right)$. Then, $f(0) = 0$ and for any $0 \leq \varepsilon \leq \frac{1}{2}$

$$\begin{aligned} f'(\varepsilon) &\geq \frac{1}{2}\ln\left(\frac{q+\varepsilon}{q}\right) + \frac{1}{2}\ln\left(\frac{1-q-\varepsilon/2}{1-q}\right) \\ &\geq \frac{1}{2}\ln\left(\frac{q(1-q)+\varepsilon(1-3q/2-\varepsilon/2)}{q(1-q)}\right) \geq 0. \end{aligned}$$

Hence, for any $0 \leq \varepsilon \leq \frac{1}{2}$, we have $f(\varepsilon) \geq 0$. ■

We now present bounds on $\varepsilon_{J,q}(n)$. To do so, we start by showing upper bounds using the function $\phi_{J,q}(n)$.

Proposition 18 *There exists a universal constant* $C_1 \geq 1$ *such that for all* $n, J \geq 1$ *and* $q \in (0, 1/2]$,

$$\varepsilon_{J,q}(n) \leq C_1 \cdot \phi_{J,q}(n).$$

Proof The proof relies on the Chernoff bound from Lemma 10. In the rest of this proof, we let $Y \sim \mathcal{B}(n,p)$, $\hat{p}_n = \frac{Y}{n}$ and $\varepsilon := \varepsilon_{J,q}(n)$. We have $\frac{c_0}{2J} < \mathbb{P}(\hat{p}_n \geq q+\varepsilon) \leq e^{-nD(q+\varepsilon\|q)}$. As a result, this shows

$$D(q+\varepsilon \parallel q) \leq \frac{\ln \frac{2J}{c_0}}{n}.$$

The upper bound given in the first regime $n \leq \frac{\ln(J+1)}{\ln\frac{1}{q}}$ is trivial. We then turn to the second regime. For convenience, we will denote $\phi(n) := \phi_{J,q}(n)$.

Regime $\frac{\ln(J+1)}{\ln\frac{1}{q}} \leq n \leq \frac{\ln(J+1)}{eq}$. We first observe that the function $n \ln \frac{\ln(J+1)}{nq}$ is non-decreasing in that regime. As a result, we always have $\phi(n) \geq \phi\left(\frac{\ln(J+1)}{eq}\right) = eq$. Next, the upper bounds are immediate if $q \geq \frac{1}{4}$ since using a constant $C_1 \geq 4$ would yield a trivial upper bound 1. We therefore suppose that $q \leq \frac{1}{4}$. Similarly, without loss of generality, suppose $\phi(n) \leq \frac{1}{40}$. As a result, by Lemma 16, for any constant $\alpha \geq 1$,

$$D(q + 3\alpha\phi(n) \parallel q) \geq \frac{3}{2}\alpha\phi(n)\ln\frac{3\alpha\phi(n)}{q}.$$

Now if $x \geq 1$ is the solution to the equation $x \ln x = 2\frac{\ln \frac{2J}{c_0}}{nq} := z \geq 2$, one has precisely

$$x \asymp \frac{z}{\ln z} \asymp \frac{\ln(J+1)}{nq \ln \frac{\ln(J+1)}{nq}}.$$

As a result, there exists a constant $\alpha \geq 1$ sufficiently large such that either $3\alpha\phi(n) \geq 1/4$ (in which case the bound for this regime is immediate for sufficiently large C_1), or

$$D(q + 3\alpha\phi(n) \parallel q) \geq \frac{\ln \frac{2J}{c_0}}{n}.$$

This implies $\varepsilon \leq 3\alpha\phi(n)$.

Regime $n \geq \frac{\ln(J+1)}{eq}$. In this regime, we have $\phi(n) \leq q\sqrt{e} \leq 2q$. Using the second estimate from Lemma 16, we have for any constant $\gamma \geq 1$,

$$D(q + \gamma\phi(n) \parallel q) \geq \frac{\gamma^2 \ln(J+1)}{2(1+2\gamma)n}.$$

As a result, there exists a universal constant $\gamma \geq 1$ such that $D(q + \gamma\phi(n) \parallel q) \geq 2\frac{\ln \frac{2J}{c_0}}{n}$, which implies $\varepsilon \leq \gamma\phi(n)$. This ends the proof of the proposition. ■

We next turn to lower bounds.

Proposition 19 *There is a universal constants $c_2 > 0$ such that for all $J, n \geq 1$ and $q \in (0, 1/2]$ satisfying $1 - (1-q)^n > \frac{c_0}{2J}$ (e.g. for $q \geq \frac{c_0}{nJ}$), one has*

$$\varepsilon_{J,q}(n) \geq \left(\frac{1}{n} - q\right) \vee c_2 \cdot \phi_{J,q}(n).$$

On the other hand, if $1 - (1-q)^n \leq \frac{c_0}{2J}$, we have $\varepsilon_{J,q}(n) = -q$.

Proof As in the previous proof, we let $\hat{p}_n = \frac{Y}{n}$ where $Y \sim \mathcal{B}(n,p)$. We compute $\mathbb{P}(\hat{p}_n \geq 1/n) = 1 - (1-q)^n$. As a result, if $1 - (1-q)^n > \frac{c_0}{2J}$, we have $q + \varepsilon_{J,q}(n) \geq \frac{1}{n}$ and otherwise, $q + \varepsilon_{J,q}(n) = 0$. We now prove that $q \geq \frac{c_0}{nJ}$ suffices to obtain $1 - (1-q)^n > \frac{c_0}{2J}$. Note that $1 - (1-q)^n \geq 1 - e^{-qn}$. If $q \geq \frac{\ln 2}{n}$, we have $\mathbb{P}(\hat{p}_n \geq 1/n) \geq \frac{1}{2} > \frac{c_0}{2J}$. Otherwise, since $qn \leq \ln 2$ and the exponential function is convex, we have

$$\mathbb{P}(\hat{p}_n \geq 1/n) \geq \frac{nq}{2\ln 2} > \frac{c_0}{2J}.$$

We assume from now on that $1 - (1-q)^n > \frac{c_0}{2J}$. Let $x(n)$ be the solution to the equation

$$D(q + x(n) \parallel q) = \frac{\ln \frac{3J}{2}}{Cn}.$$

If $q + x(n) \geq \frac{1}{n}$, Lemma 11 shows that

$$\mathbb{P}(\hat{p}_n \geq q + x(n)) \geq c_0 e^{-CnD(q+x(n)\|q)} \geq \frac{2c_0}{3J} > \frac{c_0}{2J}.$$

As a result, if $q + x(n) \geq \frac{1}{n}$, we obtain $\varepsilon_{J,q}(n) \geq x(n)$. Thus, in both cases, we obtain

$$\varepsilon_{J,q}(n) \geq \left(\frac{1}{n} - q\right) \vee x(n).$$

It remains to compute an estimate of $x(n)$. Using Lemma 16, if $x(n) \geq 8q$, we have $D(q + x(n) \parallel q) \asymp x(n) \ln \frac{x(n)}{q}$, so that similarly as in the proof of Proposition 18, we have with $z = \frac{\ln \frac{3J}{2}}{Cnq}$,

$$x(n) \asymp q \frac{z}{\ln(2+z)} \asymp \frac{\ln(J+1)}{n \ln\left(2 + \frac{\ln(J+1)}{nq}\right)}.$$

On the other hand, if $x(n) \leq 10q$, the second bounds of Lemma 16 show that $D(q + x(n) \parallel q) \asymp \frac{x(n)^2}{q}$. As a result, this yields

$$x(n) \asymp \sqrt{\frac{q \ln \frac{3J}{2}}{n}} \asymp \sqrt{\frac{q \ln(J+1)}{n}}.$$

The cutoff for $x(n)$ corresponds to $n \asymp \frac{\ln(J+1)}{eq}$, and the two estimates of $x(n)$ match in this complete regime (if $a\frac{\ln(J+1)}{eq} \leq n \leq b\frac{\ln(J+1)}{eq}$ for some universal constants $0 < a \leq b$) up to constants. Recalling that $x(n) \leq 1$, we obtained exactly $x(n) \asymp \phi(n)$. This proves that for some universal constant $c_2 > 0$, one has

$$\varepsilon_{J,q}(n) \geq \left(\frac{1}{n} - q\right) \vee c_2 \cdot \phi(n),$$

which ends the proof of the proof of the proposition. ■

Proof of Proposition 13 Propositions 18 and 19 exactly prove Proposition 13. ■

We now combine the two results when possible, to give estimates on $\Delta_n^+(\boldsymbol{p})$.

Proposition 20 *For any $p \in [0, \frac{1}{2}]_{\downarrow 0}^{\mathbb{N}}$ and $n \geq 1$ such that there exists $j \geq 1$ with $p(j) \geq \frac{c}{nj}$, we have*

$$\Delta_n^+(\boldsymbol{p}) \asymp \sup_{i \geq 1} \phi_{i,p(i)}(n).$$

Proof For any $j \geq 1$ such that $p(j) \geq \frac{c_0}{nj}$, using Proposition 13, we have $\varepsilon_{j,p(j)}(n) \asymp \phi_{j,p(j)}(n) \geq 0$. In fact, whenever $\varepsilon_{j,p(j)}(n) \geq 0$, these propositions imply $\varepsilon_{j,p(j)}(n) \asymp \phi_{j,p(j)}(n)$. Then, Proposition 12 implies that $\Delta_n^+(\boldsymbol{p}) \asymp \sup_{i \geq 1} \varepsilon_{i,p(i)}(n)$. For convenience, let $\varepsilon = \sup_{i \geq 1} \varepsilon_{i,p(i)}(n) \geq \frac{c_3}{n}$. In order to prove the theorem, given Proposition 13, it remains to prove that if $\varepsilon_{i,p(i)}(n) < 0$ for some $i \geq 0$, we have $\phi_{i,p(i)}(n) \lesssim \phi_{j,p(j)}(n)$ ($j \geq 1$ is such that $p(j) \geq \frac{c_0}{nj}$).

First, necessarily $p(i) < \frac{c_0}{ni}$. As a result, $\frac{\ln(j+1)}{\ln \frac{1}{p(i)}} \leq 1$ and hence the first regime for $\phi_{i,p(i)}$ is not present. Further,

$$\frac{\ln(i+1)}{ep(i)} \geq \frac{i \ln(i+1)}{ec_0} n \geq \frac{\ln 2}{ec_0} n.$$

This proves that either n falls the second regime for $\phi_{i,p(i)}(n)$, i.e., $\frac{\ln(i+1)}{\ln\frac{1}{p(i)}} \leq n \leq \frac{\ln(i+1)}{ep(i)}$, or $n \asymp \frac{\ln(i+1)}{ep(i)}$. In both cases,

$$\phi_{i,p(i)}(n) \asymp \frac{\ln(i+1)}{n\ln\left(2+\frac{\ln(i+1)}{np(i)}\right)} \leq \frac{\ln(i+1)}{n\ln\left(2+\frac{i\ln(i+1)}{c_0}\right)} \lesssim \frac{1}{n}.$$

As a result, there exists a universal constant $C_3 > 0$ such that $\phi_{i,p(i)}(n) \leq \frac{C_3}{n}$. Now recall that $p(j) \geq \frac{c_0}{nj}$. We first consider the case when $p(j) \leq \frac{\ln(j+1)}{en}$. In this case, $\phi_{j,p(j)}(n)$ lies in one of the two regimes. In the first regime, we have directly $\phi_{j,p(j)}(n) \asymp 1 \gtrsim \frac{1}{n}$. In the second case, we have

$$\phi_{j,p(j)}(n) = \frac{\ln(j+1)}{n\ln\frac{\ln(j+1)}{np(j)}} \geq \frac{\ln(j+1)}{n\ln\frac{j\ln(j+1)}{c_0}} \gtrsim \frac{1}{n}.$$

We now consider the case when $p(j) \geq \frac{\ln(j+1)}{en}$. In this case, $\phi_{j,p(j)}(n)$ lies in the third regime which yields

$$\phi_{j,p(j)}(n) = \sqrt{\frac{p(j)\ln(j+1)}{n}} \geq \frac{\ln(j+1)}{\sqrt{e}\cdot n} \gtrsim \frac{1}{n}.$$

As a result, there is a constant $c_3 > 0$ such that in all cases $\phi_{j,p(j)}(n) \geq \frac{c_3}{n}$. Putting everything together yields

$$\sup_{i\geq 1}\phi_{i,p(i)}(n) \asymp \sup_{j\geq 1,\varepsilon_{j,p(j)}(n)\geq 0}\phi_{i,p(i)}(n) \asymp \sup_{i\geq 1}\varepsilon_{j,p(j)}(n) \asymp \Delta_n^+(\boldsymbol{p}).$$

This ends the proof of the proposition. ■

It remains to consider the Poissonian case in which one has $p(j) \leq \frac{c}{nj}$ for all $i \geq 1$. Recall that we have $c_0 < \frac{1}{2}$.

Proposition 21 *For any $\boldsymbol{p} \in [0, \frac{1}{2}]_{\downarrow 0}^{\mathbb{N}}$ and $n \geq 1$ such that for all $j \geq 1$, one has $p(j) \leq \frac{1}{2nj}$, then*

$$\Delta_n^+(\boldsymbol{p}) \asymp \frac{1}{n} \wedge \sum_{j\geq 1} p(j).$$

Proof We first give some simple bounds on binomial tails for $q \leq \frac{1}{2n}$. We write $\hat{q}_n = \frac{Y}{n}$ for $Y \sim \mathcal{B}(n,q)$. For any $k \geq 1$,

$$\mathbb{P}\left(\hat{q}_n \geq \frac{k}{n}\right) = \sum_{l=k}^{n}\binom{n}{l}q^l(1-q)^{n-l} \leq \sum_{l=k}^{n}\frac{(nq)^l}{l!} \leq 2\frac{(nq)^k}{k!}.$$

Now let $p \in [0, \frac{1}{2}]_{\downarrow 0}^{\mathbb{N}}$ such that for all $j \geq 1$, $p(j) \leq \frac{1}{2nj}$. We define $U(\boldsymbol{p}) = \sup_{j\geq 1} njp(j) \leq \frac{1}{2}$. Letting $\hat{P}_n := \sup_{j\geq 1}\hat{p}_n(j)$, for any $k \geq 2$, the union bound implies

$$\mathbb{P}\left(\hat{P}_n \geq \frac{k}{n}\right) \leq \frac{2U(\boldsymbol{p})^k}{k!}\sum_{j\geq 1}\frac{1}{j^k} \leq \frac{\pi^2}{3}\frac{U(\boldsymbol{p})^k}{k!}.$$

Hence,

$$\mathbb{E}\left[\hat{P}_n \mathbb{1}_{\hat{P}_n \geq \frac{2}{n}}\right] \leq \sum_{k\geq 2} \frac{k}{n}\mathbb{P}\left(\hat{P}_n \geq \frac{k}{n}\right) \leq \frac{2\pi^2}{3}\frac{U(\boldsymbol{p})^2}{n} \leq \frac{\pi^2 U(\boldsymbol{p})}{3n}.$$

Now let $V(\boldsymbol{p}) = \sum_{j\geq 1} np(j)$. Note that for any $j \geq 1$, $V(\boldsymbol{p}) \geq \sum_{i\leq j} np(j) \geq njp(j)$, so that $U(\boldsymbol{p}) \leq V(\boldsymbol{p}) \wedge 1$. By linearity of the expectation, one has

$$\mathbb{E}\left[\sum_{j\geq 1} n\hat{p}_n(j)\right] = V(\boldsymbol{p}).$$

In particular, since this sum takes integer values and is nonzero whenever $\hat{P}_n \geq \frac{1}{n}$, we obtain $\mathbb{P}(\hat{P}_n \geq \frac{1}{n}) \leq V(\boldsymbol{p}) \wedge 1$. We now show that $\mathbb{P}(\hat{P}_n \geq \frac{1}{n}) \gtrsim V(\boldsymbol{p}) \wedge 1$. We have

$$\mathbb{P}\left(\hat{P}_n \geq \frac{1}{n}\right) = 1 - \prod_{j\geq 1}(1-p(j))^n \geq 1 - e^{-V(\boldsymbol{p})} \geq c_4 V(\boldsymbol{p}) \wedge 1,$$

for some universal constant $c_4 > 0$. Recall that for all $j \geq 1$, one has $p(j) \leq \frac{1}{2n}$. Therefore, whenever $\hat{P}_n \geq \frac{1}{n}$, we have $\sup_{j\geq 1} \hat{p}_n(j) - p(j) \geq \frac{1}{2n}$. The previous bound then shows that

$$\Delta_n^+(\boldsymbol{p}) \geq \frac{c_4}{2}\frac{V(\boldsymbol{p}) \wedge 1}{n}.$$

On the other hand,

$$\begin{aligned}
\Delta_n^+(\boldsymbol{p}) \leq \mathbb{E}[\hat{P}_n] &\leq \frac{1}{n}\mathbb{P}\left(\hat{P}_n \geq \frac{1}{n}\right) + \mathbb{E}\left[\hat{P}_n \mathbb{1}_{\hat{P}_n \geq \frac{2}{n}}\right] \\
&\leq \frac{V(\boldsymbol{p}) \wedge 1}{n} + \frac{\pi^2 U(\boldsymbol{p})}{3n} \\
&\leq 5\frac{V(\boldsymbol{p}) \wedge 1}{n}.
\end{aligned}$$

where in the last inequality we used $U(\boldsymbol{p}) \leq V(\boldsymbol{p}) \wedge 1$. This ends the proof of the proposition. ■

Using the previous results, we are now ready to prove the complete behavior of $\Delta_n(\boldsymbol{p})$.

Proof of Theorem 1 Propositions 20 and 21 provide the complete behavior of $\Delta_n^+(\boldsymbol{p})$. It remains to show that this is the leading term in the decomposition $\Delta_n(\boldsymbol{p}) \asymp \Delta_n^+(\boldsymbol{p}) + \Delta_n^-(\boldsymbol{p})$. We first consider the case when $p(j) \leq \frac{1}{2nj}$ for all $j \geq 1$. In that case, we have directly

$$\Delta_n^-(\boldsymbol{p}) \leq \sup_{j\geq 1} p(j) \leq \frac{1}{n} \wedge \sum_{j\geq 1} p(j).$$

Now suppose that $p(j) \geq \frac{1}{2nj}$ for $j \geq 1$. By construction of $\phi_{J,q}(n)$, one has for all $J \geq 1$ and $q \in (0, \frac{1}{2}]$,

$$\phi_{J,q}(n) \geq 1 \wedge \sqrt{\frac{q\ln(J+1)}{n}},$$

i.e. intuitively the second regime is larger than the third. As a result,

$$\sup_{j\geq 1}\phi_{j,p(j)}(n) \geq 1 \wedge \sup_{j\geq 1}\sqrt{\frac{p(j)\ln(j+1)}{n}} = 1 \wedge \sqrt{\frac{S(\boldsymbol{p})}{n}}.$$

Next, we clearly have $\Delta_n^-(\boldsymbol{p}) \leq 1$. Also, in the proof of Cohen and Kontorovich (2022, Theorem 3), the authors show that

$$\Delta_n^-(\boldsymbol{p}) \leq \sqrt{\frac{S(\boldsymbol{p})}{n}}.$$

Hence, we finally obtain

$$\Delta_n^+(\boldsymbol{p}) \asymp \sup_{j\geq 1}\phi_{j,p(j)}(n) \geq 1 \wedge \sqrt{\frac{S(\boldsymbol{p})}{n}} \geq \Delta_n^-(\boldsymbol{p}).$$

This ends the proof that $\Delta_n(\boldsymbol{p}) \asymp \Delta_n^+(\boldsymbol{p})$, which implies the desired result. ■

Appendix B. Proof of the localized Dvoretzky-Kiefer-Wolfowitz results

We first prove our local DKW result in Theorem 5. Before considering the case of general distributions and intervals, we focus on the simpler case of the uniform distribution and consider intervals of the form $[q/2, q]$.

Lemma 22 *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{U}([0,1])$ and F_m be the empirical CDF. Let $q \in (0, \frac{1}{2}]$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\sup_{x\in[\frac{q}{2},q]}|F_n(x) - F(x)| > t\sqrt{\frac{q}{n}}\right) \leq c_1 e^{-c_2\min(t^2, t\sqrt{nq})},$$

for some universal constants $c_1, c_2 > 0$.

Proof For the proof, we apply Bernstein inequalities to the number of points falling in intervals within $[\frac{q}{2}, q]$. We first treat the simple case when $t \geq \sqrt{nq}$. In that case, Bernstein's inequality shows that

$$\mathbb{P}\left(F_n(q) - q \geq t\sqrt{\frac{q}{n}}\right) \leq \exp\left(-\frac{\frac{1}{2}t^2 nq}{nq(1-q) + \frac{t\sqrt{nq}}{3}}\right) \leq e^{-t\sqrt{nq}/4}.$$

Suppose that the complementary event is met, then for any $x \in [0, q]$, we have $0 \leq F_n(x) \leq F_n(q) \leq q + t\sqrt{\frac{q}{n}} \leq 2t\sqrt{\frac{q}{n}}$. In particular, $|F_n(x) - x| \leq 2t\sqrt{\frac{q}{n}}$. Hence,

$$\mathbb{P}\left(\sup_{x\leq q}|F_n(x) - x| \geq 2t\sqrt{\frac{q}{n}}\right) \leq e^{-t\sqrt{nq}/4}.$$

This shows that for any $t \geq 2\sqrt{nq}$,

$$\mathbb{P}\left(\sup_{x\leq q}|F_n(x) - x| \geq t\sqrt{\frac{q}{n}}\right) \leq e^{-t\sqrt{nq}/8}.$$

In the rest of the proof, we suppose that $t \le \frac{3}{2}\sqrt{nq}$ which implies $3n\frac{q}{2} \ge t\sqrt{nq}$. We recall that if $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{B}(r)$ are independent Bernoulli variables with $r \in (0, \frac{1}{2}]$, the Bernstein's inequality yields for $\delta \le 3nr$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i - nr\right| \ge \delta\right) \le 2\exp\left(-\frac{\frac{1}{2}\delta^2}{nr(1-r)+\frac{\delta}{3}}\right) \le 2\exp\left(-\frac{\delta^2}{4nr}\right).$$

Now consider any $u \ge 1$ such that $\frac{3nq}{2^u} \ge t\sqrt{nq}$, and any $v \in \{0, \ldots, 2^{u-1}-1\}$. We apply the previous inequality to the points $X_1, \ldots, X_n$ falling in the interval $I_{u,v} := (\frac{q}{2} + q\frac{v}{2^u}, \frac{q}{2} + q\frac{v+1}{2^u}]$. We obtain

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i \in I_{u,v}) - \frac{q}{2^u}\right| \ge \frac{t}{2^{u/3}}\sqrt{\frac{q}{n}}\right) \le 2e^{-t^2 2^{u/3-2}}.$$

Last, using the same inequality, given that $3n\frac{q}{2} \ge t\sqrt{nq}$, we have that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(Y_i \le \frac{q}{2}\right) - \frac{q}{2}\right| \ge t\sqrt{\frac{q}{n}}\right) \le 2e^{-t^2/2}.$$

We denote by E_t the intersection of the complementary events described above. By the union bound, we have

$$1 - \mathbb{P}(E_t) \le 2e^{-t^2/2} + \sum_{u \ge 1} 2^{u-1} \cdot 2e^{-t^2 2^{u/3-2}} \le c_1 e^{-c_2 t^2}$$

for some universal constants $c_1, c_2 > 0$. We now suppose that this event is met and aim to prove an upper bound for $|F_n(x) - F(x)|$ for an arbitrary $x \in [\frac{q}{2}, q)$. To do so, we first focus on the points of the form $x_v = \frac{q}{2} + \frac{v}{2^{u_0}}$ where $u_0 \ge 1$ is the largest integer for which $\frac{3nq}{2^{u_0}} \ge t\sqrt{nq}$. In particular, we have $\frac{q}{2^{u_0}} \le \frac{2}{3}t\sqrt{\frac{q}{n}}$. We decompose v in binary encoding via $v = \sum_{u \le u_0} a_u 2^{u_0-u}$ where $a_u \in \{0,1\}$ for $u \in [u_0]$. Writing $v_u = \sum_{u' \le u} a_{u'} 2^{u-u'}$, we can write

$$nF_n(x_v) = \sum_{i=1}^{n} \mathbb{1}\left(Y_i \le x_v\right) = \sum_{i=1}^{n} \mathbb{1}\left(Y_i \le \frac{q}{2}\right) + \sum_{u=1}^{u_0}\sum_{i=1}^{n} \mathbb{1}(Y_i \in I_{u,v_u}).$$

As a result, on E, we have for any $v \in \{0, \ldots, 2^{u_0}\}$,

$$|F_n(x_v) - x_v| \le t\sqrt{\frac{q}{n}} + \sum_{u=1}^{u_0} a_u \frac{t}{2^{u/3}}\sqrt{\frac{q}{n}} \le \frac{t}{1-2^{-1/3}}\sqrt{\frac{q}{n}}.$$

Last, let $x \in (\frac{q}{2}, q]$. There exists $v \in \{0, \ldots, 2^{u_0}-1\}$ such that $\frac{v}{2^{u_0}} < x \le \frac{v+1}{2^{u_0}}$. We note that

$$\begin{aligned}
|F_n(x) - x| &\le \max\left(\left|F_n\left(\frac{v}{2^{u_0}}\right) - \frac{v}{2^{u_0}}\right|, \left|F_n\left(\frac{v+1}{2^{u_0}}\right) - \frac{v+1}{2^{u_0}}\right|\right) + \frac{1}{2^{u_0}} \\
&\le \left(\frac{1}{1-2^{-1/3}} + \frac{2}{3}\right) t\sqrt{\frac{q}{n}} \\
&\le 6t\sqrt{\frac{q}{n}}.
\end{aligned}$$

Hence, on E, we showed that

$$\sup_{x\in[\frac{q}{2},q]} |F_n(x) - x| \le 6t\sqrt{\frac{q}{n}}.$$

Hence, we showed that for any $t \le \frac{9}{2}\sqrt{nq}$, one has

$$\mathbb{P}\left(\sup_{x\in[\frac{q}{2},q]} |F_n(x) - x| > t\sqrt{\frac{q}{n}}\right) \le c_1 e^{-c_2 t^2},$$

for some universal constants $c_1, c_2 > 0$. ∎

We are now ready to prove the local DKW bound for intervals of the form $(-\infty, x_0]$.

Proof of Corollary 14 First, note that if $F(x_0) \ge \frac{1}{2}$, then we can use the classical DKW Theorem 4 to obtain the desired bound. We will therefore suppose without loss of generality that $F(x_0) \le \frac{1}{2}$. We first prove the result for the uniform distribution. Fix $q \in [0, \frac{1}{2}]$. For convenience, let $\delta = t\sqrt{\frac{q}{n}}$. We first suppose that $q \le \delta$. Then, Lemma 22 implies in particular that

$$\mathbb{P}\left(|F_n(q) - q| \le \delta\right) \le c_1 e^{-c_2 \min(t^2, t\sqrt{nq})} = c_1 e^{-c_2 t\sqrt{nq}}.$$

Note that on the event $|F_n(q) - q| \le \delta$, we have in particular for all $x \le q$ that $F_n(x) \le F_n(q) \le q + \delta \le 2\delta$. As a result, for all $x \le q$, $|F_n(x) - x| \le 2\delta$. This yields

$$\mathbb{P}\left(\sup_{x\le q} |F_n(x) - x| \le 2\delta\right) \le c_1 e^{-c_2 \min(t^2, t\sqrt{nq})} = c_1 e^{-c_2 t\sqrt{nq}}.$$

We now consider the case when $q \ge \delta$. Similarly as above, if $|F_n(\delta) - \delta| \le \delta$, then for any $x \le \delta$, we have $\sup_{x\le\delta} |F_n(x) - x| \le 2\delta$. As a result, we can focus on the interval $[\delta, q]$. We decompose the supremum on intervals of the form $[\frac{q}{2^{u+1}}, \frac{q}{2^u}]$ for $u \ge 0$. From the above arguments, it suffices to consider intervals $[\frac{q}{2^{u+1}}, \frac{q}{2^u}]$ for $u \le u_0$ such that $\frac{q}{2^{u_0+1}} \le t\sqrt{\frac{q}{n}} \le \frac{q}{2^{u_0}}$. We note that for $0 \le u \le u_0$, one has $2^u t \le \sqrt{nq}$, so that $\min(2^u t^2, t\sqrt{nq}) = 2^u t^2$. Hence, by Lemma 22,

$$\begin{aligned}
\mathbb{P}\left(\sup_{x\in[0,q]} |F_n(x) - x| > 2\delta\right) &\le \mathbb{P}\left(\sup_{x\in[\delta,q]} |F_n(x) - x| > \delta\right) \\
&\le \sum_{u=0}^{u_0} \mathbb{P}\left(\sup_{x\in[\frac{q}{2^{u+1}},\frac{q}{2^u}]} |F_n(x) - x| > 2^{u/2} t\sqrt{\frac{q}{2^u n}}\right) \\
&\le c_1 \sum_{u=0}^{u_0} e^{-c_2 \min(2^u t^2, t\sqrt{nq})} \\
&= c_1 \sum_{u=0}^{u_0} e^{-c_2 2^u t^2} \le c_3 e^{-c_4 t^2},
\end{aligned}$$

for some universal constants $c_3, c_4 > 0$. This shows that for some constants $c_3, c_4 > 0$, we have

$$\mathbb{P}\left(\sup_{x\in[0,q]} |F_n(x) - x| > 2t\sqrt{\frac{q}{n}}\right) \le c_3 e^{-c_4 \min(t^2, t\sqrt{nq})}.$$

Changing the constants appropriately ends the proof of the theorem for the uniform distribution. The result extends directly to general distributions via a change of variables. Consider a real-valued distribution μ with CDF F_X. If $U \sim \mathcal{U}([0,1])$ is uniform, then $X = F^{-1}(U) \sim \mu$, where we define $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Because the CDF F is right-continuous, we have in particular $F(F^{-1}(u)) \geq u$. Hence, $F(x) \geq u$ i.if $x \geq F^{-1}(u)$. Given n samples $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \mathcal{U}([0,1])$, we denote by $F_{n,U}$ their empirical CDF. Similarly, letting $X_i = F^{-1}(U_i)$ for $i \in [n]$, we denote by $F_{n,X}$ their empirical CDF. Now note that for any $x \in \mathbb{R}$,

$$F_{n,X}(x) - F(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(F^{-1}(U_i) \leq x) - F(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(U_i \leq F(x)) - F(x).$$

As a result, we have

$$\mathbb{P}\left(\sup_{x \leq x_0} |F_{n,X}(x) - F(x)| > t\sqrt{\frac{F(x_0)}{n}}\right) \leq \mathbb{P}\left(\sup_{x \in [0,F(x_0)]} |F_{n,U}(x) - x| > t\sqrt{\frac{F(x_0)}{n}}\right).$$

This ends the proof of the theorem. ■

Last, we now prove the main localized result for intervals $[x_0, x_1]$.

Proof of Theorem 5 If $F(x_0) \leq \frac{1}{2} \leq F(x_1)$, the standard DWK Theorem 4 gives the desired result. Otherwise, without loss of generality, we suppose that $F(x_1) \leq \frac{1}{2}$. Then, $F(x_1) \geq V = \max_{x \in [x_0,x_1]} F(x)(1-F(x)) \geq \frac{F(x_1)}{2}$. Hence, using Corollary 14,

$$\mathbb{P}\left(\sup_{x \in [x_0,x_1]} |F_n(x) - F(x)| > t\sqrt{\frac{V}{n}}\right) \leq \mathbb{P}\left(\sup_{x \leq x_1} |F_n(x) - F(x)| > t\sqrt{\frac{F(x_1)}{2n}}\right)$$
$$\leq c_1 e^{-\frac{c_2}{4}\min(t^2, 2t\sqrt{nF(x_1)})} \leq c_1 e^{-\frac{c_2}{4}\min(t^2, 2t\sqrt{nV})}.$$

This ends the proof. ■

Appendix C. Proofs of the results on the expected maximum deviation of distributions on $[0,1]$

Before proving the main result Corollary 6, we recall Bennett's inequality.

Lemma 23 (Bennett's inequality) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ, variance σ^2 and $0 \leq X_i \leq 1$ almost surely. Then for any $t > 0$ we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geq t\right) \leq e^{-n\sigma^2 \mathrm{h}\left(\frac{t}{\sigma^2}\right)},$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geq t\right) \leq e^{-n\mu \mathrm{h}\left(\frac{t}{\mu}\right)},$$

where $\mathrm{h}(u) = (1+u)\ln(1+u) - u$.

Proof For the first part, see (Boucheron et al., 2013, Theorem 2.9). For the second, note that

$$0 \le \mathbb{E}[X(1-X)] = \mathbb{E}[X] - \mathbb{E}[X^2] = \mu - \sigma^2 + \mu^2.$$

Hence, $\sigma^2 \le \mu - \mu^2 \le \mu$ and $f(x) = x\,\mathrm{h}\left(\frac{t}{x}\right)$ is decreasing on $x \in [0,\infty)$ since $f'(x) = \ln(1+\frac{t}{x}) - \frac{t}{x} \le 0$. Thus, the second statement is weaker than the first one. ■

We will use this inequality instead of the Chernoff bound (Lemma 10) that we used in the case of Bernoulli random variables.

Proof of Corollary 6 As a first step, we show that we can upper bound $\Delta_n^+(\mu)$ similarly as in Proposition 12, by replacing $p(i)$ with $\sigma^2(i)$.

As in the proof of Proposition 12, let $\varepsilon := \sup_{i\ge1} \varepsilon_{i,\sigma^2(i)}(n)$ and $\tilde\varepsilon = (\varepsilon \wedge \frac{1}{4}) \vee \frac{1}{n}$. The same proof shows that for any $i \ge 1$,

$$D(\sigma^2(i) + \tilde\varepsilon \parallel \sigma^2(i)) \ge \frac{\ln(2i)}{nC},$$

for some universal constant $C > 0$. The proof also shows that $\varepsilon \ge \frac{1}{4n}$, hence $\varepsilon \asymp \tilde\varepsilon$. Next, by Lemma 17, we have

$$\sigma^2(i)\,\mathrm{h}\left(\frac{2\tilde\varepsilon}{\sigma^2(i)}\right) \ge D(\sigma^2(i) + \tilde\varepsilon \parallel \sigma^2(i)) \ge \frac{\ln(2i)}{nC}$$

where in the last inequality, we used Lemma 17. We now use Bennett's inequality from Lemma 23 instead together with the convexity of the function $\sigma^2(i)\,\mathrm{h}(\frac{\cdot}{\sigma^2(i)})$, to obtain as in the proof of Proposition 12 that for any $k \ge 1$,

$$\mathbb{P}(\hat p_n(i) \ge p(i) + 2kC\varepsilon) \le e^{-\sigma^2(i)\,\mathrm{h}\left(\frac{2kC\varepsilon}{\sigma^2(i)}\right)} \le e^{-kC\sigma^2(i)\,\mathrm{h}\left(\frac{2\varepsilon}{\sigma^2(i)}\right)} \le \frac{1}{(2i)^k}.$$

The same union bound argument then shows that $\Delta_n^+(\mu) \le 3C\varepsilon \lesssim \varepsilon$.

In summary, this shows that if there exists $j \ge 1$ such that $\sigma^2(j) \ge \frac{1}{2nj}$, then from Proposition 19 one has $\varepsilon = \sup_{i\ge1} \varepsilon_{i,\sigma^2(i)}(n) \ge 0$. As a result, the proof of Proposition 20 shows that $\sup_{i\ge1} \varepsilon_{i,\sigma^2(i)}(n) \asymp \sup_{i\ge1} \phi_{i,\sigma^2(i)}(n)$, which gives

$$\Delta_n^+(\mu) \lesssim \sup_{i\ge1} \varepsilon_{i,\sigma^2(i)}(n) \asymp 1 \wedge \sup_{j\ge1}\left(\sqrt{\frac{\sigma^2(j)\ln(j+1)}{n}} \vee \frac{\ln(j+1)}{n\ln\left(2+\frac{\ln(j+1)}{n\sigma^2(j)}\right)}\right).$$

It now remains to bound $\Delta_n^-(\mu)$. This can be done in a completely symmetric manner, by considering the distribution $\tilde\mu$ of $(1-X_i)_{i\ge1}$ for $X \sim \mu$. We obtain directly

$$\Delta_n^-(\mu) = \Delta_n^+(\tilde\mu) \lesssim \sup_{i\ge1} \varepsilon_{i,\sigma^2(i)}(n),$$

where in the last inequality, we applied Eq (C) to $\tilde\mu$. Finally, we showed that if there exists $j \ge 1$ such that $\sigma^2(j) \ge \frac{1}{2nj}$, then

$$\Delta_n(\mu) \lesssim 1 \wedge \sup_{j\ge1}\left(\sqrt{\frac{\sigma^2(j)\ln(j+1)}{n}} \vee \frac{\ln(j+1)}{n\ln\left(2+\frac{\ln(j+1)}{n\sigma^2(j)}\right)}\right).$$

We now suppose that for all $j \geq 1$, one has $p(j) \leq \frac{1}{2nj}$. Fix $i \geq 1$. Since $\sigma^2(i) \leq \frac{1}{2ni}$, for any $k \geq 1$, one has $\frac{4k}{n} \geq 8\sigma^2(i)$. Then, Bennett's inequality in Lemma 23 together with a lower bound from Lemma 16 shows that

$$\mathbb{P}\left(\hat{p}_n(i) \geq p(i) + \frac{4k}{n}\right) \leq e^{-\sigma^2(i)\,\mathrm{h}\left(\frac{4k}{n\sigma^2(i)}\right)} \leq \left(\frac{n\sigma^2(i)}{4k}\right)^{\frac{4k}{n\sigma^2(i)}} \leq \frac{1}{(8ki)^{8ki}}.$$

As a result, the same computations as in the proof of Proposition 12 show that

$$\Delta_n^+(\mu) \leq \frac{12}{n}.$$

As before, the argument is symmetric, hence we obtain $\Delta_n^-(\mu) \leq \frac{12}{n}$ as well. This shows that

$$\Delta_n(\mu) \lesssim \frac{1}{n}.$$

Next, for any $i \geq 1$, note that $\mathrm{Var}(\hat{p}_n(i)) = \frac{\sigma^2(i)}{n}$. As a result, for any $c > 0$, Chebyshev's inequality yields

$$\mathbb{P}(|\hat{p}_n(i) - p(i)| \geq c) \leq \frac{\sigma^2(i)}{nc^2}.$$

Hence, by the union bound,

$$\mathbb{P}(\|\hat{p}_n - p\|_\infty \geq c) \leq \frac{1}{nc^2} \sum_{i \geq 1} \sigma^2(i).$$

Now suppose that $\sum_{i\geq 1} \sigma^2(i) < \infty$. For simplicity, let $\eta = \sqrt{\frac{\sum_{i\geq 1} \sigma^2(i)}{n}}$. Then,

$$\begin{aligned}
\Delta_n(\mu) = \mathbb{E}\|\hat{p}_n - p\|_\infty &\leq \eta + \sum_{k\geq 1} 2^k \eta \mathbb{P}\left(\|\hat{p}_n - p\|_\infty \geq 2^{k-1}\eta\right) \\
&\leq \eta + \eta \sum_{k \geq 1} \frac{\sum_{i\geq 1} \sigma^2(i)}{n 2^{k-2} \eta^2} = 5\eta.
\end{aligned}$$

This ends the proof that

$$\Delta_n(\mu) \lesssim \frac{1}{n} \wedge \sqrt{\frac{\sum_{i\geq 1} \sigma^2(i)}{n}},$$

which ends the proof of the result. ■

By Theorem 1, we know that the upper bounds from Corollary 6 are attained using a sequence of independent Bernoulli random variables—we recall that in this case, since $p(i) \in [0, 1/2]$, for $i \geq 1$, one has $\sigma^2(i) \asymp p(i)$—except in the case when $\sum_{j\geq 1} \sigma^2(j) \leq \frac{1}{2n}$.

In that case, changing the support from $\{0, 1\}$ to $\{0, \sqrt{2n \sum_{j\geq 1} \sigma^2(j)}\}$ achieves the desired upper bound. For convenience, define $\eta = \sqrt{2n \sum_{i\geq 1} \sigma^2(i)}$. We consider the sequence of independent variables $X_i = \eta Z_i$ where $(Z_i)_{i\geq 1}$ are independent Bernoulli variables with parameters

$\frac{\sigma^2(i)}{\eta^2}$. We denote by μ this distribution. We have that

$$\mathbb{P}\left(\hat{p}_n(i) \geq \frac{\eta}{n}\right) = 1 - \prod_{j\geq 1}\left(1 - \frac{\sigma^2(i)}{\eta^2}\right)^n \geq 1 - \exp\left(-\frac{n}{\eta^2}\sum_{j\geq 1}\sigma^2(i)\right) \geq 1 - e^{-1/2}.$$

As a result, since $p(i) = \eta\frac{\sigma^2(i)}{\eta^2} \leq \frac{\eta}{2n}$, we obtained that

$$\mathbb{E}\|\hat{p}_n - p\|_\infty \gtrsim \frac{\eta}{n} \asymp \sqrt{\frac{\sum_{i\geq 1}\sigma^2(i)}{n}}.$$

Appendix D. Proofs of the results on expected empirical deviations in ℓ^q norms

We first prove the convergence characterization from Proposition 7.

Proof of Proposition 7 Suppose that $\|\boldsymbol{p}\|_1 = \infty$. Further let $p(i) \to 0$ as $i \to \infty$; otherwise the result would be straightforward. Let $(X_i)_{i\geq 1}$ be a sequence of independent Bernoulli random variables such that $X_i \sim \mathcal{B}(p(i))$. Because $\sum_{j\geq 1} p(j) = \infty$, by Borel-Cantelli's lemma, almost surely, there is an infinite number of indices $i \geq 1$ for which $X_i = 1$. In particular, with full probability, there is an infinite number of indices i for which $\hat{p}_n(i) \geq \frac{1}{n}$ and thus infinitely many for which $|\hat{p}_n(i) - p(i)| \geq \frac{1}{2n}$ since $p(i) \to 0$ as $i \to \infty$. As a result, $\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q = \infty$ $(a.s.)$.

Now suppose that $\|\boldsymbol{p}\|_1 < \infty$. We note that

$$\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \leq \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_1 \asymp \sum_{j\geq 1} p(j) \wedge \sqrt{\frac{p(j)}{n}}.$$

In particular, for any $\varepsilon > 0$, there exists $i \geq 1$ such that $\sum_{j\geq i} p(j) < \varepsilon$. Then, for $n \geq \frac{1}{p(i)}$, we have

$$\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \lesssim \frac{1}{\sqrt{n}}\sum_{j<i}\sqrt{p(j)} + \sum_{j\geq i} p(j) \leq \frac{1}{\sqrt{n}}\sum_{j<i}\sqrt{p(j)} + \varepsilon.$$

Hence, $\limsup_{n\to\infty} \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \leq \varepsilon$. Because this holds for any ε, this shows that $\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \to 0$ as $n \to \infty$. ∎

We now provide bounds on the deviation $\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q$ when $\|\boldsymbol{p}\|_1 < \infty$. To do so, we first need estimates on the central moments of binomials.

Lemma 24 *Let $n \geq 1$ and $0 \leq p \leq \frac{1}{2}$. Let $Y \sim \mathcal{B}(n,p)$ be a binomial and $q \geq 1$. Then,*

$$\mathbb{E}|Y - np|^q \asymp_q \psi_q(n,p) := \begin{cases} (npq)^{q/2} & p \geq \frac{q}{2n} \\ \left(\frac{q}{\ln\frac{q}{np}}\right)^q & \frac{q}{ne^q} \leq p \leq \frac{q}{2n} \\ np & p \leq \frac{q}{ne^q} \end{cases}$$

where the $\asymp_q$ term hides factors $\Omega(c^q)$ and $\mathcal{O}(C^q)$ for universal constants $c, C > 0$.

Proof We consider the three different regimes separately. Before doing so, we introduce some notations. For convenience, we will use the extended factorials to define for any $l \in [-np, n(1-p)]$,

$$b_l := \binom{n}{np+l} p^{np+l}(1-p)^{n(1-p)-l} l^q.$$

Regime 1: $\frac{q}{2n} \leq p \leq \frac{1}{2}$**.** We aim to understand the sequence $(b_l)_l$ and start with the right tails when $l \geq 0$. Let $l \geq 2\sqrt{npq}$. We can use the convexity inequality $e^x \leq 1 + 3x$ for $x \in [0, 1]$ to obtain

$$\frac{b_{l+1}}{b_l} = \frac{n(1-p) - l}{np + l + 1} \cdot \frac{p}{1-p} \left(1 + \frac{1}{l}\right)^q \leq \frac{e^{q/l}}{1 + \frac{l}{np}} \leq \frac{e^{\sqrt{q/np}/2}}{1 + 2\sqrt{q/np}} \leq e^{-\sqrt{q/np}/6} \leq e^{-1/6}. \quad (6)$$

In the second-to-last inequality, we used the fact that $\frac{2}{3}\sqrt{\frac{q}{np}} \leq \frac{2\sqrt{2}}{3} \leq 1$. In particular, this shows that if $k_1 = \lceil np + 2\sqrt{npq} \rceil$,

$$\sum_{k=k_1}^{n} b_{k-np} \asymp b_{k_1 - np}. \quad (7)$$

On the other hand, if $1 \leq l \leq \frac{1}{6}\sqrt{npq}$

$$\frac{b_{l+1}}{b_l} \geq \left(1 - \frac{3l}{np}\right)\left(1 + \frac{q}{l}\right) \geq \left(1 - \frac{1}{2}\sqrt{\frac{q}{np}}\right)\left(1 + 6\sqrt{\frac{q}{np}}\right) \leq 1$$

In the last inequality, we used the fact that $\sqrt{\frac{q}{np}} \leq \sqrt{2}$. As a result, the maximum of b_l for $l \geq 0$ is achieved for $l \in [\frac{1}{6}\sqrt{npq}, 2\sqrt{npq}]$, and if $k_2 = \lceil np + \frac{1}{6}\sqrt{npq} \rceil$, we obtained

$$\sum_{k=\lceil np \rceil}^{k_2 - 1} b_{k-np} \leq \frac{1}{6}\sqrt{npq} \cdot b_{k_2 - np}. \quad (8)$$

We next show that up to exponential terms in q, b_l has same order within this range. Precisely, for $l \in [\frac{1}{6}\sqrt{npq}, 2\sqrt{npq}]$ and an integer $0 \leq r \leq 2\sqrt{npq}$, we have

$$\frac{b_{l+r}}{b_l} \leq \left(1 + \frac{r}{l}\right)^q \leq \left(1 + \frac{1}{8}\right)^q \asymp_q 1.$$

We now turn to the lower bound and now suppose $n \geq 100$q. We will treat the other case separately. We recall that $l + r \leq 4\sqrt{npq} \leq \frac{4n}{\sqrt{200}} \leq \frac{n}{4}$, so that $\frac{l+r-1}{n(1-p)} \leq \frac{1}{2}$. Next, by convexity, we have the inequality $1 - x \geq e^{-2x}$ for $x \in [0, 1/2]$. Hence,

$$\frac{b_{l+r}}{b_l} \geq \left(\frac{1 - \frac{l+r-1}{n(1-p)}}{1 + \frac{l+r}{np}}\right)^r \geq \exp\left[-r\left(8\sqrt{\frac{q}{n(1-p)}} + 4\sqrt{\frac{q}{np}}\right)\right] \geq e^{-2(8\sqrt{2}+4)q} \asymp_q 1.$$

Hence, this shows that $b_l \asymp_q b_{l+r}$. In particular, the two previous statements showed that

$$\sum_{k=k_2}^{k_1} b_{k-np} \asymp_q \sqrt{npq} \cdot b_{k_1 - np}.$$

Combining the previous equation together with Eq (7) and (8), we have,

$$\mathbb{E}\left[|Y - np|^q \mathbb{1}_{Y > np}\right] = \sum_{np \leq k \leq n} b_{k-np} \asymp_q \sqrt{npq} \cdot b_{k_1 - np}.$$

We then use Stirling's approximation formula to estimate the right-hand side. Noting that $1 \leq \sqrt{npq} \leq \frac{np}{10}$ since $n \geq 100q$, we have

$$b_{k_1-np} \asymp_q \frac{(\sqrt{npq})^q}{\sqrt{np}\left(\frac{k_1}{np}\right)^{k_1}\left(\frac{1-\frac{k_1}{n}}{1-p}\right)^{n-k_1}}.$$

Now writing $k_1 = np + l_1$, we have that $\frac{l_1}{np} \leq 2\sqrt{\frac{q}{np}} \leq 2$ and $\frac{k_1 l_1^2}{(np)^2} = \mathcal{O}(q)$. Then,

$$\left(\frac{k_1}{np}\right)^{k_1} = \exp\left(k_1 \ln\left(1+\frac{l_1}{np}\right)\right) = \exp\left(\frac{k_1 l_1}{np} + k_1\mathcal{O}\left(\frac{l_1^2}{(np)^2}\right)\right) = e^{l_1+\mathcal{O}(q)} \asymp_q e^{l_1}.$$

Similarly,

$$\begin{aligned}
\left(\frac{1-\frac{k_1}{n}}{1-p}\right)^{n-k_1} &= \exp\left((n-k_1)\ln\left(1-\frac{l_1}{n(1-p)}\right)\right) \\
&= \exp\left(-l_1 - \frac{l_1^2}{n(1-p)} + (n-k_1)\mathcal{O}\left(\frac{l_1^2}{n^2}\right)\right) \asymp_q e^{-l_1}.
\end{aligned}$$

As a result, combining all the previous estimates gives

$$\mathbb{E}\left[|Y-np|^q \mathbb{1}_{Y>np}\right] \asymp_q \sqrt{q}\cdot(npq)^{q/2} \asymp_q (npq)^{q/2}.$$

We now turn to the left tails. For $2\sqrt{npq} \leq l \leq np$,

$$\frac{b_{-l-1}}{b_{-l}} = \frac{1-\frac{l}{np}}{1+\frac{l+1}{n(1-p)}}\left(1+\frac{1}{l}\right)^q \leq \exp\left(\frac{q}{l}-\frac{l}{np}\right) \leq \exp\left(-\sqrt{\frac{q}{np}}\right) \leq \frac{1}{e}.$$

Thus, if $k_3 = \lfloor np - 2\sqrt{npq}\rfloor \geq 0$, we have that $\sum_{k=0}^{k_3} b_{k-np} \asymp b_{k_3-np}$. It suffices then to focus on the terms b_{-l} for $0 \leq l \leq 2\sqrt{npq}+1$. Going back to the previous displayed equation shows that the term $\binom{n}{np-l}p^{np-l}(1-p)^{n(1-p)+l}$ is decreasing with $l \geq 0$. As a result, for any $k_3 \leq k \leq \lfloor np \rfloor$, we have

$$b_{k-np} \leq \binom{n}{\lfloor np \rfloor} p^{\lfloor np \rfloor}(1-p)^{n-\lfloor np \rfloor}(2\sqrt{npq}+1)^q \asymp_q \frac{(npq)^{q/2}}{\sqrt{np}}.$$

In the last inequality, we used Stirling's approximation formula. As a result, $\sum_{k=k_3}^{\lfloor np \rfloor} b_{k-np} \lesssim_q \sqrt{q}\cdot(npq)^{q/2} \asymp (npq)^{q/2}$. Combining the previous equations shows that for $n \geq 100q$,

$$\mathbb{E}|Y-np|^q = \sum_{k=0}^{n} b_{k-np} \asymp_q (npq)^{q/2}.$$

We now treat the case $n \leq 100q$. In that case, $\frac{1}{100} \leq p \leq \frac{1}{2}$ so that $2^n, p^n \asymp_q 1$. Hence,

$$\mathbb{E}|Y-np|^q \asymp_q \sum_{k=0}^{n} |k-np|^q \asymp_q q^q \asymp_q (npq)^{q/2}.$$

Regime 2: $\frac{q}{ne^q} \le p \le \frac{q}{2n}$. Again, we start with the right tails. For convenience, we let

$$L := \frac{q}{\ln \frac{q}{np}}.$$

We note that $L \gtrsim np$. Using similar computations as in Eq (6), for $l \ge L$, we have

$$\frac{b_{l+1}}{b_l} \le \frac{np}{l} e^{q/l} \le \frac{np}{L} e^{q/L} = \sqrt{\frac{np}{q}} \ln \frac{q}{np} \le \frac{2}{e}.$$

Hence, after $l = L$, the decay of b_l is exponential. Hence, $\sum_{k=\lceil np+L \rceil}^{n} b_{k-np} \asymp b_{\lceil np+L \rceil - np}$. This also shows that if k_{max} is the integer for which $b_{k_{max}-np}$ is maximized and $k_{max} \ge np$, we have $k_{max} - np \le L$. As a result

$$\begin{aligned} b_{k_{max}-np} \le \mathbb{E}\left[|Y-np|^q \mathbb{1}_{Y>np}\right] &= \sum_{np \le k \le n} b_{k-np} \\ &\le (1+L) b_{k_{max}-np} + \sum_{k \ge np+L} b_{k-np} \asymp L b_{k_{max}-np}. \end{aligned}$$

Now note that $1 \lesssim L \lesssim q$, so that $L \asymp_q 1$. Thus, with $l_{max} = k_{max} - np$,

$$\mathbb{E}\left[|Y-np|^q \mathbb{1}_{Y>np}\right] \asymp_q b_{l_{max}}.$$

We recall that we already know $l_{max} \le L$. Now for any $l \in [0, L]$, by Stirling's approximation formula,

$$b_l \asymp \frac{1}{\sqrt{np}} \frac{l^q}{\left(1+\frac{l}{np}\right)^{np+l} \left(1-\frac{l}{n(1-p)}\right)^{n(1-p)-l}}$$

Now $\frac{q}{e^q} \le np \le q$ so that $\sqrt{np} \asymp_q 1$. Also, since $l \le L \lesssim q \le n$, we have

$$1 \ge \left(1 - \frac{l}{n(1-p)}\right)^{n(1-p)-l} \ge \left(1 - \frac{2L}{n}\right)^n = e^{-\mathcal{O}(L)} \asymp_q 1.$$

Next, $1 \le \left(1+\frac{l}{np}\right)^{np} \le e^l - e^{\mathcal{O}(q)} \asymp_q 1$. As a result, for $l \in [0, L]$,

$$b_l \asymp_q \frac{l^q}{\left(1+\frac{l}{np}\right)^l}.$$

Now if $l \le np$, we have $b_l \lesssim_q (np)^q \lesssim_q L^q$. On the other hand, if $l \gtrsim np$,

$$\left(1+\frac{l}{np}\right)^l = \exp\left(l\mathcal{O}\left(\ln\frac{l}{np}\right)\right) \le \exp\left(L\mathcal{O}\left(\ln\frac{q}{np}\right)\right) = e^{\mathcal{O}(q)} \asymp_q 1.$$

Hence, we obtained that for $np \lesssim l \le L$, $b_l \asymp_q L^q$, while for $0 \le l \le np$ $b_l \lesssim_q L^q$. As a result we obtained

$$\mathbb{E}\left[|Y-np|^q \mathbb{1}_{Y>np}\right] \asymp_q b_{l_{max}} \asymp L^q.$$

The left tail bound is immediate since

$$\mathbb{E}\left[|Y-np|^q \mathbb{1}_{Y<np}\right] \le (np)^q \lesssim_q L^q.$$

Combining the two previous equations gives the desired result $\mathbb{E}|Y-np|^q \asymp_q L^q$.

Regime 3: $p \le \frac{q}{ne^q}$**.** In particular, $p \le \frac{1}{2n}$ so that $(1-p)^n \asymp 1$. Hence, noting that $|k - np| \asymp k$ for any $k \ge 1$, we obtain

$$\mathbb{E}|Y - np|^q \asymp_q np + \sum_{k=2}^{n} \binom{n}{k} p^k k^q. \tag{9}$$

Now note that

$$\sum_{k=2}^{n} \binom{n}{k} p^k k^q \le \sum_{k\ge 2} \frac{(np)^k}{k!} k^q$$

Letting $a_k := \frac{(np)^k}{k!} k^q$, we have for any $k \ge 2$

$$\frac{a_{k+1}}{a_k} = \frac{np}{k+1}\left(1 + \frac{1}{k}\right)^q \le \frac{qe^{q/k}}{ke^q} \le \frac{q}{2e^{q/2}} \le \frac{1}{e}.$$

As a result,

$$\sum_{k=2}^{n} \binom{n}{k} p^k k^q \le \frac{e}{2(e-1)} (np)^2 2^q.$$

Plugging this into Eq (9) yields

$$\mathbb{E}|Y - np|^q \asymp_q np.$$

This ends the proof of the lemma. ■

We are now ready to prove the following result, which gives general bounds on $\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q$ as well the asymptotic convergence rate when $q \ge 2$ up to a factor $\Theta(\sqrt{q})$.

Proposition 25 *Let $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ such that $\|\boldsymbol{p}\|_1 < \infty$. For $q \ge 1$ and $n \ge 1$, we have*

$$\frac{1}{\sqrt{n}}\left(\sum_{p(i)\ge \frac{1}{n}} p(i)^{q/2}\right)^{1/q} + \left(\sum_{p(i)\le \frac{1}{n}} p(i)^q\right)^{1/q} \lesssim \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \lesssim \left(\frac{1}{n^q}\sum_{i\ge 1} \psi_q(n, p(i))\right)^{1/q}.$$

Proof We start by observing that by Jensen's inequality, one has

$$\left(\sum_{i\ge 1} (\mathbb{E}|\hat{p}_n(i) - p(i)|)^q\right)^{1/q} \le \mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q \le (\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q^q)^{1/q}. \tag{10}$$

The right-hand side inequality uses the convexity of $x \ge 0 \mapsto x^q$ and the left-hand side uses the convexity of $x \ge 0 \mapsto (c + x^q)^{1/q}$ for any fixed $c \ge 0$. Now using Theorem 24, we have

$$(\mathbb{E}\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_q^q)^{1/q} \asymp \left(\frac{1}{n^q}\sum_{i\ge 1} \psi_q(n, p(i))\right)^{1/q},$$

which gives the desired upper bound. Next,

$$\left(\sum_{i\geq 1}(\mathbb{E}|\hat{p}_n(i)-p(i)|)^q\right)^{1/q} \asymp \left(\sum_{i\geq 1}\left(\sqrt{\frac{p(i)}{n}}\vee p(i)\right)^q\right)^{1/q}$$
$$\asymp \frac{1}{\sqrt{n}}\left(\sum_{p(i)\geq\frac{1}{n}} p(i)^{q/2}\right)^{1/q} + \left(\sum_{p(i)\leq\frac{1}{n}} p(i)^q\right)^{1/q}.$$

This ends the proof of the proposition. ■

We are now ready to prove the asymptotic bounds from Proposition 8 using the previous result.

Proof of Proposition 8 We now analyze the asymptotic convergence of $\mathbb{E}\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_q$ when $q\geq 2$. Since $q\geq 2$, we have $\|\boldsymbol{p}\|_{q/2}^{q/2}\leq\|\boldsymbol{p}\|_1<\infty$. For any $i\geq 1$, for $n\geq 1/p(i)$, the previous result shows that

$$\mathbb{E}\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_q \gtrsim \frac{1}{\sqrt{n}}\left(\sum_{j\leq i} p(j)^{q/2}\right)^{1/q}.$$

Because this holds for any $i\geq 1$, we obtain the desired lower bound

$$\liminf_{n\to\infty}\sqrt{n}\mathbb{E}\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_q \gtrsim \left(\sum_{j\geq 1} p(j)^{q/2}\right)^{1/q} = \sqrt{\|\boldsymbol{p}\|_{q/2}}.$$

For the upper bound, we first simplify the characterization of the central moments of binomials given in Lemma 24. We obtain directly

$$\psi_q(n,p)\lesssim_q \begin{cases}(npq)^{q/2} & p\geq\frac{q}{2n}\\ npq^q & p\leq\frac{q}{2n},\end{cases}$$

where $\lesssim_q$ hides factors $\mathcal{O}(C^q)$ for a universal constant $C>0$. Here we only simplified the second regime for which $np\leq 1$. As a result, we have that

$$(\mathbb{E}\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_q^q)^{1/q} \lesssim \sqrt{\frac{q}{n}}\left(\sum_{p(i)\geq\frac{q}{2n}} p(i)^{q/2}\right)^{1/q} + \frac{1}{n^{1-1/q}}\left(\sum_{p(i)<\frac{q}{2n}} p(i)\right)^{1/q}$$
$$\leq \sqrt{\frac{q\|\boldsymbol{p}\|_{q/2}}{n}} + \frac{1}{\sqrt{n}}\left(\sum_{p(i)<\frac{q}{2n}} p(i)\right)^{1/q}.$$

It suffices to note that $\sum_{p(i)<\frac{q}{2n}} p(i)=o(1)$ as $n\to\infty$ to obtain

$$\mathbb{E}\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_q \leq (\mathbb{E}\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_q^q)^{1/q} \lesssim \sqrt{\frac{q\|\boldsymbol{p}\|_{q/2}}{n}} + o\left(\frac{1}{\sqrt{n}}\right).$$

This ends the proof of the proposition. ■

Appendix E. Proofs of the results on high-probability bounds

In this section, we provide high-probability bounds on the maximum deviation of the empirical mean for product distributions on $\{0,1\}$. We will need the following simple lemma on small empirical mean deviations of binomials.

Lemma 26 *Let $p \in (0, \frac{1}{2}]$ and $n \geq \frac{4}{p}$. Then, for $\frac{1}{\sqrt{np}} \leq \delta \leq 1$,*

$$\mathbb{P}_{Y\sim\mathcal{B}(n,p)}\left(|Y-np| \leq \delta\sqrt{np}\right) \asymp \delta.$$

In particular, for any $\delta \in (0,1]$, there is a universal constant $c>0$ for which

$$\mathbb{P}_{Y\sim\mathcal{B}(n,p)}\left(|Y-np| \vee 1 \leq c\delta\sqrt{np}\right) \leq \delta.$$

Proof Let $k \in [np-\sqrt{np}, np+\sqrt{np}]$ be an integer. We write $k = n(p+l)$ and note that $p+l \leq \frac{3}{4}$ because of the hypothesis. Then, using Stirling's formula

$$\mathbb{P}(Y=k) = \binom{n}{k} p^k (1-p)^{n-k} \asymp \frac{1}{\sqrt{n}} \left(\frac{p}{p+l}\right)^k \left(\frac{1-p}{1-p-l}\right)^{n-k}.$$

But $\left(\frac{p}{p+l}\right)^k = \exp(-n(p+l)\ln(1+l/p)) \asymp \exp(-np\ln(1+l/p)) \asymp e^{-nl}$. Also, $\left(\frac{1-p}{1-p-l}\right)^{n-k} = \exp(n(1-p-l)\ln(1+l/(1-p-l))) \asymp e^{nl}$. Hence we obtained that there exist constants $b_1, b_2 > 0$ such that for any $k \in [np-\sqrt{np}, np+\sqrt{np}] \cap \mathbb{N}$, one has

$$\frac{b_1}{\sqrt{n}} \leq \mathbb{P}(Y=k) \leq \frac{b_2}{\sqrt{n}}. \tag{11}$$

By hypothesis, this set $[np-\sqrt{np}, np+\sqrt{np}] \cap \mathbb{N}$ contains at least $2\sqrt{np}-1 \asymp 2\sqrt{np}$ elements. Now let $\frac{1}{\sqrt{np}} \leq \delta \leq 1$. Then, the interval $[np-\delta\sqrt{np}, np+\delta\sqrt{np}] \cap \mathbb{N}$ contains at least $\frac{1}{2}\delta\sqrt{np}$ elements. Together with Eq (11) this gives

$$\mathbb{P}(|Y-np| \leq \delta\sqrt{np}) \asymp \delta\mathbb{P}(|Y-np| \leq \sqrt{np}) \asymp \delta.$$

In the last inequality, we used Chernoff's bound. ∎

We prove the following bounds, which in particular include those stated in Proposition 9.

Proposition 27 *Let $\gamma \in (0, \frac{1}{2})$ and $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ such that there exists $j \geq 1$ with $p(j) \geq \frac{\gamma}{2nj}$. Then, for some universal constants $a_1, a_2, a_3 > 0$,*

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \geq a_1 \sup_{j\geq 1} \phi_{\frac{j}{\gamma}, p(j)}(n)\right) \leq \gamma \quad \textit{and} \quad \mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \geq a_2 \sup_{j\geq 1} \phi_{\frac{j}{\gamma}, p(j)}(n)\right) \geq a_3\gamma$$

Also,

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \leq a_2 \sup_{j\geq 1} \phi_{\frac{j}{\ln 1/\gamma}, p(j)}(n)\right) \leq \gamma$$

$$\textit{and} \quad \mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \leq a_1 \left(\sup_{j\geq 1} \phi_{\frac{j}{\ln 1/\gamma}, p(j)}(n) \vee \frac{1}{n}\right)\right) \geq \gamma^{3\ln\frac{1}{\gamma}}.$$

Let $\boldsymbol{p} \in [0, \frac{1}{2}]^{\mathbb{N}}_{\downarrow 0}$ such that $p(j) \geq \frac{\gamma}{2nj}$ for all $j \geq 1$. Then,

$$\mathbb{P}\left(\sup_{j\geq 1} \hat{p}_n(j) \geq \frac{2}{n}\right) \leq \gamma \quad \textit{and} \quad \mathbb{P}\left(\sup_{j\geq 1} \hat{p}_n(j) = \frac{1}{n}\right) \asymp 1 \wedge n \sum_{j\geq 1} p(j).$$

Proof We start with the first claim which gives upper bounds on $\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty$. The only difference with the proofs for the bounds in expectation is that instead of $\varepsilon_{J,q}(n)$, we will use instead.

$$\varepsilon_{J/\gamma,q}(n) = \inf\left\{\varepsilon \geq 0 : \mathbb{P}_{Y\sim\mathcal{B}(n,q)}\left(\frac{Y}{n} \geq q + \varepsilon\right) \leq \frac{\gamma c_0}{2J}\right\}.$$

Define $\varepsilon = \sup_{j\geq 1} \varepsilon_{j/\gamma,p(j)}(n;\gamma)$.

We start with the case when there exists $j \geq 1$ for which $p(j) \leq \frac{\gamma}{2nj}$. Then, Proposition 13 shows that $\varepsilon \geq 0$. We focus on the right tails. The proofs of the high-probability bounds now follow exactly the proof of Proposition 12. Let $\tilde{\varepsilon} = (\varepsilon \wedge \frac{1}{4}) \vee \frac{1}{n}$. The proof of Proposition 12 gives for any $i \geq 1$,

$$D(p(i) + \tilde{\varepsilon} \parallel p(i)) \geq \frac{\ln \frac{2i}{\gamma}}{nC},$$

then using Chernoff's bound and the union bound,

$$\mathbb{P}\left(\sup_{j\geq 1} \hat{p}_n(j) - p(j) \geq 2C\tilde{\varepsilon}\right) \leq \sum_{i\geq 1} \left(\frac{\gamma}{2i}\right)^2 \leq \frac{\gamma}{2}. \tag{12}$$

On the other hand, the proof of Proposition 12 also shows that this high-probability maximum deviation is tight up to constants. Indeed, it shows that for $i \geq 1$ such that $\varepsilon_i := \varepsilon_{i/\gamma,p(i)}(n) \geq \frac{1}{2}\varepsilon$, one has for any $j \leq i$ with $p(j) \geq \frac{1}{n}$ or $p(j) \leq \frac{\varepsilon_i}{2}$,

$$\mathbb{P}\left(\hat{p}_n(j) \geq p(j) + \frac{\varepsilon_i}{2C}\right) \geq \frac{\gamma c_0^2}{2i}.$$

If $j \leq i$, $p(j) \leq \frac{1}{n}$ and $p(j) \geq \frac{\varepsilon_i}{2}$, the proof then showed that $\mathbb{P}(\hat{p}_n(j) \geq p(j) + \frac{\varepsilon_i}{2}) \geq \frac{1}{144}$. Hence, we obtain with the same proof

$$\mathbb{P}\left(\sup_{j\geq 1} \hat{p}_n(j) - p(j) \geq \frac{\varepsilon_i}{2C}\right) \geq \frac{1}{144} \vee (1 - e^{-\gamma c_0^2/2}) \gtrsim \gamma.$$

We next turn to the left tails and show that these are dominated by the right tails. We first note that $D(p - 2C\tilde{\varepsilon} \parallel p) \geq D(p + 2C\tilde{\varepsilon} \parallel p)$. Indeed, with $f(x) = D(q + x \parallel q)$, we have $f'(x) = \ln \frac{q+x}{q} - \ln \frac{1-q-x}{1-q}$. Hence, if $x \geq 0$, $f'(x) + f'(-x) = \ln\left(1 - \frac{x^2}{q^2}\right) - \ln\left(1 - \frac{x^2}{(1-q)^2}\right) \leq 0$. Together with the fact that f achieves its minimum at $x = 0$ ends the proof of the claim. We then use the union bound together with Chernoff's bound as in Eq (12) to obtain

$$\mathbb{P}\left(\sup_{j\geq 1} p(j) - \hat{p}_n(j) \geq 2C\tilde{\varepsilon}\right) \leq \sum_{i\geq 1} e^{-nD(p-2C\tilde{\varepsilon}\|p)} \leq \sum_{i\geq 1} e^{-nD(p+2C\tilde{\varepsilon}\|p)} \leq \frac{\gamma}{2}.$$

We next turn to lower bounds on $\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty$. Again, we heavily use the proof of Proposition 12, but here we will use $\varepsilon_{j/\ln\frac{1}{\gamma},p(j)}(n)$ instead of $\varepsilon_{j,p(j)}(n)$. We define $\eta = \sup_{j\geq 2\ln\frac{1}{\gamma}/c_0^2} \varepsilon_{\frac{jc_0^2}{2\ln\frac{1}{\gamma}},p(j)}(n)$ and suppose for now that $\eta \geq 0$. Let $i \geq 2\ln\frac{1}{\gamma}/c_0^2$ be an integer such that $\eta_i := \varepsilon_{\frac{ic_0^2}{2\ln\frac{1}{\gamma}},p(i)}(n) \geq \frac{\eta}{2}$. Then, the proof of Proposition 12 shows that for all $j \leq i$, if either $p(j) \leq \frac{\eta_i}{2}$ or $p(j) \geq \frac{1}{n}$,

$$\mathbb{P}\left(\hat{p}_n(j) \geq p(j) + \frac{\eta_i}{2C}\right) \geq \frac{\ln\frac{1}{\gamma}}{i}.$$

On the other hand, if there is $j \leq i$ such that $\frac{\eta_i}{2} \leq p(j) \leq \frac{1}{n}$, then $\eta \leq 2\eta_i \leq \frac{4}{n}$. Then, we have directly that

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \geq \frac{\eta}{4}\right) = 1 \geq 1 - \gamma.$$

As a result, in both cases, this gives

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \geq \frac{\eta}{4C}\right) \geq 1 - \left(1 - \frac{\ln\frac{1}{\gamma}}{i}\right)^i \geq 1 - \gamma.$$

We can then relate this equation to the quantities $\phi_{j/\ln\frac{1}{\gamma},p(j)}(n)$, using Proposition 13, and the fact that a constant factor in J only affects these quantities up to a constant factor. We then obtain directly for some universal constant a_1,

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \geq a_1 \sup_{j\geq 2\ln\frac{1}{\gamma}/c_0^2} \phi_{j/\ln\frac{1}{\gamma},p(j)}(n)\right) \geq 1 - \gamma. \tag{13}$$

We next consider the case when $\eta < 0$. By Proposition 13, this implies that for any $j \geq 2\ln\frac{1}{\gamma}/c_0^2$, we have $p(j) \leq \frac{\ln\frac{1}{\gamma}}{c_0^2 nj}$. In particular, $\phi_{j/\ln\frac{1}{\gamma},p(j)}(n) \lesssim \frac{1}{n}$, hence obtaining Eq (13) in this case is immediate.

We now focus on the indices $i \leq 2\ln\frac{1}{\gamma}/c_0^2$. Note that if $p(i) \leq \frac{4}{n}$, then we again have $\phi_{j/\ln\frac{1}{\gamma},p(j)}(n) \lesssim \frac{1}{n}$. Without loss of generality, we can therefore suppose that $p(i) \geq \frac{4}{n}$. Lemma 26 implies that for some constant $c_1 > 0$, for any $j \leq i$,

$$\mathbb{P}\left(|\hat{p}_n(j) - p(j)| \vee \frac{1}{n} \leq c_1\gamma^{1/i}\sqrt{\frac{p(i)}{n}}\right) \leq \gamma^{1/i}.$$

As a result,

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \leq c'\gamma^{1/i}\sqrt{\frac{p(i)}{n}}\right) \leq \gamma.$$

In particular, this shows that for some universal constant $c_2 > 0$, we have

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \leq c_2 \sup_{j\leq 2\ln\frac{1}{\gamma}/c_0^2} \phi_{j/\ln\frac{1}{\gamma},p(j)}(n)\right) \leq \gamma.$$

Together with Eq (13), we showed the desired bound for some constant $c_3 > 0$,

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \vee \frac{1}{n} \leq c_3 \sup_{j\geq 1} \phi_{j/\ln\frac{1}{\gamma},p(j)}(n) \right) \leq \gamma.$$

Next, we turn to the second inequality for the lower bound on $\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty$. The proof of Proposition 12 shows that with $\tilde{\varepsilon}' = \frac{1}{n} \vee \sup_{j\geq \ln\frac{1}{\gamma}} \varepsilon_{j/\ln\frac{1}{\gamma},p(j)}(n)$, for $j \geq \ln\frac{1}{\gamma}$, we have

$$D(p(j) + 2C\tilde{\varepsilon}' \parallel p(j)) \geq \frac{2\ln(2j/\ln\frac{1}{\gamma})}{n}.$$

As a result, using Chernoff's bound and the fact that $D(p(j) - 4C\tilde{\varepsilon}' \parallel p(j)) \geq D(p(j) + 4C\tilde{\varepsilon}' \parallel p(j))$, we obtain

$$\mathbb{P}(|\hat{p}_n(i) - p(i)| \geq 4C\tilde{\varepsilon}') \leq 2e^{-nD(p(j)+4C\tilde{\varepsilon}'\|p(j))} \leq \frac{\ln^4\frac{1}{\gamma}}{8i^4}.$$

Using the the inequality $\ln(1-x) \geq -2x$ for $x \in [0, \frac{1}{2}]$, we obtain

$$\mathbb{P}\left(\sup_{j\geq 1+\ln\frac{1}{\gamma}} |\hat{p}_n(j) - p(j)| \leq 4C\tilde{\varepsilon}' \right) \geq \exp\left(-\sum_{j\geq 1+\ln\frac{1}{\gamma}} \frac{\ln^4\frac{1}{\gamma}}{4j^4} \right) \geq \exp\left(\frac{\ln\frac{1}{\gamma}}{12} \right) = \gamma^{1/12}.$$

From Proposition 13, we have that

$$\tilde{\varepsilon} \asymp \frac{1}{n} \vee \sup_{j\geq \ln\frac{1}{\gamma}} \phi_{j/\ln\frac{1}{\gamma},p(j)}(n).$$

Hence, it only remains to focus on the indices $i \leq 1 + \ln\frac{1}{\gamma}$. First, note that in this case $\phi_{j/\ln\frac{1}{\gamma},p(j)}(n) \asymp \gamma^{1/i}\sqrt{\frac{p(i)}{n}}$. By Lemma 26, for some constant $C_4 > 0$, we have

$$\mathbb{P}\left(\sup_{j\leq 1+\ln\frac{1}{\gamma}} |\hat{p}_n(j) - p(j)| \leq C_4 \left(\sup_{j\leq 1+\ln\frac{1}{\gamma}} \phi_{j/\ln\frac{1}{\gamma},p(j)}(n) \vee \frac{1}{n} \right) \right) \geq \gamma^{2\ln\frac{1}{\gamma}}.$$

Putting everything together yields the desired lower bound for some constant $C_5 > 0$ sufficiently large

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_\infty \leq C_5 \left(\sup_{j\geq 1} \phi_{j/\ln\frac{1}{\gamma},p(j)}(n) \vee \frac{1}{n} \right) \right) \geq \gamma^{3\ln\frac{1}{\gamma}}.$$

We now treat the case when $p(j) \leq \frac{\gamma}{2nj}$ for all $j \geq 1$, for which we use the proof of Proposition 21. It directly gives with $\hat{P}_n := \sup_{j\geq 1} \hat{p}_n(j)$ and $U(\boldsymbol{p}) = \sup_{j\geq 1} njp(j) \leq \frac{\gamma}{2}$ that

$$\mathbb{P}\left(\hat{P}_n \geq \frac{2}{n} \right) \leq \frac{\pi^2}{6} U(\boldsymbol{p})^2 \leq \frac{\gamma}{2}.$$

On the other hand with $V(\boldsymbol{p}) := n\sum_{j\geq 1} p(j)$ which satisfies $V(\boldsymbol{p}) \wedge 1 \geq U(\boldsymbol{p})$, we have

$$c_4 V(\boldsymbol{p}) \wedge 1 \leq \mathbb{P}\left(\hat{P}_n \geq \frac{1}{n} \right) \leq V(\boldsymbol{p}) \wedge 1,$$

for some constant $c_4 > 0$. ∎