Uniform Convergence Beyond Glivenko-Cantelli

Tanmay Devale TDEVALE@PURDUE.EDU
Pramith Devulapalli PDEVULAP@PURDUE.EDU
Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

We characterize conditions under which collections of distributions on $\{0,1\}^{\mathbb{N}}$ admit uniform estimation of their mean. Prior work from Vapnik and Chervonenkis (1971) has focused on uniform convergence using the empirical mean estimator, leading to the principle known as $P-$ Glivenko-Cantelli. We extend this framework by moving beyond the empirical mean estimator and introducing Uniform Mean Estimability, also called UME-learnability, which captures when a collection permits uniform mean estimation by any arbitrary estimator. We work on the space created by the mean vectors of the collection of distributions. For each distribution, the mean vector records the expected value in each coordinate. We show that separability of the mean vectors is a sufficient condition for UME-learnability. However, we show that separability of the mean vectors is not necessary for UME-learnability by constructing a collection of distributions whose mean vectors are non-separable yet UME-learnable using techniques fundamentally different from those used in our separability-based analysis. Finally, we establish that countable unions of UME-learnable collections are also UME-learnable, solving the conjecture posed in Cohen et al. (2025).

Keywords: Glivenko-Cantelli, Uniform Convergence, Uniform Mean Estimation

1. Introduction

The seminal work of Vapnik and Chervonenkis (1971) establishes that for any binary function class $\mathcal{F}$, finite VC dimension guarantees uniform convergence independent of the distribution. However, in settings where $\mathcal{F}$ admits infinite VC dimension, uniform convergence can still hold for some distributions provided specific properties are satisfied; in such cases, we say $\mathcal{F}$ satisfies the $P-$Glivenko-Cantelli property, which is described as

$$\mathbb{E}_{S\sim P^n}\left[\sup_{f\in\mathcal{F}}|\mathbb{P}_n f - Pf|\right] \xrightarrow{n\to\infty} 0, \tag{1}$$

where S is a set of n i.i.d. data points sampled from P, $\mathbb{P}_n f$ is the empirical mean of f computed using S, and Pf is the true mean of f. The result of Vapnik and Chervonenkis (1971) characterizes the distributions that satisfy the $P-$ Glivenko-Cantelli property for any binary function class $\mathcal{F}$.
The work of Cohen et al. (2025) raised the direction of going beyond the empirical mean estimator and posited a conjecture, which we presently resolve and prove a stronger result in this paper. In this work, we consider collections of distributions defined over a countable function class $\mathcal{F}$ that admit uniform mean estimation even if it fails to satisfy the $P-$ Glivenko-Cantelli property. Our goal is to identify conditions under which the following holds:

$$\mathbb{E}_{S\sim P^n}\left[\sup_{f\in\mathcal{F}}|\mathscr{P}_n f - Pf|\right] \xrightarrow{n\to\infty} 0, \tag{2}$$

where $\mathscr{P}_n f$ is an arbitrary estimator that uses S to estimate Pf the true mean of f.
For countable, binary-valued concept classes, $P-$Glivenko–Cantelli can also be equivalently expressed as a distribution on $\{0,1\}^{\mathbb{N}}$ where the j^{th} coordinate of the distribution is equivalent to the output of the j^{th} function. The works of Cohen and Kontorovich (2023); Cohen et al. (2025) can be understood through this lens, and it is the one we adopt. This alternative formulation allows us to work directly with distributions over $\{0,1\}^{\mathbb{N}}$, thereby abstracting away the explicit choice of a concept class. Formally speaking, equation (1) can be rewritten as equation (3) below in the following way:

$$\mathbb{E}_{S\sim\mu^n}\left[\sup_{j\in\mathbb{N}}|\hat{q}_j - q_j|\right] \;=\; \mathbb{E}_{S\sim\mu^n}\left\|\hat{q}-q\right\|_\infty \tag{3}$$

where μ is a distribution on $\{0,1\}^{\mathbb{N}}$, S is a sample of size n, and $\hat{q}_j$ and q_j are the empirical mean and true mean respectively for the j^{th} function in the countable function class.
In this paper, we characterize properties of a collection of distributions $\mathcal{Q}$ on $\{0,1\}^{\mathbb{N}}$ that ensure (3) converges to 0 as $n\to\infty$, and we address the broader question of replacing the empirical estimator $\hat{q}$ with an arbitrary estimator $\tilde{q}$. More technically, we define Uniform Mean Estimation (UME) learnability[1] in the following way: there exists an algorithm $\mathcal{A}$ such that for any ground truth distribution $\mu^*\in\mathcal{Q}$, given n data points, it produces an estimate $\tilde{q}$ of the true mean vector $q\in[0,1]^{\mathbb{N}}$ satisfying

$$\mathbb{E}_{S\sim\mu^n}\left\|\mathcal{A}(S)-q\right\|_\infty = \mathbb{E}_{S\sim\mu^n}\left\|\tilde{q}-q\right\|_\infty \xrightarrow{n\to\infty} 0. \tag{4}$$

The motivation for this framework can be seen in the limitations of the empirical mean as observed in Cohen and Kontorovich (2023). Consider the collection of distributions $\mathcal{Q}=\{\mu\}$ where μ is a product measure and $\text{Mean}(\mu)=\left(\frac{1}{2},\frac{1}{2},\ldots\right)$. As the coordinates of μ are independent, we obtain $X_j\sim\text{Bernoulli}\left(\frac{1}{2}\right)$. An obvious algorithm is to return the mean of the only distribution in the collection, which is $\left(\frac{1}{2},\frac{1}{2},\ldots\right)$. But even for such a trivial collection, the empirical mean estimator fails. The probability of obtaining all 0s or all 1s for n data points at a particular coordinate is positive. There are infinitely many coordinates at which this could occur; hence, it will almost surely occur. Hence, the empirical mean estimator cannot be used to estimate this collection of distributions even though it consists of only one distribution.

Our Contributions

- **Separability implies Learnability:** We study when collections of distributions $\mathcal{Q}$ are UME-learnable and prove that if the collection of mean vectors corresponding to $\mathcal{Q}$ is separable, then $\mathcal{Q}$ is UME-learnable (Theorem 7).
- **Closed under Unions:** We prove that any countable union of UME-learnable collections of distributions is also UME-learnable (Theorem 11). In particular, this resolves the conjecture of Cohen et al. (2025) for the union of two families and extends it to countably many families.
- **Beyond Separability:** A natural question we tackle is whether separability is necessary for UME-learnability. We illustrate that it is not a necessary condition by constructing a collection $\mathcal{Q}$ whose space of mean vectors is non-separable, yet it is UME-learnable (Proposition 9). Moreover, our construction utilizes techniques fundamentally different from those used in Theorem 7, which may be of independent interest.

1. The term "uniform" here refers to uniformity over indices, not uniformity over distributions (though see Appendix C).

2. Related Works

Classical Empirical Process Theory Our work stems from classical empirical process theory, which aims to characterize the conditions under which the empirical estimator converges to the true mean uniformly over a class of functions. For binary functions, Vapnik and Chervonenkis (1971) provides necessary and sufficient conditions that are independent of the underlying distribution guaranteed by the finiteness of a combinatorial quantity known as the VC dimension. They also characterize distributions for which the empirical mean estimator is a uniform estimator for the true mean using VC entropy. The subsequent work Vapnik and Chervonenkis (1981) obtains that sub-exponential growth of the empirical covering numbers is also necessary and sufficient for uniform convergence. Modern expositions and refinements of these results can be found in Vapnik (2006); van der Vaart and Wellner (2023).

Product Measures on $\{0,1\}^{\mathbb{N}}$ Cohen and Kontorovich (2023) study product measures on $\{0,1\}^{\mathbb{N}}$ that are uniformly estimable by the empirical mean estimator. They identify the largest collection of estimable product measures, which they call the LGC class. They show that LGC consists of exactly those distributions whose mean vectors q satisfy $T(q) = \sup_{j\in\mathbb{N}} \frac{\log(j+1)}{\log(1/q_j)}$ is finite. We are motivated by their framework in developing the notion of UME-learnability over countable function classes, but we differ in two respects: we drop the reliance on product measures and assume any arbitrary mean estimator.

Dependent Coordinates and Arbitrary Estimators The more recent works of Blanchard et al. (2024) and Cohen et al. (2025) extend the analysis from Cohen and Kontorovich (2023). Blanchard et al. (2024) drop the assumption of product measures and analyze necessary and sufficient conditions of uniform convergence of the empirical mean estimator to the true mean when different coordinates can be correlated. On the other hand, Cohen et al. (2025) explores other arbitrary mean estimators besides the empirical mean estimator while keeping their attention focused on product measures. They derive specific conditions that a product measure must satisfy for it to be UME-learnable by the empirical mean estimator. They also provide non-trivial extensions of the LGC class when certain restrictions are relaxed.

Infinite-Dimensional Exponential Families Sriperumbudur et al. (2013) studies an infinite dimensional exponential family of densities and constructs an estimator that can effectively predict the unknown density. Our setting is fundamentally different because we do not assume a common reference measure on which to define a density. We work with collections of measures defined on the space $\{0,1\}^{\mathbb{N}}$ without assuming any common dominating measure. As a result, our analysis falls outside the scope of Sriperumbudur et al. (2013).

2.1. Notation

For any $k \in \mathbb{N}$, we write $[k] = \{i \in \mathbb{N} : i \leq k\}$. All logarithms are base e unless otherwise specified. The floor and ceiling functions are denoted by $\lfloor t \rfloor$ and $\lceil t \rceil$ for $t \in \mathbb{R}$ mapping t to the nearest integer below or above, respectively. Unspecified constants c, c' may change from line to line.

We denote our collection of distributions with $\mathcal{Q}$. For any distribution μ with mean q, a data point is denoted by X, indicates $X \sim \mu$. The realization denoted by $X_j^{(i)}$ refers to the j^{th} coordinate of the i^{th} data point. We overload our notation and use the superscript to enumerate a countable collection of distributions. For example, if $\mathcal{Q}$ is a countable collection then μ_j^i denotes the j^{th}

coordinate of the i^{th} distribution in the collection. The same convention applies to means. The measure-theoretic nuisances of defining distributions on $\{0,1\}^{\mathbb{N}}$ have been addressed in Cohen and Kontorovich (2023).
Some of the results in this paper and the literature are specific to product measures, hence we say $\mu = \text{Prod}(q)$ if $X \sim \mu$ is equivalent to $X_j \sim \text{Bernoulli}(q_j)$ for every $j \in \mathbb{N}$. We say a collection of distributions is a collection of product measures if for every $\mu \in \mathcal{Q}, \mu = \text{Prod}(q)$ for some $q \in [0,1]^{\mathbb{N}}$.

3. Definitions and Main Results

For any distribution μ on $\{0,1\}^{\mathbb{N}}$, letting $X \sim \mu$, $\text{Mean}(\mu) = \mathbb{E}X$ denote its mean vector, and for each coordinate $j \in \mathbb{N}$, $[\text{Mean}(\mu)]_j = \mathbb{E}X_j$. Specifically for our setting, we define an estimator $\tilde{q}$ as a mapping from $(\{0,1\}^{\mathbb{N}})^n$ to $[0,1]^{\mathbb{N}}$ where n is the number of data points. The estimator will be the output of some algorithm $\mathcal{A}$.

Definition 1 *We say a collection of distributions $\mathcal{Q}$ is **Uniform Mean Estimation (UME) learnable** by algorithm $\mathcal{A}$ if for any distribution $\mu \in \mathcal{Q}$, the algorithm $\mathcal{A}$ returns an estimate $\tilde{q}$ using n i.i.d. data points $S = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ obtained from μ such that*

$$\mathbb{E}_{S\sim\mu^n} \left\|\mathcal{A}(S) - q\right\|_\infty = \mathbb{E}_{S\sim\mu^n} \left\|\tilde{q} - q\right\|_\infty \xrightarrow{n\to\infty} 0$$

where $q = \text{Mean}(\mu)$. A collection of distributions $\mathcal{Q}$ is UME-learnable if there exists an algorithm $\mathcal{A}$ such that $\mathcal{Q}$ is UME-learnable by $\mathcal{A}$.

For a collection of distributions $\mathcal{Q}$, we can define the corresponding collection of mean vectors as

$$\text{Mean}(\mathcal{Q}) = \{q \in [0,1]^{\mathbb{N}} : q = \text{ Mean}(\mu) \text{ for some } \mu \in \mathcal{Q}\} \tag{5}$$

Definition 2 *We say a collection of distributions $\mathcal{Q}$ has a **countable** $\varepsilon-$**cover** for its mean if there is some $\mathcal{Q}_\varepsilon$ such that for any $q \in \text{Mean}(\mathcal{Q})$, there exists $q_\varepsilon \in \mathcal{Q}_\varepsilon$ such that $\left\|q - q_\varepsilon\right\|_\infty < \varepsilon$ and $\mathcal{Q}_\varepsilon$ is countable.*

Definition 3 *We say a collection of distributions $\mathcal{Q}$ has **separable** mean vectors if for every $\varepsilon > 0$ there exists a countable $\varepsilon-$cover for $\text{Mean}(\mathcal{Q})$. A collection of distributions $\mathcal{Q}$ has **non-separable** mean vectors if for some $\varepsilon > 0$ there does not exist a countable $\varepsilon-$cover for $\text{Mean}(\mathcal{Q})$.*

Definition 4 *Given a collection $\mathcal{Q}$, let $\mathcal{B}(q,\varepsilon)$ denote the ball of radius ε around the vector q under the ℓ_∞ norm defined as follows:*

$$\mathcal{B}(q,\varepsilon) = \left\{q' \in \mathcal{Q} : \left\|q - q'\right\|_\infty < \varepsilon\right\}.$$

Main Results Here is a summary of our main results:

- If $\mathcal{Q}$ is countable then $\mathcal{Q}$ is UME-learnable (Theorem 6).
- If $\mathcal{Q}$ has separable means vectors then $\mathcal{Q}$ is UME-learnable (Theorem 7).
- UME-learnability is closed under countable unions (Theorem 11).
- A UME-learnable collection of distributions with non-separable mean (Proposition 9).
- A discussion of UME-learnability that is uniform over the collection of distributions in Appendix C.

4. Separability is Sufficient for UME-learnability

In this section, we are interested in finding conditions on the collection of distributions $\mathcal{Q}$ to guarantee UME-learnability. We focus on the collection of mean vectors, Mean($\mathcal{Q}$), and show that their separability is a sufficient condition for UME-learnability. By Definition 3, for every $\varepsilon > 0$ we obtain a countable $\varepsilon-$cover of Mean($\mathcal{Q}$). We use this cover to provide an $\varepsilon-$ approximation of the true underlying distribution using Algorithm 1.

Algorithm 1 $\varepsilon-$approximate $(\mathcal{Q}, n > 0, \varepsilon \geq 0)$

Initialize $\mathcal{Q}_\varepsilon = \{q^1, q^2, \ldots\}$ as the countable $\varepsilon-$cover of Mean($\mathcal{Q}$)
Let $i \leftarrow 1$ and $\hat{q}$ be the empirical mean estimator computed using the training data.
while there exists $j < n$ with $|q^i_j - \hat{q}_j| > \sqrt{\frac{3\log n}{n}} + \varepsilon$ and $i \leq n$ **do** $i \leftarrow i + 1$
return q^i

Given an $\varepsilon-$ cover of Mean($\mathcal{Q}$), Algorithm 1 will find the first vector in the cover that is $\varepsilon-$close to the true mean vector. We leverage the fact that a vector that is not $\varepsilon-$close to the true mean vector is ε far in at least one coordinate. We can rule out incorrect vectors by focusing on coordinates where the empirical mean closely matches the true mean. We focus on the first n coordinates, as Hoeffding's inequality provides strong concentration guarantees for them. We return the first vector that is within the confidence bound provided by Hoeffding's inequality on the first n coordinates.

Lemma 5 *If collection of distributions $\mathcal{Q}$ has a countable $\varepsilon-$cover for Mean($\mathcal{Q}$) then for any $\mu \in \mathcal{Q}$ with probability 1 there exists a data size n_0 such that for all $n > n_0$ the estimator $\tilde{q}$ returned by Algorithm 1 satisfies*

$$\|\tilde{q} - q\|_\infty \leq \varepsilon$$

where $q = Mean(\mu)$.

Proof Let $\mathcal{Q}$ be a collection of distributions and $\varepsilon > 0$ be given. Let $\mathcal{Q}_\varepsilon$ be the countable $\varepsilon-$cover of Mean($\mathcal{Q}$) under the ℓ_∞ norm. Let μ^* be the true underlying distribution and let $q^* = \text{Mean}(\mu^*)$. Let $q^{i^*_\varepsilon} \in \mathcal{Q}_\varepsilon$ be the first vector such that $\left\|q^* - q^{i^*_\varepsilon}\right\|_\infty \leq \varepsilon$. We refer to $q^{i^*_\varepsilon}$ as the $\varepsilon-$approximating vector. Let $q^1, q^2, \ldots, q^{i^*_\varepsilon - 1}$ be the vectors appearing before $q^{i^*_\varepsilon}$. Hence, by definition, there exists some coordinate for which the deviation is at least ε. Therefore, for $i < i^*_\varepsilon$, we define

$$j_i = \min\{j \in \mathbb{N} : \left|q^i_j - q^*_j\right| > \varepsilon\}$$

Our task is to find $q^{i^*_\varepsilon}$. When we set the deviation between the empirical mean from the true mean at a particular coordinate as $\sqrt{\frac{3\log n}{n}}$ by Hoeffding inequality (Hoeffding (1963)), we obtain with probability at least $1 - \frac{2}{n^6}$, $\left|q^*_j - \hat{q}_j\right| < \sqrt{\frac{3\log n}{n}}$ for any coordinate i. Consequently, in Algorithm 1 we test the first n coordinates. And as the $\varepsilon-$approximating vector is ε far from the true vector, we allow an ε slack. Thus, we have the following test.

$$\forall j < n, \left|q^i_j - \hat{q}_j\right| < \sqrt{\frac{3\log n}{n}} + \varepsilon$$

For all $i < i^*_\varepsilon$ let

$$\gamma_i = \left|q^i_{j_i} - q^*_{j_i}\right| - \varepsilon$$

We note $\gamma_i > 0$ by definition of the $\varepsilon-$approximating vector and j_i.
As we wish to focus on the first n coordinates, we need to ensure n is large enough to include the coordinates that differentiate the vectors from the true mean vector by at least ε. We also need to ensure that for any $i < i^*_\varepsilon, q^i$ is not accidentally accepted due to the confidence bound given by Hoeffding inequality. In addition, $q^{i^*_\varepsilon}$ should be analyzed by the algorithm. Hence, n should be sufficiently large such that

$$n \geq i^*_\varepsilon \qquad n \geq \max_{i<i^*_\varepsilon} j_i \quad \text{and} \quad \min_{i<i^*_\varepsilon} \gamma_i > 2\sqrt{\frac{3\log n}{n}} \tag{6}$$

We want to ensure that the event F_n, that the algorithm returns any of the vectors preceding the $\varepsilon-$approximating vector, and the event G_n that the algorithm does not return the $\varepsilon-$ approximating vector after obtaining n data points, do not occur infinitely often.
We start by analyzing the probability of F_n. We apply the union bound together with the second constraint in equation (6).

$$\mathbb{P}\left(\exists i < i^*_\varepsilon : \forall j < n, \left|q^i_j - \hat{q}_j\right| < \sqrt{\frac{3\log n}{n}} + \varepsilon\right) \leq \sum_{i=1}^{i^*_\varepsilon - 1} \mathbb{P}\left(\left|q^i_{j_i} - \hat{q}_{j_i}\right| < \sqrt{\frac{3\log n}{n}} + \varepsilon\right) \tag{7}$$

We use the third constraint and triangle inequality to obtain

$$\left|q^i_{j_i} - \hat{q}_{j_i}\right| \geq \left|q^i_{j_i} - q^*_{j_i}\right| - \left|\hat{q}_{j_i} - q^*_{j_i}\right| = \gamma_i + \varepsilon - \left|\hat{q}_{j_i} - q^*_{j_i}\right| \geq 2\sqrt{\frac{3\log n}{n}} + \varepsilon - \left|\hat{q}_{j_i} - q^*_{j_i}\right| \tag{8}$$

We combine equations (7),(8), use the first constraint and apply Hoeffding inequality to obtain

$$\sum_{i=1}^{i^*_\varepsilon - 1} \mathbb{P}\left(\left|q^i_{j_i} - \hat{q}_{j_i}\right| < \sqrt{\frac{3\log n}{n}} + \varepsilon\right) \leq \sum_{i=1}^{i^*_\varepsilon - 1} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| > \sqrt{\frac{3\log n}{n}}\right) \leq \frac{2(i^*_\varepsilon - 1)}{n^6} \leq \frac{2}{n^5}$$

Similarly, we can analyze the probability of the event G_n using the union bound and Hoeffding inequality to obtain

$$\mathbb{P}\left(\exists j < n : \left|q^{i^*_\varepsilon}_j - \hat{q}_j\right| \geq \sqrt{\frac{3\log n}{n}}\right) \leq \sum_{j=1}^{n} \mathbb{P}\left(\left|q^{i^*_\varepsilon}_j - \hat{q}_j\right| \geq \sqrt{\frac{3\log n}{n}}\right) \leq \sum_{j=1}^{n} \frac{2}{n^6} \leq \frac{2}{n^5}$$

We define the event E_n as the occurrence of either F_n or G_n. By our previous analysis we obtain $\mathbb{P}(E_n) \leq \frac{4}{n^5}$. We note that $\sum_{n=1}^{\infty} \mathbb{P}(E_n) \leq \sum_{n=1}^{\infty} \frac{4}{n^5} < \infty$. Hence, we can use the First Borel-Cantelli Lemma to conclude that with probability 1 there exists $n_0 > 0$ such that for all $n > n_0$ the algorithm successfully finds $q^{i^*}_\varepsilon$. ■

As a direct by-product, we can show that any countable collection of distributions is UME-learnable.

Theorem 6 *If $\mathcal{Q}$ is countable then $\mathcal{Q}$ is UME-learnable by Algorithm 1 with $\varepsilon = 0$.*

Proof Let $\mathcal{Q}$ be a countable collection of distributions. We note that $\mathcal{Q}$ is a $0-$cover of itself. We use Lemma 5 with $\varepsilon = 0$ to obtain with probability 1, for any $\mu \in \mathcal{Q}$ with $q = \text{Mean}(\mu)$, there exists n_0 such that for all $n > n_0$ the estimate $\tilde{q}$ returned by Algorithm 1 satisfies $\|\tilde{q} - q\| = 0$. As $\|\tilde{q} - q\|_\infty \leq 1$ by the Dominated Convergence Theorem we obtain $\mathbb{E}\|\tilde{q} - q\|_\infty \xrightarrow{n\to\infty} 0$. ■

We can now consider a collection of distributions that have separable mean vectors and show that they are UME-learnable by Algorithm 2.

Algorithm 2 Separable $(\mathcal{Q}, n > 0)$

Initialize $\mathcal{P} \leftarrow \text{Mean}(\mathcal{Q})$ where $\text{Mean}(\mathcal{Q})$ is as in equation (5)
Let $\tilde{q} \leftarrow \emptyset, k \leftarrow 1$
while $\mathcal{P}$ is not empty and $k \leq \log n$ **do**
  $\varepsilon_k \leftarrow \frac{1}{2^k}, \tilde{q} \leftarrow$ any $q \in \mathcal{P}$
  Run Algorithm 1$(\mathcal{Q}, n, \varepsilon_k)$ to obtain q^k
  $\mathcal{P} \leftarrow \mathcal{P} \cap \mathcal{B}(q^k, \varepsilon_k)$
  $k \leftarrow k + 1$
return $\tilde{q}$

For a collection of distributions that have separable mean vectors, we run Algorithm 1 at countably many resolutions $\varepsilon_k = 2^{-k}$ and take a vector that lies in the intersection of ε_k-balls around the vectors returned by Algorithm 1. Let K be the value such that $\left\|q^k - q^*\right\|_\infty \leq \varepsilon_k$ for every $k \leq K$ where Algorithm 1 returns q^k for ε_k resolution. Hence, q^*(the true mean vector) is in the intersection of these balls. The algorithm selects a vector in the last non-empty intersection, thus yielding a $2\varepsilon_K$ approximation of the true mean vector. Increasing n yields finer approximations, ensuring asymptotic convergence.

Theorem 7 *If the collection of distributions $\mathcal{Q}$ has separable mean vectors, then $\mathcal{Q}$ is UME-learnable by Algorithm 2.*

Proof We prove the theorem by presenting an algorithm that returns an estimate arbitrarily close to the true underlying mean. The analysis relies on obtaining a sufficiently large training set. As we increase the size of the training set, we obtain increasingly accurate approximations of the true mean vector. From Lemma 5, we know that if a countable $\varepsilon-$cover exists, then we can find an $\varepsilon-$approximation of the true mean vector. Here, we exploit Algorithm 1 to establish UME-learnability for a separable collection of distributions.
Let $\mu^* \in \mathcal{Q}$ be the true distribution, and let $q^* = \text{Mean}(\mu^*)$. Let n denote the number of data points obtained. We define $\varepsilon_k = 2^{-k}$. We denote the countable ε_k-cover of $\text{Mean}(\mathcal{Q})$ by $\mathcal{Q}_k$, and let q^k be the estimate returned by Algorithm 1 for $\varepsilon = \varepsilon_k$.By Lemma 5, with probability 1 there exists n_k such that for every $n > n_k$, the estimator q^k satisfies $\left\|q^k - q^*\right\|_\infty \leq \varepsilon_k$.
When $n > n_k$, we say Algorithm 2 has converged for ε_k. Let K be the largest value such that for all $k \leq K$ the algorithm has converged for ε_k. Let $q \in \text{Mean}(\mathcal{Q}) \cap \bigcap_{k \leq K} \mathcal{B}(q^k, \varepsilon_k)$ be any vector in the intersection of the balls for the converged values for ε_k. Also note that, since this algorithm has converged for all $k \leq K$, the true mean vector lies in the intersection, making it non-empty. Furthermore, using the triangle inequality, we obtain

$$\|q - q^*\|_\infty \leq \left\|q - q^K\right\|_\infty + \left\|q^K - q^*\right\|_\infty \leq 2\varepsilon_K \tag{9}$$

Note that the algorithm does not necessarily stop at $k = K$; rather, it continues until the intersection of the balls around the vectors returned by Algorithm 1 becomes empty. Let $\mathscr{K}$ be the largest k such that the intersection of the balls is non-empty. The algorithm then returns $\tilde{q} \in \text{Mean}(\mathcal{Q}) \cap \bigcap_{k \leq \mathscr{K}} \mathcal{B}(q^k, \varepsilon_k)$. Since the intersection for the first K balls is non-empty, it follows that $\mathscr{K} \geq K$.

In particular, it further implies that $\tilde{q}$ is in $\text{Mean}(\mathcal{Q}) \cap \bigcap_{k \leq K} \mathcal{B}(q^k, \varepsilon_k)$. Thus, using equation (9) we conclude that

$$\|\tilde{q} - q^*\|_\infty < 2\varepsilon_K$$

Lemma 5 holds simultaneously for all $k \in \mathbb{N}$ by union bound. Hence, with probability 1 we obtain,

$$\|\tilde{q} - q^*\|_\infty \leq 2 \min \{\varepsilon_k : n > n_k\}$$

and since $n_k < \infty$ for every $k \in \mathbb{N}$, $\lim_{n\to\infty} \min \{\varepsilon_k : n > n_k\} = 0$.
And as $\|\tilde{q} - q\|_\infty \leq 1$ by the Dominated Convergence Theorem we obtain $\mathbb{E} \|\tilde{q} - q\|_\infty \xrightarrow{n\to\infty} 0$. ■

5. Examples

Section 4 shows us the sufficiency of separability in the mean as a characterization for UME-learnability. We consider proposition 1 from Cohen and Kontorovich (2023), which shows that the following collection of distributions is UME-learnable. We show it is also separable and therefore UME-learnable.

$$\mathcal{Q}_{prop} = \left\{ \mu : \mu = \text{Prod}(q) \text{ such that for all } j \in \mathbb{N}, \left| q_j - \frac{1}{2} \right| \leq \frac{c}{\sqrt{j}} \right\}$$

for a universal constant $c > 0$.

Proposition 8 *$\mathcal{Q}_{prop}$ has separable mean vectors.*

Proof We wish to show that for every $\varepsilon > 0$ we can provide a countable $\varepsilon-$cover.
Let $\varepsilon > 0$ be given. Let $j_\varepsilon = \left\lceil \frac{c^2}{\varepsilon^2} \right\rceil$ we define the $\varepsilon-$covering set $\mathcal{Q}_\varepsilon$ as follows:

$$\mathcal{Q}_\varepsilon = \left\{ p \in [0,1]^{\mathbb{N}} : p_j \in \mathbb{Q} \text{ if } j \leq j_\varepsilon \text{ and } p_j = \frac{1}{2} \text{ otherwise} \right\}$$

Note that for any vector $q \in \text{Mean}(\mathcal{Q})$ and any vector $p \in \mathcal{Q}_\varepsilon$,

$$||p - q||_\infty = \max \left\{ \max_{i \leq j_\varepsilon} |p_i - q_i|, \sup_{j > j_\varepsilon} |p_i - q_i| \right\} \leq \max \left\{ \max_{i \leq j_\varepsilon} |p_i - q_i|, \varepsilon \right\}$$

As rationals can arbitrarily approximate any real, there exists $p \in \mathcal{Q}_\varepsilon$ such that $\max_{j \leq j_\varepsilon} |p_i - q_i| < \varepsilon$. Hence $\mathcal{Q}_\varepsilon$ is an $\varepsilon-$cover of $\mathcal{Q}$, and as it consists of rational numbers for finite coordinates, it is countable. ■

The sufficiency of separability for UME-learnability leads to a natural question.

Is separability necessary for UME-learning?

We answer this in the negative. Consider the following collection of distributions

$$\mathcal{Q}_{bin} = \left\{ \mu : \text{Mean}(\mu) \in \{0,1\}^{\mathbb{N}} \right\}$$

The collection of distributions whose means are the set of all binary vectors is trivially UME-learnable. With one data point, we know the exact underlying distribution used for sampling, as the

realization will be 0 only if the mean value was 0, and it will be 1 only if the mean value was 1.
We also note that $\mathcal{Q}_{bin}$ has non-separable mean vectors. If possible, $\mathcal{Q}_{bin}$ have separable mean vectors. Let $\bar{\mathcal{Q}}$ be a countable $\frac{1}{2}-$cover of Mean$(\mathcal{Q}_{bin})$. We note that for any $q, q' \in$ Mean$(\mathcal{Q})$, if $q \in \mathcal{B}(p, \frac{1}{2})$ for some $p \in \bar{\mathcal{Q}}$ then $q' \notin \mathcal{B}(p, \frac{1}{2})$ as $\|q - q'\|_\infty = 1$. Hence, there is only one element of $\mathcal{Q}$ in every ball of radius $\frac{1}{2}$ around any $p \in \bar{\mathcal{Q}}$, making $\bar{\mathcal{Q}}$ uncountable and thus contradicting our assumption.
Although the above example provides a trivial counterexample to the necessity of separability in the mean for UME-learnability, there are non-separable collections of distributions that are not UME-learnable. For instance, consider the following collection of distributions.

$$\mathcal{Q}_{tert} = \left\{ \mu : \mu = \text{Prod}(q) \text{ where } q \in \left\{ \frac{1}{3}, \frac{2}{3} \right\}^{\mathbb{N}} \right\}$$

We note that $\mathcal{Q}_{tert}$ is also non-separable as for any $q, q' \in \mathcal{Q}_{tert}, \|q - q'\|_\infty = \frac{1}{3}$. Hence, we can apply an argument similar to the one used to show the non-separability of $\mathcal{Q}_{bin}$. We also refer to the proof of Theorem 1 in Cohen et al. (2025), which implicitly proves that $\mathcal{Q}_{tert}$ is not UME-learnable.
The collection of distributions whose means are binary vectors is a trivial example of a non-separable but UME-learnable collection of distributions. We present a non-trivial example of non-separable classes that is UME-learnable using techniques fundamentally different from those used in the separability-based analysis or methods used in the literature.
We define a collection of distributions $\mathcal{Q}_{tree}$ using their respective mean vectors. We consider a binary tree and label it using a mean vector by traversing it in a level order fashion. Formally, for every mean vector $q = (q_1, q_2, \ldots)$, the root corresponds to q_1, the left child of q_1 corresponds to q_2, the right child of q_1 corresponds to q_3, the left child of q_2 corresponds to q_4, the right child of q_2 corresponds to q_5 and so on. Because the mean vector is infinite, the binary tree has infinite depth. Finally, for any branch, i.e., a root-to-leaf path in the tree, we assign all their corresponding coordinates with the value $\frac{2}{3}$, whereas all other coordinates are given a value of $\frac{1}{3}$.
Hence, we can define $\mathcal{Q}_{tree}$ as a collection of distributions for all such mean vectors as follows

$$\mathcal{Q}_{tree} = \{\mu : \mu = \text{Prod}(q) \text{ where } q \text{ satisfies the structure given above}\}$$

We note that $\mathcal{Q}_{tree}$ has non-separable mean vectors as for any $q, q' \in \mathcal{Q}_{tree}, \|q - q'\|_\infty = \frac{1}{3}$. Hence, we can use an argument similar to the one used to show the non-separability of Mean$(\mathcal{Q}_{bin})$.

Proposition 9 *$\mathcal{Q}_{tree}$ is UME-learnable.*

Proof Sketch To learn $\mathcal{Q}_{tree}$, we leverage the tree structure embedded in the collection of mean vectors. We note that finding the true underlying mean vectors is equivalent to identifying the branch of the tree labeled $\frac{2}{3}$. We calculate the limiting average of the empirical means along every branch of the tree and return the branch for which it is exactly $\frac{2}{3}$. The algorithm works as we can show that for all branches other than the true branch, the limiting average is not $\frac{2}{3}$ uniformly. The formal proof, along with the motivating idea, is provided in Appendix A. ■.
In many learning-theory problems, the notion of a bad structure arises, and if such a structure appears, the learning problem is considered hard or not learnable. We can regard $\mathcal{Q}_{tert}$ as a bad structure for UME-learnability, but we can show that it is a substructure of another problem that is UME-learnable, as seen in the following example.

Consider $\mathcal{Q}_{round}$ defined as follows:

$$\mathcal{Q}_{round} = \left\{ \mu : \mu = \text{Prod}(q) \text{ such that } q_{2n-1} \in \left\{ \frac{1}{3}, \frac{2}{3} \right\} \text{ and } q_{2n} = \mathbb{1}\left[q_{2n-1} = \frac{2}{3} \right] \text{ for } n \in \mathbb{N} \right\}$$

Proposition 10 *$\mathcal{Q}_{round}$ is UME-learnable.*

Proof Let $\mu^* \in \mathcal{Q}_{round}$ be the underlying distribution. Let $q^* = \text{Mean}(\mu^*)$. Let $X \sim \mu^*$ be a data point. As $q_{2n} \in \{0, 1\}$ we can find them using the value of X_{2n} for every $n \in \mathbb{N}$. And as $q_{2n} = \mathbb{1}\left[q_{2n-1} = \frac{2}{3}\right]$, q_{2n-1} can be inferred using q_{2n}. ■

6. UME-learnability is closed under countable unions.

The conjecture from Cohen et al. (2025) looks at countable collections of distributions $\mathcal{Q}$ with some specific properties to ascertain the UME-learnability of LGC$\cup\mathcal{Q}$. We claim a collection of distributions $\mathcal{Q} = \cup_{i\in\mathbb{N}}\mathcal{Q}_i$ where $\mathcal{Q}_i$ is UME-learnable by algorithm $\mathcal{A}_i$ which returns an estimate $\tilde{q}^i$ is also UME-learnable using the following algorithms,

Algorithm 3 Survival Test $\left(i, \varepsilon, n, \left(\tilde{q}^1, \tilde{q}^2, \ldots, \tilde{q}^n\right), \hat{q}\right)$

Initialize wins $\leftarrow 0$
for t goes from 1 to n **do**
  if for every $j \in \mathbb{N} \left|\tilde{q}_j^t - \tilde{q}_j^i\right| \leq 4\varepsilon$ **then** wins $\leftarrow$ wins $+1$
  else
    $J = \min\{j \in \mathbb{N} : \left|\tilde{q}_j^t - \tilde{q}_j^i\right| > 4\varepsilon\}$
    if $\left|\hat{q}_J - \tilde{q}_J^i\right| < \varepsilon + \sqrt{\frac{3\log n}{n}}$ **then** wins $\leftarrow$ wins $+1$
if wins is equal to n **then** return "pass" **otherwise** return "fail"

Algorithm 4 Countable union $(\mathcal{Q}, 2n > 0, (\mathcal{A}_1, \mathcal{A}_2, \ldots))$

We split the $2n$ training data into a training set S_1 and a validation set S_2 each of size n.
$\mathcal{P} \leftarrow$Mean$(\mathcal{Q}), \tilde{q} \leftarrow \emptyset, k \leftarrow 1$
We consider the first n algorithms and run $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ on S_1 to obtain $\tilde{q}^1, \tilde{q}^2, \ldots, \tilde{q}^n$ resp.
We compute the empirical mean estimator using S_2 to obtain $\hat{q}$
while $\mathcal{P}$ is not empty **do**
  $\varepsilon \leftarrow \frac{1}{2^k}, \tilde{q} \leftarrow$ any $q \in \mathcal{P}$
  for i goes from 1 to n **do**
  if Algorithm 3 $(i, \varepsilon, n, (\tilde{q}^1, \ldots, \tilde{q}^n), \hat{q})$ returns "pass" **then**
    $\mathcal{P} \leftarrow \mathcal{P} \cap \mathcal{B}(\tilde{q}^i, 5\varepsilon)$
    $k \leftarrow k + 1$
return $\tilde{q}$

An estimator that survives Algorithm 3 will be a $5\varepsilon-$approximation of the true underlying mean vector for any $\varepsilon > 0$ with high probability for a sufficiently large amount of training data. This algorithm is used as a subroutine for the algorithm 4. Similar to Algorithm 1, Algorithm 4 focuses on the first n algorithms and, like Algorithm 2, it chains ε_k-approximations of the true underlying distribution to guarantee UME-learnability of countable unions.

Theorem 11 *UME-learnability is closed under countable unions.*

Proof Let $\mathcal{Q} = \cup_{i\in\mathbb{N}}\mathcal{Q}_i$ be our collection of distributions where $\mathcal{Q}_i$ is UME-learnable by algorithm $\mathcal{A}_i$ using the estimator $\tilde{q}^i$. Let $\mu^* \in \mathcal{Q}_{i^*}$ be our true underlying distribution. Let $q^* = \text{Mean}(\mu^*)$. We have been provided with $2n$ data points. We split the data into a training set (S_1) and a validation set (S_2), each of size n. Let $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ denote the first n algorithms and let $\tilde{q}^1, \tilde{q}^2, \ldots, \tilde{q}^n$ denote the estimates returns by running the respective algorithm on S_1. Let $\hat{q}$ denote the empirical mean estimator calculated using S_2. Algorithm 3 checks whether a particular estimator could approximate the true underlying mean vector. We say an estimator $\tilde{q}^1$ wins against another estimator $\tilde{q}^2$ if $\left\|\tilde{q}^1 - \tilde{q}^2\right\|_\infty \leq 4\varepsilon$ or $\left|\tilde{q}^1_J - \hat{q}_J\right| < \varepsilon + \sqrt{\frac{3\log n}{n}}$ where $\left|\tilde{q}^1_J - \tilde{q}^2_J\right| > 4\varepsilon$.
Let $n > i^*$ and $(\varepsilon^*)_n = \sqrt{\mathbb{E}_{S_1}\left\|\tilde{q}^{i^*} - q^*\right\|_\infty}$. For any $\varepsilon > \max\left\{(\varepsilon^*)_n, \sqrt[4]{\frac{9}{n}}\right\}$, let $I_1 = \{i \in [n] : \left\|\tilde{q}^{i^*} - \tilde{q}^i\right\|_\infty \leq 4\varepsilon\}$ and $I_2 = [n] \setminus I_1$. For any $i \in I_2$ let $j_i = \min\left\{j \in \mathbb{N} : \left|\tilde{q}^i_j - \tilde{q}^{i^*}_j\right| > 4\varepsilon\right\}$. As $\mathcal{Q}_{i^*}$ is UME-learnable by $\mathcal{A}_{i^*}$, we focus on $\tilde{q}^{i^*}$. $\tilde{q}^{i^*}$ will win against every $\tilde{q}^i$ for $i \in I_1$ by definition of I_1. Therefore, we focus on the event E_n of $\tilde{q}^{i^*}$ not winning against $\tilde{q}^i$ for some $i \in I_2$. We analyze the probability of E_n by conditioning on the event $\left\|\tilde{q}^{i^*} - q^*\right\|_\infty < \varepsilon$. We apply the union bound to focus on comparing $\tilde{q}^{i^*}$ with $\tilde{q}^i$ for $i \in \mathbb{N}$ and the triangle inequality to focus on the deviation from the true underlying mean. Next, we use the condition. We finally apply Markov inequality (Markov (1884)) and Hoeffding inequality (Hoeffding (1963)), as detailed below.

$$\mathbb{P}(E_n) \leq \mathbb{P}\left(\exists i \in I_2 : \left|\hat{q}_{j_i} - \tilde{q}^{i^*}_{j_i}\right| \geq \varepsilon + \sqrt{\frac{3\log n}{n}} \,\middle|\, \left\|\tilde{q}^{i^*} - q^*\right\|_\infty < \varepsilon\right) + \mathbb{P}\left(\left\|\tilde{q}^{i^*} - q^*\right\|_\infty \geq \varepsilon\right)$$

$$\leq \sum_{i=1}^{n} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| + \left|q^*_{j_i} - \tilde{q}^{i^*}_{j_i}\right| \geq \varepsilon + \sqrt{\frac{3\log n}{n}} \,\middle|\, \left\|\tilde{q}^{i^*} - q^*\right\|_\infty < \varepsilon\right) + \mathbb{P}\left(\left\|\tilde{q}^{i^*} - q^*\right\|_\infty \geq \varepsilon\right)$$

$$\leq n \cdot 2\exp\left(-2n\left(\sqrt{\frac{3\log n}{n}}\right)^2\right) + \frac{\mathbb{E}\left\|\tilde{q}^{i^*} - q^*\right\|_\infty}{\varepsilon} \leq \frac{2}{n^5} + (\varepsilon^*)_n \tag{10}$$

We also note that if Algorithm 3 "passes" $\tilde{q}^{i^*}$ then none of $i \in I_2$ can pass the test because for any $i \in I_2$, $\left|\tilde{q}^i_{j_i} - \hat{q}_{j_i}\right| \geq \left|\tilde{q}^i_{j_i} - \tilde{q}^{i^*}_{j_i}\right| - \left|\tilde{q}^{i^*}_{j_i} - \hat{q}_{j_i}\right| \geq 2\varepsilon$ since $n > \frac{9}{\varepsilon^4}$. $\tilde{q}^i$ for some $i \in I_1$ might survive algorithm 3. But since $\left\|\tilde{q}^i - \tilde{q}^{i^*}\right\|_\infty \leq 4\varepsilon$, with probability $1 - (\varepsilon^*)_n - \frac{2}{n^5}$ any estimator that survives algorithm 3 is a $5\varepsilon-$approximation of the true mean vector.
Algorithm 4 exploits all the $5\varepsilon_k-$approximating vectors obtained on running Algorithm 3 for the first n algorithms for $\varepsilon_k = 2^{-k}, k \in \mathbb{N}$ and obtains a vector which is a $10\varepsilon_k-$approximating vector for all $k \in \mathbb{N}$ simultaneously asymptotically as we argue as follows.
Let $K = \min\left\{\left\lfloor \frac{1}{\sqrt{(\varepsilon^*)_n + \frac{2}{n^5}}} \right\rfloor, \left\lfloor \log\left(\frac{1}{(\varepsilon^*)_n}\right) \right\rfloor, \left\lfloor \frac{1}{4}\log\frac{n}{9} \right\rfloor\right\}$ and let $\mathcal{R} = \text{Mean}(\mathcal{Q}) \cap \bigcap_{k\leq K} \mathcal{B}(q^k, 5\varepsilon_k)$ where q^k is the estimator that "passes" Algorithm 3 for $\varepsilon = \varepsilon_k$. By union bound we obtain with probability $1 - K\left((\varepsilon^*)_n + \frac{2}{n^5}\right) \geq 1 - \sqrt{(\varepsilon^*)_n + \frac{2}{n^5}}$ for every $k \in [K], \left\|q^k - q^*\right\| < 5\varepsilon_k$.
Hence with probability $1 - \sqrt{(\varepsilon^*)_n + \frac{2}{n^5}}$, $\mathcal{R}$ is non-empty as the true mean vector will be in $\mathcal{R}$.
The algorithm does not halt after the first K rounds; rather, it continues until the intersection of the balls around the vectors returned by Algorithm 3 becomes empty. Let $\mathscr{K}$ be the largest k such

that the intersection is non-empty. Let $\mathcal{T} = \text{Mean}(\mathcal{Q}) \cap \bigcap_{k \leq \mathscr{K}} \mathcal{B}(q^k, 5\varepsilon_k)$. Due to our previous argument, $\mathcal{T} \subset \mathcal{R}$. Hence, the estimate $\tilde{q}$ returned by the algorithm is in $\mathcal{R}$. Therefore, we can conclude with probability $1 - \sqrt{(\varepsilon^*)_n + \frac{2}{n^5}}$

$$\|\tilde{q} - q^*\|_\infty \leq \left\|\tilde{q} - q^K\right\|_\infty + \left\|q^K - q^*\right\|_\infty \leq 10\varepsilon_K. \tag{11}$$

We note that by definition of UME-learnability (Definition 1), $(\varepsilon^*)_n \xrightarrow{n\to\infty} 0$ and $K \xrightarrow{n\to\infty} \infty$. Therefore, by Equation (11), $\mathbb{E}\|\tilde{q} - q^*\|_\infty \leq 10\varepsilon_K + \sqrt{(\varepsilon^*)_n + \frac{2}{n^5}} \xrightarrow{n\to\infty} 0$. ■

7. Conclusion

In this paper, we discuss uniform convergence beyond the paradigm of $P-$Glivenko-Cantelli by studying more general types of estimators than the empirical mean estimator. We introduced UME-learnability to characterize when collections of distributions on $\{0,1\}^{\mathbb{N}}$ admit uniform mean estimation by arbitrary estimators. We showed that if a collection of distributions $\mathcal{Q}$ has separable mean vectors, then $\mathcal{Q}$ is UME-learnable. We further demonstrated that separability is not necessary by constructing a non-separable, tree-structured collection that is nevertheless UME-learnable via techniques distinct from the separability-based analysis. Finally, we proved that UME-learnability is closed under countable unions, thereby resolving the conjecture of Cohen et al. (2025) and extending it beyond the two-collection setting considered there.
Uniform convergence is often used in the design and analysis of algorithms for problems such as classification. One natural application of our more general estimators would be as an alternative to empirical risk minimization in those learnable problems by minimizing the estimated mean losses beyond empirical means.

8. Extensions and open problems

This work opens several natural directions for further investigation. Some partial progress on these questions is already included in the appendix, while others remain open and appear to require new ideas.

- Throughout this work, we focus on distributions indexed by a countable coordinate set. A natural extension we have studied in Appendix B is to allow an uncountable coordinate set. In Theorem 16, we show that separability of the mean space remains a sufficient condition for UME-learnability even in this more general setting.
- While we show that separability of the mean space implies UME-learnability, Proposition 9 demonstrates that this condition is not necessary. This raises the problem of identifying necessary and sufficient conditions for UME-learnability when we do not restrict the mean vectors of a collection of distributions to be separable. An especially challenging open question is to characterize UME-learnability in the non-separable regime when the coordinate set is uncountable.
- Another extension, discussed in Appendix C, is regarding uniform convergence over the function class as well as the underlying collection of distributions. When the mean vectors of a collection of distributions are totally bounded, we can provide an upper bound on the expected estimation error. An open problem is to provide a complete characterization of optimal *uniform* and *universal* rates of UME-learnability.

References

Moïse Blanchard, Doron Cohen, and Aryeh Kontorovich. Correlated Binomial Process, 2024. URL https://arxiv.org/abs/2402.07058.

Doron Cohen and Aryeh Kontorovich. Local Glivenko-Cantelli, 2023. URL https://arxiv.org/abs/2209.04054.

Doron Cohen, Aryeh Kontorovich, and Roi Weiss. The Empirical Mean is Minimax Optimal for Local Glivenko-Cantelli, 2025. URL https://arxiv.org/abs/2410.02835.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963. ISSN 1537-274X. doi: 10.1080/01621459.1963.10500830. URL http://dx.doi.org/10.1080/01621459.1963.10500830.

Andrey A. Markov. On some applications of algebraic calculus to probabilities. *Proceedings of the Kazan Physical-Mathematical Society*, 2:3–20, 1884. Originally published in Russian as "O nekotorykh prilozheniyakh algebraicheskogo ischisleniya k veroyatnostyam".

Walter Rudin. *Functional Analysis*. McGraw-Hill, New York, 2 edition, 1991. ISBN 978-0-07-054236-5.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families, 2013. URL https://arxiv.org/abs/1312.3516.

A. W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer International Publishing, 2023. ISBN 9783031290404. doi: 10.1007/978-3-031-29040-4. URL http://dx.doi.org/10.1007/978-3-031-29040-4.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL https://doi.org/10.1137/1116025.

Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer New York, 2006. ISBN 9780387342399. doi: 10.1007/0-387-34239-7. URL http://dx.doi.org/10.1007/0-387-34239-7.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications*, 26(4):532–553, 1981.

Appendix A. An interesting UME-learnable collection of distributions that has non-separable mean vectors

The UME-learnability of the collection of distributions with the binary vectors as their corresponding mean vectors provides a trivial counterexample towards separability in the mean being a necessary condition for UME-learnability of a collection of distributions. The non UME-learnability of the collection of product measures with their mean in $\left\{\frac{1}{3}, \frac{2}{3}\right\}^{\mathbb{N}}$ as proven implicitly in Theorem 1 of Cohen et al. (2025) warrants further investigation in the setting of a collection of distributions that have non-separable mean vectors.

We show that the collection of distributions with non-separable mean vectors possesses an inherent structure in which the mean vectors can be infinitely sequentially fat-shattered. More formally, consider a complete binary tree of depth d. The nodes of the tree are labeled by integers that correspond to the coordinates of the collection of mean vectors. Each node is associated with a value $r_i \in (0, 1)$. At a node with label i, the left edge indicates the value of the mean vector at the i^{th} coordinate is less than or equal to $r_i - \gamma$, whereas the right edge indicates the value is greater than or equal to $r_i + \gamma$ for some $\gamma > 0$. We say a tree of depth d is shattered by the mean vectors if for every branch in the tree, there exists a mean vector that follows the path as set by the nodes in the branch (For example, refer to Fig. 1). We say the mean vectors are infinitely shattered if for every $d \in \mathbb{N}$ there exists a tree of depth d that is shattered by the mean vectors.

Theorem 12 *If a collection of distributions $\mathcal{Q}$ has non-separable mean vectors, then there exists $\gamma > 0$ such that the mean vectors Mean$(\mathcal{Q})$ can be infinitely sequentially fat-shattered.*

To prove Theorem 12, we first develop the necessary machinery. As $\mathcal{Q}$ has a non-separable mean vectors, there exists $\gamma > 0$ such that $\mathcal{Q}$ does not have a countable $3\gamma-$cover for its means. We will use this γ to show that Mean$(\mathcal{Q})$ is infinitely sequentially fat-shattered. We will also use lemma 13 which shows that there exists a coordinate i for which there are two subsets of $\mathcal{Q}$ such that for one of the collection at coordinate i the mean value is less than or equal to $r_i - \gamma$, another collection for which the mean value is more than or equal to $r_i + \gamma$ for some $r_i \in (0, 1)$ and the mean vectors of these two collections do not possess a countable $3\gamma-$cover.

Lemma 13 *For a collection of distributions $\mathcal{Q}$ that does not possess a countable $3\gamma-$ cover for its mean, there exists a coordinate i and a value $r_i \in (0, 1)$ for which there exist two collections $\mathcal{Q}_1, \mathcal{Q}_2$ that also do not possess a countable $3\gamma-$cover for their means and if $q \in$ Mean$(\mathcal{Q}_1)$ then $q_i \geq r_i + \gamma$ and if $q \in$ Mean$(\mathcal{Q}_2)$ then $q_i \leq r_i - \gamma$.*

Proof We define an operation *subset selection* across a coordinate i with a value r_i for deviation γ performed on the collection of distributions $\mathcal{Q}$ in which we create two collections of distributions $\mathcal{Q}_{1,i}$ and $\mathcal{Q}_{2,i}$ such that for any $q \in \text{Mean}(\mathcal{Q}_{1,i})$, $q_i \leq r_i - \gamma$ and for any $q \in \text{Mean}(\mathcal{Q}_{2,i})$, $q_i \geq r_i + \gamma$. We first consider the case in which, when we perform *subset selection* for deviation γ across all coordinates i, there is some value r_i that produces two collections which possess a countable $3\gamma-$cover. Let $\mathcal{Q}_{1,i}, \mathcal{Q}_{2,i}$ denote the collections obtained after *subset selection* across coordinate i with value r_i for deviation γ. Let $\bar{\mathcal{Q}}_{1,i}, \bar{\mathcal{Q}}_{2,i}$ denote their respective countable $3\gamma-$covers. Let $\mathcal{Q}_1 = \cup_{i\in\mathbb{N}}\mathcal{Q}_{1,i}$ and $\mathcal{Q}_2 = \cup_{i\in\mathbb{N}}\mathcal{Q}_{2,i}$. We note that $\bar{\mathcal{Q}}_1 = \cup_{i\in\mathbb{N}}\bar{\mathcal{Q}}_{1,i}$ is a countable $3\gamma-$cover for Mean$(\mathcal{Q}_1)$ and $\bar{\mathcal{Q}}_2 = \cup_{i\in\mathbb{N}}\bar{\mathcal{Q}}_{2,i}$ is a countable $3\gamma-$cover for Mean$(\mathcal{Q}_2)$. If $\mu \in \mathcal{Q}$ does not belong to either $\mathcal{Q}_1$ or $\mathcal{Q}_2$ then due to our *subset selection* operation for every $i \in \mathbb{N}, |q_i - r_i| < \gamma$ where $q = \text{Mean}(\mu)$. Hence $\bar{\mathcal{Q}}_1 \cup \bar{\mathcal{Q}}_2 \cup \{(r_1, r_2, \ldots)\}$ is a countable $3\gamma-$cover of $\mathcal{Q}$ contradicting our assumption that

Mean($\mathcal{Q}$) does not possess a countable $3\gamma-$cover.
We now consider the case in which for all coordinates i with any $r_i \in (0,1)$, at most one of the collections obtained after *subset selection* for deviation γ does not possess a countable $3\gamma-$cover. Let the collections obtained after *subset selection* across coordinate i with $r_i = \frac{1}{2}$ for deviation γ be $\mathcal{Q}_{1,i}, \mathcal{Q}_{2,i}$. Without loss of generality, let for all $i \in \mathbb{N}$, Mean($\mathcal{Q}_{1,i}$) not possess a countable $3\gamma-$cover whereas Mean($\mathcal{Q}_{2,i}$) possess a countable $3\gamma-$cover labeled as $\bar{\mathcal{Q}}_{2,i}$. Let $\mathcal{Q}_2^1 = \cup_{i\in\mathbb{N}}\mathcal{Q}_{2,i}$ and $\bar{\mathcal{Q}}_2^1 = \cup_{i\in\mathbb{N}}\bar{\mathcal{Q}}_{2,i}$. We note that $\bar{\mathcal{Q}}_2^1$ is a countable $3\gamma-$cover of $\mathcal{Q}_2^1$. We note that for any $\mu \in \mathcal{Q}_2^1, q_i \geq \frac{1}{2}+\gamma$ for every $i \in \mathbb{N}$ where $q = \text{Mean}(\mu)$. Hence for any $\mu \notin \mathcal{Q}_2^1, q_i < \frac{1}{2}+\gamma$ for every $i \in \mathbb{N}$ where $q = \text{Mean}(\mu)$.
We can now similarly repeat the produce of performing *subset selection* on $\mathcal{Q}_1^1$ for every coordinate $i \in \mathbb{N}$ with $r_i = \frac{1}{2}\left(\frac{1}{2}+\gamma\right)$ and deviation γ and create $\mathcal{Q}_1^2$ and $\mathcal{Q}_2^2$ where Mean($\mathcal{Q}_1^2$) does not possess a countable $3\gamma-$cover and Mean($\mathcal{Q}_2^2$) has a countable $3\gamma-$cover $\bar{\mathcal{Q}}_2^2$.
We recursively repeat the procedure for $K = \left\lceil \log_2\left(\frac{1-2\gamma}{\gamma}\right)\right\rceil$ iterations to obtain K countable covers $\bar{\mathcal{Q}}_2^1, \bar{\mathcal{Q}}_2^2, \ldots, \bar{\mathcal{Q}}_2^K$ for Mean($\mathcal{Q}_2^1$), Mean($\mathcal{Q}_2^2$), $\ldots$, Mean($\mathcal{Q}_2^K$) respectively and $\mathcal{Q}_1^K$ such that for any $\mu \in \mathcal{Q}_1^K, q_i < \frac{1}{2^K} + \left(1+\frac{1}{2}+\ldots+\frac{1}{2^{K-1}}\right)\gamma \leq 3\gamma$ where $q = \text{Mean}(\mu)$. Consequently, Mean($\mathcal{Q}_1^K$) can be covered by $(0,0,\ldots)$. Due to our application of the *subset selection* procedure recursively, we obtain $\mathcal{Q}_1^K \cup \bigcup_{k\in[K]}\mathcal{Q}_2^k = \mathcal{Q}$. Hence $\cup_{k\in[K]}\bar{\mathcal{Q}}_2^k \cup (0,0,\ldots)$ is a countable $3\gamma-$cover for Mean($\mathcal{Q}$) which contradiction our assumption. ■

Proof (Theorem 12) We show that Mean($\mathcal{Q}$) is infinitely sequentially fat-shattered by providing an infinite-depth tree such that each branch is realized by some mean vector in Mean($\mathcal{Q}$). As previously argued, as $\mathcal{Q}$ has non-separable mean vectors, there exists $\gamma > 0$ such that there does not exist a countable $3\gamma-$cover.
We can build the tree recursively. At the root, Lemma 13 provides a coordinate i_1 that splits $\mathcal{Q}$ into two collections of distributions that do not possess a countable $3\gamma-$cover. These subsets constitute the collections used to construct the left and right subtrees of the tree. We now consider the left and right subtrees separately. As the collections of distributions do not possess a countable $3\gamma-$cover, we can repeat the previous step. Therefore, we can use Lemma 13 at every depth of the tree, hence obtaining an infinitely fat shattered tree. ■

This inherent structure of infinite fat shattering of the mean vectors of a collection of distributions that have non-separable mean vectors produces an interesting example. We define a collection of distributions $\mathcal{Q}_{tree}$ using their respective mean vectors. We consider a binary tree and label it using a mean vector by traversing it in a level order fashion. Formally, for every mean vector $q = (q_1, q_2, \ldots)$, the root corresponds to q_1, the left child of q_1 corresponds to q_2, the right child of q_1 corresponds to q_3, the left child of q_2 corresponds to q_4, the right child of q_2 corresponds to q_5 and so on. As our mean vector is infinite, the binary tree has an infinite depth. Finally, for any branch, i.e., a root-to-leaf path in the tree, we assign all their corresponding coordinates with the value $\frac{2}{3}$, whereas all other coordinates are given a value of $\frac{1}{3}$.
Hence, we can define $\mathcal{Q}_{tree}$ as a collection of distributions for all such mean vectors as follows

$$\mathcal{Q}_{tree} = \{\mu : \mu = \text{Prod}(q) \text{ where } q \text{ satisfies the structure given above}\}$$

We note that $\mathcal{Q}_{tree}$ has non-separable mean vectors as for any $q, q' \in \mathcal{Q}_{tree}, \|q-q'\|_\infty = \frac{1}{3}$. Hence we can use an argument similar to the one used to show the non-separability of Mean($\mathcal{Q}_{bin}$).

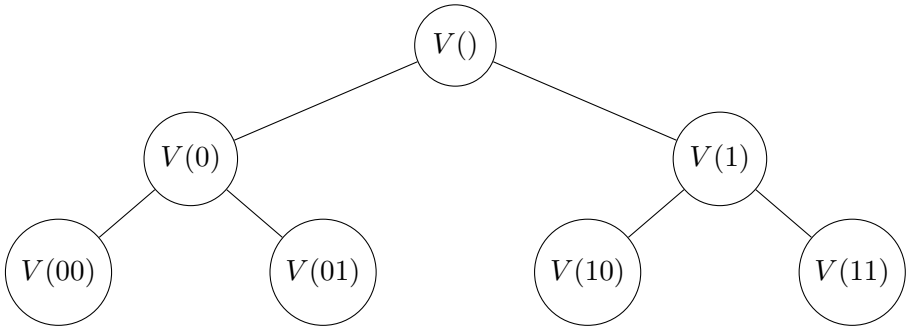

Figure 1: Labeling for tree of depth 2

We will now establish the notation for demonstrating that $\mathcal{Q}_{tree}$ is UME-learnable.
For $\mathcal{Q}_{tree}$, we will use a tree-specific notation. The root is labeled as $V()$. A branch is identified using a bit string where 0 indicates a left node and 1 indicates a right node. At depth d, we consider a $d-$dimensional binary vector $(b_1, b_2, \ldots, b_d)$ which provides the root-to-node path. This node is labeled as $V(b_1, b_2, \ldots, b_d)$. For example, for depth 2, we have the labeling according to Figure 1.
To UME-learn $\mathcal{Q}_{tree}$, we note that finding the underlying distribution is equivalent to finding the branch labeled with $\frac{2}{3}$. We will refer to this branch as the true branch. We consider the following algorithm,

Algorithm 5 Tree$(\mathcal{Q}_{tree}, n)$

$\forall b \in \{0,1\}^{\mathbb{N}}$ compute $\phi(b) = \liminf_{d\to\infty} \frac{1}{d}\sum_{j=1}^{d} \frac{1}{n}\sum_{i=1}^{n} X^{(i)}_{V(b_1,\ldots,b_j)}$
Return $\tilde{b}$ such that $\phi(\tilde{b}) = \frac{2}{3}$

The algorithm computes the limiting average of the empirical mean using the n data points available along every branch of the tree. We will show that the value converges to $\frac{2}{3}$ only for the true branch, whereas the value uniformly converges to a value other than $\frac{2}{3}$ for all other branches.

Proposition 14 *$\mathcal{Q}_{tree}$ is UME-learnable by Algorithm 5.*

Proof We note that due to our construction of $\mathcal{Q}_{tree}$, finding the true underlying distribution is equivalent to finding the branch that has been labeled as $\frac{2}{3}$. Consequently, we refer to this as the true branch and denote it by b^*. Also note that if a parent node is labeled $\frac{1}{3}$, then the child node must be labeled $\frac{1}{3}$ on any branch. This provides us with the core idea of the algorithm. We use this temporal relation to devise the test

$$\phi(b) = \liminf_{d\to\infty} \frac{1}{d}\sum_{j=1}^{d} \frac{1}{n}\sum_{i=1}^{n} X^{(i)}_{V(b_1,\ldots,b_j)}$$

We note that for the true branch $b^*, \phi(b^*) = \frac{2}{3}$ by the law of large numbers. Algorithm 5 will UME-learn $\mathcal{Q}_{tree}$ if with probability 1, for all $b \neq b^*, \phi(b) \neq \frac{2}{3}$ which can be equivalently proven if

$$\sup_b |\phi(b) - \mathbb{E}\phi(b)| = \sup_b \lim_{d\to\infty} \left| \frac{1}{d}\sum_{j=1}^{d} \frac{1}{n}\sum_{i=1}^{n} \left(X^{(i)}_{V(b_1,\ldots,b_j)} - q_{V(b_1,\ldots,b_j)} \right) \right| = 0$$

We consider the tree up to depth d and analyze the partial tree. We can apply a union bound over the 2^d branches of the partial tree. As these variables are independent (not identically distributed) random variables, we can use Hoeffding inequality (Hoeffding (1963)) to obtain,

$$\mathbb{P}\left(\sup_b \left|\frac{1}{d}\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)}_{V(b_1,\ldots,b_j)} - q_{V(b_1,\ldots,b_j)}\right)\right| > \frac{1}{\sqrt{n}}\right) \leq 2^d \cdot 2e^{-2dn\left(\frac{1}{\sqrt{n}}\right)^2} < 2^{-cd} \quad (12)$$

We define event E_d as obtaining a deviation of more than $\frac{1}{\sqrt{n}}$ between the estimated mean of the partial branch at depth d and its true mean. By equation (12) we know that $\mathbb{P}(E_d) < 2^{-cd}$. We note that $\sum_{d=0}^{\infty}\mathbb{P}(E_d) < \sum_{d=0}^{\infty} 2^{-cd} = \frac{1}{1-2^{-c}} < \infty$ Hence, by the First Borel-Cantelli lemma, we obtain that with probability $1, \exists d_0 < \infty$ such that

$$\text{For any } d > d_0,\ \sup_b \left|\frac{1}{d}\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)}_{V(b_1,\ldots,b_j)} - q_{V(b_1,\ldots,b_j)}\right)\right| < \frac{1}{\sqrt{n}} \quad (13)$$

i.e., every sufficiently large depth has deviations that are at most $\frac{1}{\sqrt{n}}$. Hence, using equation (13) we obtain

$$\lim_{d\to\infty}\sup_b \left|\frac{1}{d}\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)}_{V(b_1,\ldots,b_j)} - q_{V(b_1,\ldots,b_j)}\right)\right| < \frac{1}{\sqrt{n}}$$

We can further apply Fatou's Lemma (Rudin (1991)) to obtain

$$\sup_b \lim_{d\to\infty} \left|\frac{1}{d}\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}\left(X^{(i)}_{V(b_1,\ldots,b_j)} - q_{V(b_1,\ldots,b_j)}\right)\right| < \frac{1}{\sqrt{n}}$$

For our specific example of $\mathcal{Q}_{tree}$, for all non-true branches, the limiting average of the means along a branch is $\frac{1}{3}$. Therefore, we obtain,

$$\text{For any } b \neq b^*, \phi(b) \leq \frac{1}{3} + \frac{1}{\sqrt{n}}$$

So, if $n \geq 36$,

$$\text{For any } b \neq b^*, \phi(b) \leq \frac{1}{2}$$

Hence, with probability 1, for all $b \neq b^*, \phi(b) \neq \frac{2}{3}$ and for the true branch $\phi(b^*) = \frac{2}{3}$. ■

Appendix B. UME-learnability for uncountable coordinate sets

Our work is motivated by the framework adopted in Cohen and Kontorovich (2023), which considers the $P-$Glivenko-Cantelli setting for a countable coordinate set. In section 4, we show that collections of distributions that have separable mean vectors are UME-learnable. The technique used in the algorithm to claim UME-learnability (Algorithm 2) is to eliminate candidates of the $\varepsilon-$approximation of the mean vector. Whenever a candidate vector deviates excessively from the empirical mean of the first n coordinates, we eliminate it. This approach fundamentally relies on

the ability to inspect finitely many coordinates and therefore does not directly extend to uncountable coordinate sets. To overcome this obstacle, we revisit the strategy developed in Section 6, where we are able to eliminate a candidate estimator using a *single* informative coordinate. We leverage this idea to show that separability of the mean space remains sufficient for UME-learnability even when the coordinate set is uncountable.

Throughout this section, we assume that all measure-theoretic subtleties can be resolved. We begin by defining an oracle that compares two mean vectors under the ℓ_∞ norm. Given two vectors q^1, q^2 and a tolerance ε, the oracle either certifies that the vectors are $\varepsilon-$close or returns a coordinate on which they differ by more than ε.

Algorithm 6 ℓ_∞-oracle(q^1, q^2, ε)

if $\left\|q^1 - q^2\right\|_\infty < \varepsilon$ **then return** 'close'
else return $J \in \{j : \left|q_j^1 - q_j^2\right| > \varepsilon\}$

Using the oracle in Algorithm 6, we modify the survival test as seen in Algorithm 3. Given a countable $\varepsilon-$cover of the mean space, we conduct a 1-vs-n tournament among the first n candidate vectors. A candidate that wins against all others is declared the winner, yielding an $\varepsilon-$approximation of the true mean vector. Crucially, the oracle allows us to select which coordinate is tested, thereby extending UME-learnability to uncountable coordinate sets.

Algorithm 7 Modified $\varepsilon-$approximate$(\mathcal{Q}, n, \varepsilon)$

Initialize $\mathcal{Q}_\varepsilon = \{q^1, q^2, \ldots\}$ as the countable $\varepsilon-$cover of Mean$(\mathcal{Q})$
Let $\hat{q}$ be the empirical mean computed using the training data.
for s goes from 1 to n **do**
  Let wins $\leftarrow 0$
  for t goes from 1 to n **do**
    Let $J = \ell_\infty-$oracle$\left(q^s, q^t, 4\varepsilon\right)$
    if J is 'close' **then** wins $\leftarrow$ wins $+1$
    else if $|\hat{q}_J - q_J^s| < \varepsilon + \sqrt{\frac{3\log n}{n}}$ **then** wins $\leftarrow$ wins $+1$
  if wins is equal to n **then** return q^s

Lemma 15 *If collection of distributions $\mathcal{Q}$ with an uncountable coordinate set has a countable $\varepsilon-$cover for its mean then for any $\mu \in \mathcal{Q}$, with probability 1 there exists a data size n_0 such that for all $n > n_0$ the estimator $\tilde{q}$ returned by Algorithm 7 satisfies*

$$\|\tilde{q} - q\|_\infty \leq 5\varepsilon$$

where $q =$Mean$(\mathcal{Q})$

Proof Let a collection of distributions $\mathcal{Q}$ and $\varepsilon > 0$ be given. Let $\mathcal{Q}_\varepsilon$ be a countable $\varepsilon-$cover of Mean$(\mathcal{Q})$ under the ℓ_∞ norm. Let μ^* be the true underlying distribution and let $q^* =$Mean(μ^*). Let $q^{i_\varepsilon^*}$ be vector in $\mathcal{Q}_\varepsilon$ such that $\left\|q^* - q^{i_\varepsilon^*}\right\|_\infty \leq \varepsilon$. We refer to $q^{i_\varepsilon^*}$ as the $\varepsilon-$approximating vector. Let $n > i_\varepsilon^*$. Let $I_1 = \{i \in [n] : \left\|q^{i_\varepsilon^*} - q^i\right\|_\infty \leq 4\varepsilon\}$ and $I_2 = [n] \setminus I_1$. For any $i \in I_2$ let $j_i \in \{j : \left|q_j^{i_\varepsilon^*} - q_j^i\right| > 4\varepsilon\}$.

Let event E_n denote $q^{i^*_\varepsilon}$ failing the tournament against any of the n other vectors. By definition, $q^{i^*_\varepsilon}$ will wins against any vector q^i such that $i \in I_1$. Hence, we focus on winning against q^i such that $i \in I_2$. To analyze the probability of E_n, we use union bound and triangle inequality to obtain

$$\mathbb{P}\left(\exists i \in I_2 : \left|\hat{q}_{j_i} - q^{i^*_\varepsilon}_{j_i}\right| > \varepsilon + \sqrt{\frac{3\log n}{n}}\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| + \left|q^*_{j_i} - q^{i^*_\varepsilon}_{j_i}\right| > \varepsilon + \sqrt{\frac{3\log n}{n}}\right) \tag{14}$$

We further use the fact that $\left\|q^* - q^{i^*_\varepsilon}\right\|_\infty < \varepsilon$ to obtain

$$\sum_{i=1}^{n} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| + \left|q^*_{j_i} - q^{i^*_\varepsilon}_{j_i}\right| > \varepsilon + \sqrt{\frac{3\log n}{n}}\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| > \sqrt{\frac{3\log n}{n}}\right) \tag{15}$$

Applying Hoeffding's inequality we get,

$$\sum_{i=1}^{n} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| > \sqrt{\frac{3\log n}{n}}\right) \leq n \cdot 2e^{-2n\left(\sqrt{\frac{3\log n}{n}}\right)^2} = \frac{2}{n^5}. \tag{16}$$

Let $\tilde{q}$ denote the vector returned by Algorithm 7. A vector q^i such that $i \in I_1$ could also win the tournament. By our previous analysis with probability at least $1 - \frac{2}{n^5}$, the index of $\tilde{q}$ is in I_1. But as $\left\|q^i - q^{i^*_\varepsilon}\right\|_\infty < 4\varepsilon$, therefore by our previous analysis with probability at least $1 - \frac{2}{n^5}$ $\tilde{q}$ will be $5\varepsilon-$approximation of the true underlying mean vector.
We note that $\sum_{n=1}^{\infty} \mathbb{P}(E_n) \leq \sum_{n=1}^{\infty} \frac{2}{n^5} < \infty$. Hence, we can apply the First Borel-Cantelli Lemma to conclude that with probability 1 there exists $n_0 > 0$ such that for all $n > n_0$ the algorithm successfully finds a $5\varepsilon-$approximating vector. ■

We now modify Algorithm 2 by using Algorithm 7 instead of Algorithm 1, thereby extending UME-learnability to a collection of distributions indexed by an uncountable set.

Algorithm 8 Modified Separable $(\mathcal{Q}, n > 0)$

Initialize $\mathcal{P} \leftarrow \text{Mean}(\mathcal{Q})$ where $\text{Mean}(\mathcal{Q})$ is as in equation (5)
$\tilde{q} \leftarrow \emptyset, k \leftarrow 1$
while $\mathcal{P}$ is not empty **do**
  $\varepsilon_k \leftarrow \frac{1}{2^k}, \tilde{q} \leftarrow$ any $q \in \mathcal{P}$
  Run Algorithm 7$(\mathcal{Q}, n, \varepsilon_k)$ to obtain q^k
  $\mathcal{P} \leftarrow \mathcal{P} \cap \mathcal{B}(q^k, \varepsilon_k)$
  $k \leftarrow k + 1$
return $\tilde{q}$

Theorem 16 *If the collection of distributions with an uncountable coordinate set $\mathcal{Q}$ has separable mean vectors, then $\mathcal{Q}$ is UME-learnable by Algorithm 8.*

The proof is similar to the proof of Theorem 7. We use Lemma 15 instead of Lemma 5.

Appendix C. Uniform UME-learnability

In our work, we focus on uniform convergence over a function class and not over the collection of distributions, and we analyze UME-learnability asymptotically. In this section, we show that if the mean vectors of a collection of distributions are totally bounded, then we can provide non-asymptotic bounds on the expected loss using algorithm 9. We say the mean vectors of a collection of distributions $\mathcal{Q}$ are totally bounded if for every $\varepsilon > 0$ there exists a *finite* $\varepsilon-$cover for Mean$(\mathcal{Q})$.

Algorithm 9 Totally Bounded $\varepsilon-$approximate$(\mathcal{Q}_{TB}, n)$

Let N be the $\varepsilon-$covering number for Mean$(\mathcal{Q}_{TB})$ under the ℓ_∞ norm.
Let $\mathcal{Q}_\varepsilon = \{q^1, q^2, \ldots, q^N\}$ as the countable $\varepsilon-$cover of Mean$(\mathcal{Q}_{TB})$
Let $\hat{q}$ be the empirical mean computed using the training data.
for s goes from 1 to N **do**
  Let wins $\leftarrow 0$
  for t goes from 1 to N **do**
    if for every $j \in \mathbb{N} \left|q_j^s - q_j^t\right| \leq 4\varepsilon$ **then** wins $\leftarrow$ wins $+1$
    else
      $J = \min\{j \in \mathbb{N} : \left|q_j^s - q_j^t\right| > 4\varepsilon\}$
      if $|\hat{q}_J - q_J^s| < 2\varepsilon$ **then** wins $\leftarrow$ wins $+1$
  if wins is equal to n **then** return q^s

Algorithm 9 is a modification of Algorithm 3 in which for every $\varepsilon > 0$ as the $\varepsilon-$cover is *finite* we can find a $5\varepsilon-$approximation of the true underlying distribution with probability at least $1 - 2\varepsilon$ by comparing all the vectors against each other after obtaining a sufficiently large amount of data points.

Theorem 17 *Let $\mathcal{Q}_{TB}$ be a collection of distributions such that Mean$(\mathcal{Q}_{TB})$ is totally bounded. Let $N(\varepsilon)$ denote the $\varepsilon-$covering number of Mean$(\mathcal{Q}_{TB})$. $\mathcal{Q}_{TB}$ is UME-learnable using Algorithm 9 such that for every $\mu \in \mathcal{Q}_{TB}$,*

$$\mathbb{E}_{S\sim\mu^n} \|\tilde{q} - q\|_\infty \leq 7 \inf_{\varepsilon>0} \left\{\varepsilon : n > \frac{1}{2\varepsilon^2} \log\left(\frac{N(\varepsilon)}{\varepsilon}\right)\right\}$$

where q =Mean$(\mathcal{Q})$ and $\tilde{q}$ = Totally Bounded $\varepsilon-$approximate$(\mathcal{Q}_{TB}, n)$

Proof Let $\mathcal{Q}_{TB}$ be a collection of distributions such that Mean$(\mathcal{Q}_{TB})$ is totally bounded. Let $\mathcal{Q}_\varepsilon$ be a finite $\varepsilon-$cover of Mean$(\mathcal{Q}_{TB})$ under the ℓ_∞ norm. Let N denote the $\varepsilon-$covering number of Mean$(\mathcal{Q}_{TB})$. Let μ^* be the true underlying distribution and let q^* =Mean(μ^*). Let $q^{i_\varepsilon^*}$ be vector in $\mathcal{Q}_\varepsilon$ such that $\left\|q^* - q^{i_\varepsilon^*}\right\|_\infty \leq \varepsilon$. We refer to $q^{i_\varepsilon^*}$ as the $\varepsilon-$approximating vector.
Let $n > \frac{1}{2\varepsilon^2} \log\left(\frac{N}{\varepsilon}\right)$. Let $I_1 = \{i \in [N] : \left\|q^{i_\varepsilon^*} - q^i\right\|_\infty \leq 4\varepsilon\}$ and $I_2 = [N] \setminus I_1$. For any $i \in I_2$ let $j_i = \min\{j \in \mathbb{N} : \left|q_j^{i_\varepsilon^*} - q_j^i\right| > 4\varepsilon\}$.
We analyze the probability that $q^{i_\varepsilon^*}$ loses a comparison against some $q^i \in \mathcal{Q}_\varepsilon$. By using union bound and triangle inequality, we obtain

$$\mathbb{P}\left(\exists i \in I_2 : \left|\hat{q}_{j_i} - q_{j_i}^{i_\varepsilon^*}\right| > 2\varepsilon\right) \leq \sum_{i=1}^{N} \mathbb{P}\left(\left|\hat{q}_{j_i} - q_{j_i}^*\right| + \left|q_{j_i}^* - q_{j_i}^{i_\varepsilon^*}\right| > 2\varepsilon\right) \tag{17}$$

We further use the fact that $\left\|q^* - q^{i^*_\varepsilon}\right\|_\infty < \varepsilon$ to obtain

$$\sum_{i=1}^{N} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| + \left|q^*_{j_i} - q^{i^*_\varepsilon}_{j_i}\right| > 2\varepsilon\right) \leq \sum_{i=1}^{N} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| > \varepsilon\right) \tag{18}$$

We further apply Hoeffding inequality (Hoeffding (1963)),

$$\sum_{i=1}^{N} \mathbb{P}\left(\left|\hat{q}_{j_i} - q^*_{j_i}\right| > \varepsilon\right) \leq N \cdot 2e^{-2n\varepsilon^2} \leq 2Ne^{-2\frac{1}{2\varepsilon^2}\log\left(\frac{N}{\varepsilon}\right)\varepsilon^2} = 2\varepsilon \tag{19}$$

Let $\tilde{q}$ be the vector returned by running Algorithm 9 on $\mathcal{Q}_{TB}$ using sufficiently large amount of training data $\left(\text{i.e. } n \geq \frac{1}{2\varepsilon^2}\log\left(\frac{N}{\varepsilon}\right)\right)$. We note that with probability at least $1 - 2\varepsilon$, any vector in I_1 could have been returned by the algorithm. Thus, the algorithm will return a $5\varepsilon-$approximation of the true underlying mean vector with probability of error at most 2ε.
Therefore, we note that

$$\mathbb{E}\left\|\tilde{q} - q\right\|_\infty \leq 5\varepsilon \cdot \mathbb{P}\left(\left\|\tilde{q} - q\right\|_\infty \leq 5\varepsilon\right) + 1 \cdot \mathbb{P}(\left\|\tilde{q} - q\right\|_\infty > 5\varepsilon) \leq 7\varepsilon$$

Therefore if we have been provided with n data points, we can optimize for ε to obtain

$$\mathbb{E}\left\|\tilde{q} - q\right\|_\infty \leq 7 \inf_{\varepsilon > 0}\left\{\varepsilon : n > \frac{1}{2\varepsilon^2}\log\left(\frac{N(\varepsilon)}{\varepsilon}\right)\right\}$$

■