Uniform Laws of Large Numbers in Product Spaces

Ron Holzman HOLZMAN@TECHNION.AC.IL
Department of Mathematics, Technion

Shay Moran SMORAN@TECHNION.AC.IL
Departments of Mathematics, Computer Science, and Data and Decision Sciences, Technion and Google Research

Alexander Shlimovich ASHLIMOVICH@CAMPUS.TECHNION.AC.IL
Department of Mathematics, Technion

Editors: Steve Hanneke and Tor Lattimore

Abstract

Uniform laws of large numbers form a cornerstone of Vapnik–Chervonenkis theory, where they are characterized by the finiteness of the VC dimension. In this work, we study uniform convergence phenomena in *cartesian product spaces*, under assumptions on the underlying distribution that are compatible with the product structure. Specifically, we assume that the distribution is absolutely continuous with respect to the product of its marginals, a condition that captures many natural settings, including product distributions, sparse mixtures of product distributions, distributions with low mutual information, and more.

We show that, under this assumption, a uniform law of large numbers holds for a family of events if and only if the *linear VC dimension* of the family is finite. The linear VC dimension is defined as the maximum size of a shattered set that lies on an *axis-parallel line*, namely, a set of vectors that agree on all but at most one coordinate. This dimension is always at most the classical VC dimension, yet it can be arbitrarily smaller. For instance, the family of convex sets in $\mathbb{R}^d$ has linear VC dimension 2, while its VC dimension is infinite already for $d \geq 2$. Our proofs rely on an estimator that departs substantially from the standard empirical mean estimator and exhibits a more intricate structure. We show that such deviations from the standard empirical mean estimator are unavoidable in this setting. Throughout the paper, we propose several open questions, with a particular focus on quantitative sample complexity bounds.

Keywords: uniform convergence, VC Dimension, product spaces, distribution learning.

1. Introduction

A central theme in statistics and learning theory is to understand when empirical averages provide reliable *uniform* approximations to their population counterparts. Such uniform laws of large numbers lie at the heart of statistical learning theory: they underlie generalization guarantees, sample complexity bounds, and the analysis of learning algorithms. Classical results show that, for families of events, uniform convergence is characterized by the Vapnik–Chervonenkis (VC) dimension Vapnik and Chervonenkis (1971), while for real-valued function classes it is governed by complexity measures such as Rademacher averages or fat-shattering dimensions Kearns and Schapire (1994); Alon et al. (1997); Bartlett and Long (1998). These results play a basic role in learning theory and in the study of generalization.

Much of the classical theory treats the underlying distribution as arbitrary and makes no use of additional structure of the domain. In many applications, however, the domain naturally carries a product structure, with data points represented as vectors of features or measurements. In such

settings, it is often reasonable to restrict attention to distributions that are compatible with this structure, for instance through independence or weak dependence across coordinates. The focus of this work is to understand how such product structure can be leveraged in the uniform estimation problem, and to identify complexity measures that govern uniform convergence under these structural assumptions.

1.1. Uniform Estimation

We begin by presenting an abstract *uniform estimation* problem. This framework captures a broad range of questions in statistics and learning theory and provides a common language for classical uniform laws of large numbers as well as for the results developed in this work.

Uniform Estimability

Let $\mathcal{X}$ be a domain, let $\mathcal{P}$ be a family of distributions over $\mathcal{X}$, and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Under what conditions does there exist an estimator which, given i.i.d. samples from an unknown distribution $P \in \mathcal{P}$, approximates the expectations

$$P(F) := \mathbb{E}_{X\sim P}[F(X)]$$

uniformly and simultaneously for all $F \in \mathcal{F}$?

For simplicity, and in line with much of the learning-theoretic literature, we focus on the case where $\mathcal{F}$ consists of indicator functions of measurable subsets of $\mathcal{X}$. Thus, each $F \in \mathcal{F}$ corresponds to an event, and the quantity of interest is the probability $P(F)$. This setting encompasses the classical uniform convergence problem studied in Vapnik–Chervonenkis theory. While our presentation focuses on events, the ideas and techniques developed in this paper extend naturally to more general classes of real-valued functions.

From a statistical perspective, the goal is to estimate these probabilities from finitely many independent samples. An estimation algorithm $\mathcal{A}$ is therefore a mapping $\mathcal{A} : \bigcup_{m=1}^{\infty} \mathcal{X}^m \to [0,1]^{\mathcal{F}}$.

Given a sample $x_1, \dots, x_m \in \mathcal{X}$ drawn independently from an unknown distribution $P \in \mathcal{P}$, the algorithm outputs estimates $\widehat{P}_m(F) \in [0,1]$ for the probabilities $P(F)$, simultaneously for all $F \in \mathcal{F}$. A canonical example is the empirical estimator, which assigns $\widehat{P}_m(F) := \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[x_i \in F]$. When it is possible to approximate the probabilities of all events in $\mathcal{F}$ uniformly over all target distributions in $\mathcal{P}$, we say that the pair $(\mathcal{F}, \mathcal{P})$ is *uniformly estimable*.

Definition 1 (Uniform Estimability) *We say that a pair $(\mathcal{F}, \mathcal{P})$ is* uniformly estimable *if there exist a sample complexity bound $m : (0,1)^2 \to \mathbb{N}$ and an algorithm $\mathcal{A}$ such that for every $\varepsilon, \delta \in (0,1)$ and every distribution $P \in \mathcal{P}$, the following holds: if $x_1, \dots, x_m$ are drawn independently from P for $m \geq m(\varepsilon, \delta)$, then with probability at least $1-\delta$,*

$$\forall F \in \mathcal{F} : \quad \left|\widehat{P}_m(F) - P(F)\right| \leq \varepsilon,$$

where $\widehat{P}_m = \mathcal{A}(x_1, \dots, x_m)$ denotes the estimator output by $\mathcal{A}$.

Before presenting our main results, we illustrate the scope of the uniform estimability framework by revisiting two classical extremes in statistical learning theory. The first corresponds to the case where the family of distributions $\mathcal{P}$ is unrestricted, leading to Vapnik–Chervonenkis theory.

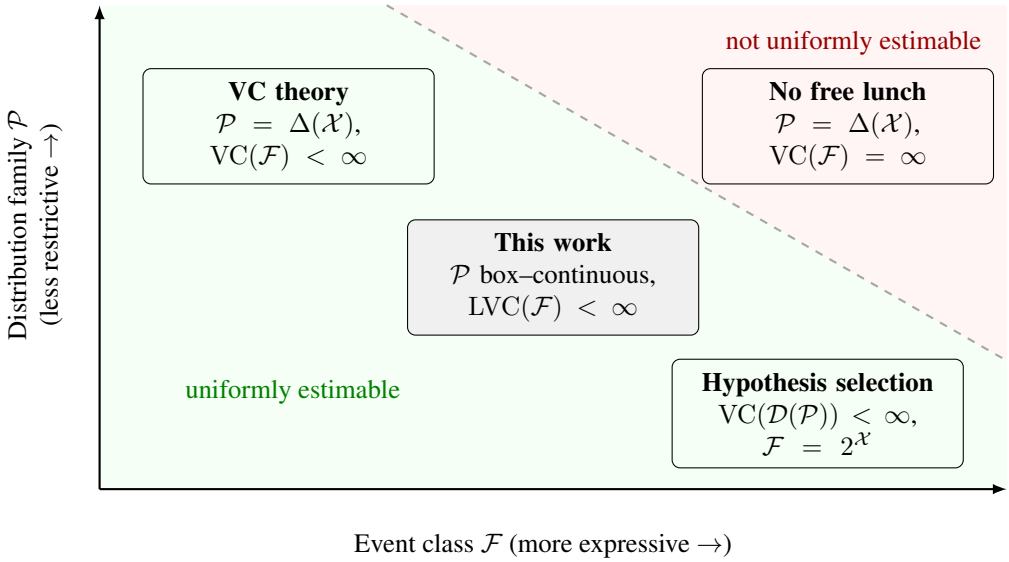

Figure 1: Uniform estimation regimes. The dashed curve separates uniformly and non-uniformly estimable regimes based on the richness of the event class $\mathcal{F}$ and distribution family $\mathcal{P}$.

The second corresponds to the case where the family of events $\mathcal{F}$ is unrestricted, which is intimately related to the *hypothesis selection* problem. These examples delineate the classical boundaries of uniform estimation and serve as reference points for the structured distribution families studied in this work.

THE FAMILY OF DISTRIBUTIONS $\mathcal{P}$ IS UNRESTRICTED (VC THEORY)

Consider the classical distribution-free setting in which the family of distributions is unrestricted, $\mathcal{P} = \Delta(\mathcal{X})$. The central question of Vapnik–Chervonenkis theory is to characterize those (possibly infinite) classes $\mathcal{F}$ for which $(\mathcal{F}, \Delta(\mathcal{X}))$ is uniformly estimable.[1] The seminal work of Vapnik and Chervonenkis (1971) shows that this holds if and only if $\mathcal{F}$ has finite VC dimension d. Moreover, the empirical distribution is minimax–optimal in this setting, achieving the sharp sample complexity $m(\varepsilon, \delta) = \Theta\big((d + \log(1/\delta))/\varepsilon^2\big)$ (Blumer et al., 1989; Talagrand, 1994).

THE FAMILY OF EVENTS $\mathcal{F}$ IS UNRESTRICTED (HYPOTHESIS SELECTION)

Consider the opposite extreme, in which the family of events is unrestricted, $\mathcal{F} = 2^{\mathcal{X}}$. In this case, uniform estimation coincides with learning the underlying distribution in total variation distance. When the family of distributions $\mathcal{P} = \{P_1, \dots, P_n\}$ is finite, this problem admits a classical solution via the hypothesis selection method of Yatracos (1985). The method is based on the *Yatracos sets* $D_{i,j} := \{x \in \mathcal{X} : P_i(x) \geq P_j(x)\}$, and the associated class $\mathcal{D}(\mathcal{P}) := \{D_{i,j} : i, j \in [n]\}$. These sets satisfy

$$\mathrm{TV}(P_i, P_j) = \max_{D \in \mathcal{D}} \big|P_i(D) - P_j(D)\big| \qquad \text{for all } i, j \in [n].$$

1. Classical VC theory focuses on uniform convergence of the empirical mean estimator and does not explicitly consider arbitrary estimators. However, the no-free-lunch lower bound underlying the necessity of finite VC dimension applies to *any* estimator, and thus the characterization extends beyond the empirical mean.

As a consequence, with $m = \Theta\big((\log n + \log(1/\delta))/\varepsilon^2\big)$ samples, the empirical distribution $\hat{p}_m$ uniformly approximates the probabilities of all sets in $\mathcal{D}$. Selecting a distribution $p_i \in \mathcal{P}$ that minimizes $\max_{D\in\mathcal{D}}\big|P_i(D) - \hat{P}_m(D)\big|$ yields a sound estimator. More generally, the same approach extends to infinite families $\mathcal{P}$ whenever the associated Yatracos class $\mathcal{D}(\mathcal{P})$ has finite VC dimension d, in which case the sample complexity scales as $m = \Omega\big((d + \log(1/\delta))/\varepsilon^2\big)$.

In contrast to the VC setting discussed above, a general characterization of those infinite families of distributions $\mathcal{P}$ for which the pair $(\mathcal{F} = 2^{\mathcal{X}}, \mathcal{P})$ is uniformly estimable is not known. Moreover, it is known that VC-type combinatorial characterizations are impossible in this regime (Lechner and Ben-David, 2024).

Roadmap. In Section 2, we state our main result in qualitative form. Within this section, Section 2.1 presents a collection of examples illustrating the scope and applicability of our result. In Section 2.2, we discuss the quantitative sample complexity bounds that follow from our analysis and formulate several open questions aimed at closing the remaining gaps. In Section 3, we provide a high-level overview of the proof and describe the resulting estimation procedure. Full proofs are deferred to the appendix.

2. Main Result

We now turn to uniform estimation on product domains $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$. To state our main result, we introduce two notions. The first is a distributional regularity condition, termed *uniform box–continuity*. The second is a combinatorial parameter of $\mathcal{F}$, called the *linear VC dimension*. With these definitions in place, we then state our main theorem.

Uniform Box–Continuity. We begin by recalling a quantitative formulation of absolute continuity. Let P and Q be probability distributions on a measurable space $(\mathcal{X}, \Sigma)$. Recall that P is absolutely continuous with respect to Q if for every $\alpha > 0$ there exists $\beta > 0$ such that for every measurable set E,

$$P(E) \geq \alpha \implies Q(E) \geq \beta.$$

This condition is the usual notion of absolute continuity, commonly denoted in measure theory by $P \ll Q$. We now specialize this notion to product spaces. Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, and let $P \in \Delta(\mathcal{X})$. Denote by P_i the marginal of P on $\mathcal{W}_i$, and by $P_\square := \prod_{i=1}^d P_i$ the product of marginals.

Definition 2 (Uniform box–continuity) *A family of distributions $\mathcal{P} \subseteq \Delta(\mathcal{X})$ is* uniformly box–continuous *if for every $\alpha > 0$ there exists $\beta > 0$ such that for every $P \in \mathcal{P}$ and every measurable set $E \subseteq \mathcal{X}$, $P(E) \geq \alpha \implies P_\square(E) \geq \beta$.*

Whenever a family $\mathcal{P}$ is uniformly box–continuous, we refer to any function $\beta\colon (0,1] \to (0,1]$ satisfying Definition 2 as a *modulus of box–continuity* for $\mathcal{P}$. We denote by $\mathcal{P}_\beta$ the family of all distributions that are box–continuous with modulus β. This condition is analogous to uniform continuity in analysis: the same $\alpha \mapsto \beta(\alpha)$ relationship applies uniformly to all distributions in the family $\mathcal{P}$. For example, the family of all product distributions on $\mathcal{X}$ is uniformly box–continuous, with $\beta(\alpha) = \alpha$. We present additional examples of uniformly box–continuous families after stating the main result.

Linear VC Dimension. We now introduce the combinatorial parameter governing uniform estimation in product spaces. A set $L \subseteq \mathcal{X}$ is called an *axis–parallel line* if there exists an index $i \in [d]$ and elements $w_j \in \mathcal{W}_j$ for all $j \neq i$ such that

$$L = \big\{(x_1, \ldots, x_d) \in \mathcal{X} : x_j = w_j \text{ for all } j \neq i\big\}.$$

A set $S \subseteq \mathcal{X}$ is called *colinear* if it is contained in some axis-parallel line. Equivalently, S is a set of points in $\mathcal{X}$ that agree on all but one coordinate.

Definition 3 (Linear VC dimension) *The* linear VC dimension *of a class $\mathcal{F} \subseteq 2^{\mathcal{X}}$, denoted $\mathrm{LVC}(\mathcal{F})$, is the largest integer k for which there exists a colinear set $S \subseteq \mathcal{X}$ of size k that is* shattered *by $\mathcal{F}$, meaning that for every labeling $S \to \{0,1\}$, there exists a set $F \in \mathcal{F}$ that realizes this labeling on S. If no such finite k exists, we set $\mathrm{LVC}(\mathcal{F}) = \infty$.*

Theorem 4 (Main Result) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ be a product measurable space, and let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ be a family of events. Then, the following are equivalent:*

1. $\mathrm{LVC}(\mathcal{F}) < \infty$.
2. *For every uniformly box–continuous family of distributions $\mathcal{P} \subseteq \Delta(\mathcal{X})$, the pair $(\mathcal{F}, \mathcal{P})$ is uniformly estimable.*

The necessity of finite linear VC dimension is relatively straightforward and follows from standard no–free–lunch lower bounds for VC classes. The more substantive part of the theorem is the converse direction, showing that finite linear VC dimension is also sufficient for uniform estimability under uniformly box–continuous distributions. In this sense, Theorem 4 identifies linear VC dimension as the only obstruction to uniform estimation in product spaces within this regime. Establishing sufficiency requires a combination of techniques, including packing arguments and the construction of grid–based estimators that explicitly exploit the underlying product structure. We provide a high–level overview of the proof strategy in Section 3.

The special case in which the family $\mathcal{P}$ consists solely of product distributions has been studied in several prior works (Cai and Daskalakis, 2017; Guo et al., 2020; Harms and Yoshida, 2022; Livni and Mansour, 2019; Coregliano and Malliaris, 2024). In particular, the results of Cai and Daskalakis (2017), Livni and Mansour (2019), and Coregliano and Malliaris (2024) imply our characterization in this regime. The conclusion of Cai and Daskalakis (2017) is weaker, as uniform estimation there is implied by the finiteness of the VC dimension of the coordinatewise projections of $\mathcal{F}$, a parameter that is always at least as large as the linear VC dimension and can be infinite even when $\mathrm{LVC}(\mathcal{F})$ is finite. By contrast, Livni and Mansour (2019) were the first to introduce a combinatorial dimension essentially equivalent to the linear VC dimension and use it to characterize uniform estimation in the symmetric i.i.d. setting; this perspective was further developed in Coregliano and Malliaris (2024), which replaces the i.i.d. assumption by weaker symmetry notions related to *exchangeability*.

Unlike these works, our results do not rely on symmetry assumptions or exact independence, and apply to any distribution families exhibiting weak independence as formalized by uniform box–continuity. For example, our framework covers finite mixtures of product distributions as well as distributions with low mutual information, such as non-degenerate multivariate Gaussians, which fall outside the scope of previous work. Additional examples illustrating this broader regime are discussed in the following section.

2.1. Examples

In this section we present several basic examples that illustrate the scope of Theorem 4 and aim to provide intuition for the notions of linear VC dimension and uniform box-continuity. Examples 1 and 2 present natural distribution families satisfying uniform box–continuity. Example 3 shows a natural class with infinite VC dimension and finite linear VC dimension. Example 4 highlights a setting in which the classical empirical estimator fails, motivating the need for estimators that explicitly exploit product structure.

Example 1 (Mixtures of product distributions). We begin with a basic example of uniform box-continuity beyond exact product distributions, given by finite mixtures of product distributions. Establishing uniform box-continuity in this case already requires some care.

Proposition 5 (Mixtures of product distributions) *Let P be a probability distribution on $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ that can be written as a mixture of at most k product distributions. Then P is uniformly box–continuous with modulus*

$$\beta(\alpha) \;=\; \frac{\alpha^d}{(k-1+\alpha)^{\,d-1}}.$$

We prove Proposition 5 in Appendix F.1, where we also show that the modulus is almost optimal, as $\beta(\alpha) \le \alpha^d/(k-1)^{\,d-1}$ is unavoidable.

Consequently, all mixtures with at most k components share a common modulus $\beta(\alpha) = \alpha^d/(k-1+\alpha)^{\,d-1}$ and form a uniformly box–continuous family. In the special case $k = 1$, this reduces to the setting of product distributions, for which $P = P_\square$ and uniform box–continuity holds with the identity modulus $\beta(\alpha) = \alpha$.

Example 2 (Bounded mutual information). Another natural source of uniform box–continuity arises from information–theoretic constraints. Recall that for a probability measure P on $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, the box projection $P_\square$ denotes the product distribution obtained from the one–dimensional marginals of P. The *total correlation* (also known as *multi–information*) is defined by $\mathrm{TC}(P) \;:=\; \mathrm{KL}(P \,\|\, P_\square)$, where KL denotes the Kullback-Leibler divergence. For $d = 2$, the total correlation coincides with the usual mutual information.

Proposition 6 (Bounded total correlation implies uniform box–continuity) *Fix $C \ge 0$ and let $\mathcal{P}_C := \{P : \mathrm{TC}(P) \le C\}$. Then the family $\mathcal{P}_C$ is uniformly box–continuous with modulus*

$$\beta(\alpha) \;=\; \exp\!\left(\frac{-H(\alpha) - C}{\alpha}\right),$$

where $H(\alpha) := -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ denotes the binary entropy function. In particular, every $P \in \mathcal{P}_C$ is box–continuous with this modulus.[2]

2. When $C = 0$, one has $\mathrm{TC}(P) = \mathrm{KL}(P \,\|\, P_\square) = 0$, which forces $P = P_\square$ and hence $\mathcal{P}_0$ coincides with the family of product distributions. In this case the optimal modulus of box–continuity is $\beta(\alpha) = \alpha$. Moreover, plugging $C = 0$ into the displayed formula gives $\beta(\alpha) = \exp(-H(\alpha)/\alpha) = \frac{\alpha}{e}\,(1 + o(1))$ as $\alpha \to 0$. The entropy–based bound stated here remains valid for $C = 0$, but is conservative and does not capture this sharp behavior.

Consequently, for every fixed $C > 0$, the class of distributions on $\mathcal{X}$ whose total correlation is at most C is uniformly box–continuous. The proof of Proposition 6 is given in Appendix F.2.

Gaussian case. Multivariate Gaussian distributions provide a concrete illustration of this phenomenon. Let $X = (X_1, \dots, X_d)$ be a Gaussian random vector in $\mathbb{R}^d$ with covariance matrix Σ. In this case, the total correlation admits the explicit closed–form expression

$$\mathrm{TC}(X) \;=\; \frac{1}{2}\log\!\left(\frac{\det(\operatorname{diag}\Sigma)}{\det(\Sigma)}\right),$$

where Σ is the covariance matrix of X and $\operatorname{diag}\Sigma$ is obtained by zeroing out all off–diagonal entries of Σ; see, e.g., Cover and Thomas (2012, Section 8.3). This quantity measures the deviation from independence in terms of the volume ratio between the product of marginals and the joint distribution. In particular, any family of Gaussian distributions for which this quantity is uniformly bounded has bounded total correlation, and therefore falls within the scope of Theorem 4.

In the bivariate case $d = 2$, this expression simplifies to $\mathrm{TC}(X) = \frac{1}{2}\log\!\left(\frac{1}{1-\rho^2}\right)$, where ρ is the correlation coefficient between X_1 and X_2. Thus, uniform box–continuity holds for bivariate Gaussians as long as ρ^2 is bounded away from 1.

Example 3 (Convex sets in $\mathbb{R}^d$). The power of Theorem 4 is most clearly seen when combined with hypothesis classes that lie far beyond the reach of classical VC theory. Let $\mathcal{X} = \mathbb{R}^d$ and let $\mathcal{F}$ be the family of all convex subsets of $\mathbb{R}^d$. Then

$$\mathrm{VC}(\mathcal{F}) = \infty \qquad \text{but} \qquad \mathrm{LVC}(\mathcal{F}) = 2.$$

To see that the classical VC dimension of $\mathcal{F}$ is infinite, note that any finite set of points in convex position in $\mathbb{R}^d$ is shattered by convex sets. When $d \geq 2$, there exist arbitrarily large such sets (for example, points on a circle), and hence $\mathrm{VC}(\mathcal{F}) = \infty$. On the other hand, the linear VC dimension satisfies $\mathrm{LVC}(\mathcal{F}) = 2$, since restricting convex sets to any line in $\mathbb{R}^d$ yields the class of intervals, which has VC dimension 2. Consequently, while convex sets are not uniformly estimable under arbitrary distributions, Theorem 4 implies that $(\mathcal{F}, \mathcal{P})$ is uniformly estimable for every uniformly box–continuous family $\mathcal{P}$.

In fact, to bound the linear VC dimension, it suffices to control the behavior of the class on axis–parallel lines, rather than on all affine lines. For example, consider the class $\mathcal{F}_{\mathrm{stair}}$ of *stair–convex* subsets of $\mathbb{R}^d$, where a set is stair–convex if for every two points that differ in a single coordinate and belong to the set, the entire axis–parallel line segment between them is also contained in the set. Restricting any stair–convex set to a coordinate axis yields an interval, and hence $\mathrm{LVC}(\mathcal{F}_{\mathrm{stair}}) = 2$.

Example 4 (Permutation graphs: why new machinery is needed). Let $\mathcal{X} = [n] \times [n]$, and let $\mathcal{F} = \{F_\pi : \pi \in S_n\}$ be the family of *permutation graphs*, where each F_π corresponds to a permutation $\pi : [n] \to [n]$ and is defined by

$$F_\pi := \{(i, \pi(i)) : i \in [n]\}.$$

Thus, each set in $\mathcal{F}$ contains exactly one point in every row and every column and hence $\mathrm{LVC}(\mathcal{F}) = 1$. By Theorem 4, $\mathcal{F}$ is therefore uniformly estimable with respect to any uniformly box–continuous family of distributions.

Nevertheless, the most natural estimator - namely, the empirical mean - fails even in this simple setting. To see this, consider the uniform distribution on $\mathcal{X} = [n] \times [n]$, which is a product distribution and hence uniformly box–continuous with $\beta(\alpha) = \alpha$. Suppose we draw $m \ll \sqrt{n}$ samples from this distribution. By the birthday paradox, with high probability no two samples share the same row and no two share the same column. Consequently, there exists a permutation π such that all sampled points lie in F_π. For the empirical estimator, this yields $\widehat{P}_m(F_\pi) = 1$, whereas the true probability satisfies $P(F_\pi) = 1/n$.

Since n can be taken arbitrarily large, this shows that the empirical mean does not uniformly estimate $\mathcal{F}$, despite the fact that $\mathrm{LVC}(\mathcal{F}) = 1$. This example highlights the need for estimators that explicitly exploit product structure, as predicted by our main result. Formal proofs of the claims above are given in Section F.3.

2.2. Quantitative Bounds

Beyond the qualitative characterization provided by Theorem 4, it is natural to ask for *quantitative* guarantees. In this subsection we discuss the sample complexity of uniform estimation and state explicit upper and lower bounds on the sample size needed to achieve uniform estimation.

2.2.1. UPPER AND LOWER BOUNDS

To state our quantitative bounds, we use the standard sample–complexity notation for uniform estimation. Let $m_{\mathcal{F},\mathcal{P}}(\varepsilon, \delta)$ denote the smallest $m \in \mathbb{N}$ for which there exists an estimator $\widehat{P}_m$, as in Definition 1, such that

$$\Pr_{S \sim p^m}\left[\sup_{F \in \mathcal{F}} \left|\widehat{P}_m(S)(F) - P(F)\right| \le \varepsilon\right] \;\ge\; 1 - \delta \qquad \text{for every } P \in \mathcal{P}.$$

Our proof of Theorem 4 is constructive and therefore yields explicit bounds on $m_{\mathcal{F},\mathcal{P}}(\varepsilon, \delta)$. In particular, we obtain the following general lower and upper bounds. Let $g := \mathrm{LVC}(\mathcal{F})$. Our lower bound shows that even when $\mathcal{P}$ is the class of product distributions one necessarily has

$$m_{\mathcal{F},\mathcal{P}}(\varepsilon, \delta) \;\ge\; \Omega\left(\frac{g + \log(1/\delta)}{\varepsilon^2}\right). \tag{1}$$

On the other hand, the main estimator gives a general upper bound. To state it compactly, set

$$T_1(\varepsilon, \delta) := \frac{\left(g + \log \frac{1}{\delta}\right)^{d-1}}{\beta(\varepsilon/2)^{2(d-1)}}$$

and

$$T_2(\varepsilon, \delta) := \frac{\left((g+d)\log\left(\frac{d^2(g+d)}{\beta(\varepsilon/2)}\right) + \log\frac{1}{\delta}\right)^{d-1}}{\beta(\varepsilon/2)^{d-1}}.$$

Then

$$m_{\mathcal{F},\mathcal{P}}(\varepsilon, \delta) \le \tilde{O}\left(\frac{\bar{C}_0^{d-1} d^{2(d-1)} g}{\varepsilon^2} \min\{T_1(\varepsilon, \delta), T_2(\varepsilon, \delta)\}\right). \tag{2}$$

Here $\bar{C}_0 > 0$ is a universal constant, d is the width of the product space $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, and β is the modulus of uniform box–continuity of $\mathcal{P}$. That is, for every measurable event $E \subseteq \mathcal{X}$, if $P(E) \geq \varepsilon/2$, then

$$P_{\square}(E) \geq \beta(\varepsilon/2).$$

The bound is proved in Appendix A; see in particular Lemma 12. The two terms in the minimum correspond to the two estimates derived in Appendix B and Appendix C, respectively.

Special case: product distributions. When $\mathcal{P}$ consists solely of product distributions, a sharper upper bound on the sample complexity can be obtained:

$$m_{\mathcal{F},\mathcal{P}}(\varepsilon,\delta) \;\leq\; O\left(\frac{d^2}{\varepsilon^2}\left(g + \log\tfrac{1}{\delta}\right)\right).$$

This bound is also recovered as a special case of our analysis (see Remark 17 in Appendix A), and in the product case it can be achieved using a similar estimator that exploits the exact product structure, without the need to handle uniform box–continuity. The same bound can be derived from the arguments of Cai and Daskalakis (2017); Livni and Mansour (2019), while one can obtain a conceptually similar guarantee using Coregliano and Malliaris (2024).

A comparison of (1) and (2) reveals a substantial gap between our current lower and upper bounds. Closing this gap leads to the following quantitative question.

Question 7 (Optimal dependence) *Let $\mathcal{P}$ be a uniformly box–continuous family with modulus β, and let $\mathcal{F}$ satisfy $g := \mathrm{LVC}(\mathcal{F})$. What are the tightest bounds on the sample complexity $m_{\mathcal{F},\mathcal{P}}(\varepsilon,\delta)$ that can be stated in terms of $d, g, \varepsilon, \delta$, and β?*

Among the various parameter gaps, the dependence on the width d stands out: our lower bound does not involve d, whereas the current upper bound has a super-exponential $\exp(\Theta(d \log d))$ dependence on d.

Tightness of the lower bound. We first observe that the lower bound (1) is already tight, up to constant factors, in its dependence on the parameters g, ε, and δ. This is witnessed by degenerate settings in which the product structure plays no essential role and the problem effectively reduces to a one–dimensional VC class. Specifically, if $\mathcal{F}$ is contained in a single axis–parallel line, then $\mathrm{VC}(\mathcal{F}) = \mathrm{LVC}(\mathcal{F}) = g$, and uniform estimation reduces to the classical one–dimensional setting. In this case, the optimal sample complexity satisfies $m_{\mathcal{F},\mathcal{P}}(\varepsilon,\delta) = \Theta\left(\frac{g+\log(1/\delta)}{\varepsilon^2}\right)$.

Tightness of the upper bound. The following example shows that for certain hypothesis classes, a dependence on the width d is unavoidable, even under product distributions.

Example 5 (Unavoidable dependence on the width d). Let $\mathcal{X} = \{0,1\}^d$ and $\mathcal{F} = 2^{\mathcal{X}}$. Then $\mathrm{LVC}(\mathcal{F}) = 2$, since every axis-parallel line in $\{0,1\}^d$ has size 2 and $\mathcal{F}$ restricted to such a line is the full power set. Moreover, the quantity $\sup_{F\in\mathcal{F}} |\widehat{P}(F) - P(F)|$ coincides with the total variation distance. Proposition 49 shows that for the family of product measures

$$\mathcal{P} = \left\{ \bigotimes_{i=1}^{d} Ber\left(\tfrac{1}{2} + \theta_i \nu\right) : \theta \in \{\pm 1\}^d \right\}, \qquad \nu = \Theta\left(\sqrt{\tfrac{\varepsilon}{d}}\right),$$

any estimator that is uniformly ε–accurate in total variation must use at least $m = \Omega(d/\varepsilon)$ samples.

This example shows that bounded $\mathrm{LVC}(\mathcal{F})$ does not rule out a dependence on the width d. At the same time, there are settings in which dimension–free rates are achievable. Understanding which additional structure of $\mathcal{F}$ distinguishes these regimes is a natural next question.

Question 8 (Rate-controlling invariants beyond LVC**)** *Is there a simple invariant of $\mathcal{F}$, extending or refining* $\mathrm{LVC}(\mathcal{F})$*, that governs the* sample–complexity rate *of uniform estimation in product spaces? More concretely, can one characterize when explicit dependence on the width d is necessary, and when it can be avoided?*

2.2.2. INFINITE PRODUCT SPACES

To probe these issues further, we move to the infinite–product setting. Here the width is unbounded, so any phenomenon that forces sample complexity to grow with d can no longer be hidden in quantitative constants and instead becomes a qualitative obstruction to uniform estimability. This makes infinite products a natural test case for identifying what features of a class $\mathcal{F}$ (beyond $\mathrm{LVC}(\mathcal{F})$) control estimability. In particular, characterizing uniform estimability on infinite product spaces provides an extreme starting point for addressing Question 8.

We focus on the infinite Boolean cube $\mathcal{X} = \{0,1\}^{\mathbb{N}}$ equipped with its canonical product σ–algebra. In this setting, we take $\mathcal{F}$ to be the family of all measurable events (i.e., the full product σ-algebra on $\mathcal{X}$), and let $\mathcal{P}$ denote the family of all product distributions on $\mathcal{X}$. Example 5 already implies that in this setting bounded linear VC dimension is insufficient to guarantee uniform estimability. Thus, in the infinite-dimensional product setting the linear VC dimension loses its role: while it yields a sharp characterization in finite-dimensional product spaces, it becomes too weak to control uniform estimability once the product has infinitely many coordinates.

This raises a natural question: what, if anything, should replace it? One possibility is that in infinite product spaces the benefit of product structure disappears altogether. For instance, on the infinite Boolean cube, could uniform estimability with respect to product measures already be as hard as uniform estimability with respect to arbitrary distributions? Equivalently, might uniform estimability in the infinite boolean cube be governed simply by the classical VC dimension? The following example shows that this is also false.

Example 6 (Infinite VC dimension with trivial uniform estimation). Let $\mathcal{X} = \{0,1\}^{\mathbb{N}}$ and let $\mathcal{P}_{\mathrm{prod}}$ denote the family of all product measures on $\mathcal{X}$. For $\alpha \in [0,1]$, define the event

$$E_\alpha := \Big\{ x \in \mathcal{X} : \lim_{n\to\infty} \tfrac{1}{n} \sum_{i=1}^{n} x_i = \alpha \Big\},$$

and let $\mathcal{F}$ be the class generated by finite unions of such events with rational parameters α. Then $\mathrm{VC}(\mathcal{F}) = \infty$, yet the pair $(\mathcal{F}, \mathcal{P}_{\mathrm{prod}})$ is uniformly estimable. In fact, uniform estimation is trivial in this case: a single sample suffices, so $m(\varepsilon,\delta) = 1$. This follows from Kolmogorov's 0-1 law, which implies that under any product measure every event in $\mathcal{F}$ has probability either 0 or 1, and is therefore almost surely determined by a single draw. See Appendix F.4 for details.

These examples show that neither $\mathrm{LVC}(\mathcal{F})$ nor $\mathrm{VC}(\mathcal{F})$ alone governs uniform estimability on $\{0,1\}^{\mathbb{N}}$. We conclude with the following open problem.

Question 9 (Uniform estimability in infinite product spaces) *Is there a natural invariant that characterizes uniform estimability in infinite product spaces?*

3. Technical Overview

We provide a high-level overview of the proof of Theorem 4, focusing on the sufficiency direction. The necessity of finite linear VC dimension follows from standard no-free-lunch lower bounds for VC classes and is therefore deferred to section A.

Upper bound: proof idea. Our upper-bound estimator proceeds in two steps (see also the pseudocode below):

- We first discretize the class $\mathcal{F}$, using a finite grid, thereby reducing the problem to a finite subclass.
- We then apply uniform estimation to this finite restriction and extend the resulting estimates back to the full class $\mathcal{F}$.

Fix accuracy and confidence parameters $\varepsilon, \delta \in (0,1)$, and let $S \sim P^m$ be an i.i.d. sample from P. Our estimator first splits S into two disjoint subsamples $S = (S^{(0)}, S^{(1)})$. The subsample $S^{(0)}$ is used for the first phase, and $S^{(1)}$ is used for the second phase.

Phase 1: discretization via representatives. The key idea of this phase is to obtain a finite discretization of the class $\mathcal{F}$ by constructing a grid G, which in turn is used to define a finite representative subclass of $\mathcal{F}$. The role of the grid is to ensure that agreement on G forces closeness in probability under the target distribution P. Using the initial sample $S^{(0)}$, we form a product grid

$$G = G_1 \times \cdots \times G_d, \qquad G_i := \pi_i(S^{(0)}) \subseteq \mathcal{W}_i,$$

where π_i denotes projection onto the i-th coordinate. That is, G is obtained by projecting the sample onto each coordinate separately and taking the Cartesian product of the resulting coordinate sets.

The crucial property we establish is a *hitting property* of the grid G with respect to the distribution P: with high probability over the draw of $S^{(0)}$, the grid intersects the symmetric difference of any two sets whose P-measure is large. Equivalently, G is an ε–net for the family of symmetric differences induced by $\mathcal{F}$ under the pseudo-metric $d_P(F, F') := P(F \triangle F')$. Formally, we show that, simultaneously for all $F, F' \in \mathcal{F}$,

$$P(F \triangle F') > \varepsilon/2 \quad \Longrightarrow \quad (F \triangle F') \cap G \neq \emptyset.$$

The proof proceeds in two steps. First, we establish the analogous statement with respect to the product-of-marginals distribution $P_\square$. Second, uniform box-continuity allows us to transfer this guarantee from $P_\square$ to P: since $P(F \triangle F') > \varepsilon/2$ implies $P_\square(F \triangle F') > \beta(\varepsilon/2)$, it suffices to ensure that G intersects every symmetric difference with $P_\square$-measure at least $\beta(\varepsilon/2)$. The required bounds are provided by Lemma 22, proved in the appendix. As a consequence, the grid G induces an equivalence relation on $\mathcal{F}$ given by

$$F \sim_G F' \quad \Longleftrightarrow \quad F \cap G = F' \cap G.$$

Let $\mathcal{F}_G \subseteq \mathcal{F}$ denote a choice of one representative from each $\sim_G$-equivalence class. This yields a finite representative family that discretizes $\mathcal{F}$ at resolution $\varepsilon/2$ with respect to P.

Phase 2: estimation on the representative family. Having constructed the finite representative family $\mathcal{F}_G$, we now turn to the second phase, in which uniform estimation over $\mathcal{F}$ is reduced to uniform

estimation over $\mathcal{F}_G$. For each $F \in \mathcal{F}$, let $F^\star \in \mathcal{F}_G$ denote its representative under $\sim_G$. On the high-probability event from Phase 1, the hitting property implies that agreement on G forces proximity under P: if $F \sim_G F^\star$, then $(F \triangle F^\star) \cap G = \emptyset$, and hence $P(F \triangle F^\star) \le \varepsilon/2$. Consequently, every set $F \in \mathcal{F}$ can be *rounded* to its representative $F^\star \in \mathcal{F}_G$ at a cost of $\le \varepsilon/2$ difference in probability. Using the independent subsample $S^{(1)}$, we estimate $P(F^\star)$ for each $F^\star \in \mathcal{F}_G$ by the empirical estimator $\widehat{P}_{\mathrm{diag}}$. Since $\mathcal{F}_G$ is finite, standard uniform convergence bounds for finite classes (e.g., Hoeffding's inequality combined with a union bound) imply that, for a suitable choice of m_1, with probability at least $1 - \delta/2$, $\sup_{F^\star \in \mathcal{F}_G} \left|\widehat{P}_{\mathrm{diag}}(F^\star) - P(F^\star)\right| \le \varepsilon/2$. We define the final estimator by extension through representatives,

$$\widehat{P}(F) := \widehat{P}_{\mathrm{diag}}(F^\star), \qquad F \in \mathcal{F}.$$

Combining the rounding error from Phase 1 with the estimation error above yields, on the intersection of the two high-probability events, $|P(F) - \widehat{P}(F)| \le \varepsilon$, for all $F \in \mathcal{F}$. We summarize the resulting algorithm below.

Product-Grid Estimation Algorithm.
Input: accuracy $\varepsilon \in (0,1)$, confidence $\delta \in (0,1)$, sample $S = (Z_1, \dots, Z_m) \in \mathcal{X}^m$.
Phase I: Discretization via representatives.

1. Construct a product grid. Split the sample into two disjoint parts $S = (S^{(0)}, S^{(1)})$. For each coordinate $i \in [d]$, let $G_i := \pi_i(S^{(0)})$ denote the projection of $S^{(0)}$ onto the i-th coordinate, and define the grid

$$G := G_1 \times \cdots \times G_d \subseteq \mathcal{X}.$$

Phase II: Discretization, estimation, and extension.

2. Discretization via representatives. Partition $\mathcal{F}$ into equivalence classes according to their traces on G,

$$F \sim F' \iff F \cap G = F' \cap G,$$

and let $\mathcal{F}_G$ be a set of representatives.

3. Estimation on the representative family. For each $F^\star \in \mathcal{F}_G$, estimate $P(F^\star)$ by the empirical mean over $S^{(1)}$:

$$\widehat{P}(F^\star) = \frac{1}{|S^{(1)}|} \sum_{z \in S^{(1)}} \mathbf{1}\{z \in F^\star\}.$$

Output: for each $F \in \mathcal{F}$, let $F^\star \in \mathcal{F}_G$ satisfy $F \cap G = F^\star \cap G$, and return $\widehat{P}(F) := \widehat{P}(F^\star)$.

Sample complexity bound. The sample complexity bound decomposes according to the two phases of the procedure.
Discretization via representatives. The first subsample of size m_0 is used to construct a product grid G with the property that agreement on G implies closeness in probability. Lemma 22 shows that, under the uniform box-continuity assumption, this holds with probability at least $1 - \delta/2$ provided that

$$m_0 \gtrsim \frac{d^2}{\beta(\varepsilon/2)^2}\left(g + \log \tfrac{1}{\delta}\right).$$

Estimation on the representative family. The discretization step yields a finite representative family $\mathcal{F}_G$. Once $\mathcal{F}_G$ is fixed, estimating the probabilities of all $F^\star \in \mathcal{F}_G$ reduces to a standard finite-class uniform estimation problem. Standard uniform convergence bounds for finite classes imply that $m_1 = O\big((\log|\mathcal{F}_G| + \log(1/\delta))/\varepsilon^2\big)$ samples suffice. Accordingly, the second-stage sample complexity depends only on $\log|\mathcal{F}_G|$, the number of distinct traces induced by $\mathcal{F}$ on the grid G.

Thus, controlling the overall sample complexity reduces to upper bounding $|\mathcal{F}_G|$, the number of distinct traces induced by $\mathcal{F}$ on the grid G. To achieve this, we prove a variant of the Sauer–Shelah–Perles lemma (Sauer, 1972; Shelah, 1972) in terms of the linear VC dimension, which we discuss next.

Grid SSP lemma. Given a grid $N = A_1 \times A_2 \ldots \times A_d$, where each $A_i \subseteq \mathcal{W}_i$ we seek to upper bound the number of distinct traces $\{F \cap N : F \in \mathcal{F}\}$ in terms of the linear VC dimension.

Lemma 10 (Grid Sauer-Shelah-Perles bound, informal) *If* $\mathrm{LVC}(\mathcal{F}) = g < \infty$*, then for any grid* $N = A_1 \times \cdots \times A_d$ *one has*

$$\log_2\big|\{F \cap N : F \in \mathcal{F}\}\big| \;\le\; O\Big(g\, n^{d-1}\log(n/g)\Big),$$

where $n := \max_i |A_i|$*. Moreover, for every* d, g *and* n*, there exist set families* $\mathcal{F}$ *with* $\mathrm{LVC}(\mathcal{F}) = g$ *for which*

$$\log_2\big|\{F \cap N : F \in \mathcal{F}\}\big| \;\ge\; \Omega\Big(g\, n^{d-1}\log(n/g)\Big).$$

In particular, the dependence on n in Lemma 10 is optimal up to constant factors. As an example, consider the case $g = 1$. Let $N = [n]^d$, and let $\mathcal{F}$ consist of all subsets $F \subseteq N$ that contain at most one point on every axis-parallel line. Equivalently, for each choice of a coordinate $i \in [d]$ and every fixing of the remaining $d-1$ coordinates, F contains at most one point on the corresponding line. Such sets are commonly referred to as $(d-1)$-dimensional permutations. In the special case $d = 2$, this notion reduces to permutation matrices: subsets of $[n] \times [n]$ that contain one point in each row and each column.

This class satisfies $\mathrm{LVC}(\mathcal{F}) = 1$. By a result of Keevash (2018) (Theorem 1.8), the number of $(d-1)$-dimensional permutations grows as

$$\big|\mathcal{F}\big| \;=\; \Big(\tfrac{n}{e^{d-1}} + o(n)\Big)^{n^{d-1}},$$

which matches the upper bound in Lemma 10 up to a constant. Lower bounds for general g are obtained by taking unions of g such $(d-1)$-dimensional permutations. These constructions yield classes with $\mathrm{LVC}(\mathcal{F}) = g$ whose number of distinct traces on N grows as $2^{\Omega(g\, n^{d-1}\log(n/g))}$, as stated above.

Acknowledgments

Shay Moran and Alexander Shlimovich are supported by the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Shay Moran is also supported by Israel PBC-VATAT, and by the Technion Center for Machine Learning and Intelligent Systems (MLIS).

References

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, July 1997. ISSN 0004-5411. doi: 10.1145/263867.263927.

Noga Alon, Alon Gonen, Elad Hazan, and Shay Moran. Boosting Simple Learners. *TheoretiCS*, Volume 2:9253, June 2023. ISSN 2751-4838. doi: 10.46298/theoretics.23.8.

Peter L. Bartlett and Philip M. Long. Prediction, Learning, Uniform Convergence, and Scale-sensitive Dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, April 1998. ISSN 00220000. doi: 10.1006/jcss.1997.1557.

Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, October 1989. ISSN 0004-5411, 1557-735X. doi: 10.1145/76359.76371.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, November 2005. ISSN 1292-8100, 1262-3318. doi: 10.1051/ps:2005018.

Yang Cai and Constantinos Daskalakis. Learning Multi-item Auctions with (or without) Samples, September 2017.

Leonardo N. Coregliano and Maryanthe Malliaris. High-arity PAC learning via exchangeability, September 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, November 2012. ISBN 978-1-118-58577-1.

Richard Durrett. *Probability: Theory and Examples*. Cambridge University Press, April 2019. ISBN 978-1-108-47368-2.

Chenghao Guo, Zhiyi Huang, Zhihao Gavin Tang, and Xinzhi Zhang. Generalizing Complex Hypotheses on Product Distributions: Auctions, Prophet Inequalities, and Pandora's Problem, July 2020.

Nathaniel Harms and Yuichi Yoshida. Downsampling for Testing and Learning in Product Distributions. In Mikołaj Bojańczyk, Emanuela Merelli, and David P. Woodruff, editors, *49th International Colloquium on Automata, Languages, and Programming (ICALP 2022)*, volume 229 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 71:1–71:19, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-235-8. doi: 10.4230/LIPIcs.ICALP.2022.71.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, June 1994. ISSN 0022-0000. doi: 10.1016/S0022-0000(05)80062-5.

Peter Keevash. The existence of designs II, February 2018.

Tosca Lechner and Shai Ben-David. Inherent limitations of dimensions for characterizing learnability of distribution classes. In *Proceedings of Thirty Seventh Conference on Learning Theory*, pages 3353–3374. PMLR, June 2024.

Roi Livni and Yishay Mansour. Graph-based Discriminators: Sample Complexity and Expressiveness, June 2019.

N Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, July 1972. ISSN 0097-3165. doi: 10.1016/0097-3165(72)90019-2.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, September 2009. ISBN 978-0-470-31719-8.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. ISBN 978-1-107-05713-5.

Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, January 1972. ISSN 0030-8730.

Albert N. Shiryaev. *Probability*. Springer, New York, 1996. ISBN 978-0-387-94549-1.

M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28–76, January 1994. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176988847.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. ISBN 978-0-521-78450-4. doi: 10.1017/CBO9780511802256.

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, January 1971. ISSN 0040-585X. doi: 10.1137/1116025.

Yannis G. Yatracos. Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy. *The Annals of Statistics*, 13(2):768–774, June 1985. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176349553.

Appendix A. Proof of Theorem 4

Overview. This section collects the proofs underlying Theorem 4. The implication $(1) \Rightarrow (2)$ is established by Proposition 11, which shows that finite linear VC dimension suffices for uniform estimability under any uniformly box–continuous family of distributions, together with a quantitative sample–size bound proved in Lemma 12. The converse implication $(2) \Rightarrow (1)$ follows from Proposition 14, which constructs a family of product distributions under which uniform estimation is impossible when $\mathrm{LVC}(\mathcal{F}) = \infty$.

Theorem 4. *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ be a product measurable space, and let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ be a family of events. Then, the following are equivalent:*

1. $\mathrm{LVC}(\mathcal{F}) < \infty$.

2. *For every uniformly box–continuous family of distributions $\mathcal{P} \subseteq \Delta(\mathcal{X})$, the pair $(\mathcal{F}, \mathcal{P})$ is uniformly estimable.*

Proof We prove the two implications separately.

(1)⇒(2). Assume that $\mathrm{LVC}(\mathcal{F}) = g < \infty$. Let $\mathcal{P} \subseteq \Delta(\mathcal{X})$ be an arbitrary uniformly box–continuous family of distributions. By definition, there exists a nondecreasing modulus $\beta : (0,1] \to (0,1]$ such that every $P \in \mathcal{P}$ is box–continuous with modulus β, i.e., $\mathcal{P} \subseteq \mathcal{P}_\beta$.

Fix $\varepsilon, \delta \in (0,1)$. Applying Proposition 11 with accuracy ε and confidence δ to the family $\mathcal{P}_\beta$, we obtain a sample size $m_0 = m(\varepsilon, \delta, g, d, \beta(\varepsilon/2))$ and an estimator $\widehat{P} : \mathcal{X}^\star \to [0,1]^{\mathcal{F}}$ such that for every $m \geq m_0$ and every $P \in \mathcal{P}_\beta$, with probability at least $1 - \delta$ over the sample $S \sim P^m$,

$$\sup_{F \in \mathcal{F}} \left| P(F) - \widehat{P}(S)(F) \right| \leq \varepsilon.$$

Since $\mathcal{P} \subseteq \mathcal{P}_\beta$, the same estimator and sample size satisfy the same guarantee for every $P \in \mathcal{P}$. As $\varepsilon, \delta \in (0,1)$ were arbitrary, this shows that $(\mathcal{F}, \mathcal{P})$ is uniformly estimable.

(2)⇒(1). We prove the contrapositive. Assume that $\mathrm{LVC}(\mathcal{F}) = \infty$. Let $\mathcal{P}_0$ denote the family of all product distributions on $\mathcal{X}$; this family is uniformly box–continuous. By Proposition 14 (Non–estimability when $\mathrm{LVC}(\mathcal{F}) = \infty$), the pair $(\mathcal{F}, \mathcal{P}_0)$ is not uniformly estimable. Hence statement (2) fails, since it requires uniform estimability for *every* uniformly box–continuous family. This proves the contrapositive, and therefore $(2) \Rightarrow (1)$.

Combining the two implications completes the proof. ■

Proposition 11 (Uniform estimability under bounded linear VC dimension) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ be a product measurable space, and let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ be a family of measurable sets with $\mathrm{LVC}(\mathcal{F}) = g < \infty$. Let $\mathcal{P} \subseteq \Delta(\mathcal{X})$ be a family of probability distributions that is uniformly box–continuous with modulus $\beta : (0,1] \to (0,1]$.*

Then for every $\varepsilon, \delta \in (0,1)$ there exists a sample size $m_0 = m(\varepsilon, \delta, g, d, \beta(\varepsilon/2))$ and an estimator $\widehat{P} : \mathcal{X}^\star \to [0,1]^{\mathcal{F}}$[3] such that for every $m \geq m_0$ and every $P \in \mathcal{P}$, with probability at least $1 - \delta$ over the sample $S \sim P^m$,

$$\sup_{F \in \mathcal{F}} \left| P(F) - \widehat{P}(S)(F) \right| \leq \varepsilon.$$

In particular, the pair $(\mathcal{F}, \mathcal{P})$ is uniformly estimable.

Proof Fix $\varepsilon, \delta \in (0,1)$. We construct an estimator $\widehat{P} : \mathcal{X}^\star \to [0,1]^{\mathcal{F}}$ and show that there exists a sample size $m_0 = m(\varepsilon, \delta, g, d, \beta(\varepsilon/2))$ such that for every $m \geq m_0$ and every $P \in \mathcal{P}$,

$$\Pr_{S \sim P^m} \left[\sup_{F \in \mathcal{F}} \left| P(F) - \widehat{P}(S)(F) \right| \leq \varepsilon \right] \geq 1 - \delta.$$

We choose integers m_0 and m_1 (specified below) and set

$$m := m_0 + m_1.$$

3. Here $\mathcal{X}^\star := \bigcup_{n \in \mathbb{N}} \mathcal{X}^n$ denotes the set of all finite samples.

Given a sample $S \sim P^m$, we write $S = (S^{(0)}, S^{(1)})$, where

$$S^{(0)} = (X^{(1)}, \dots, X^{(m_0)}) \sim P^{m_0}, \qquad S^{(1)} = (Y^{(1)}, \dots, Y^{(m_1)}) \sim P^{m_1}$$

are disjoint subsamples.

Let $G := G(S^{(0)})$ be the sample grid induced by $S^{(0)}$ (Definition 18). Choose m_0 large enough so that Lemma 22 holds with parameters $(\varepsilon/2, \delta/2)$. On the corresponding event, which has probability at least $1 - \delta/2$, the following holds simultaneously for all $F, F' \in \mathcal{F}$:

$$F \cap G = F' \cap G \quad \Longrightarrow \quad |P(F) - P(F')| \le \varepsilon/2.$$

Define an equivalence relation on $\mathcal{F}$ by $F \sim_G F'$ if and only if $F \cap G = F' \cap G$, and let $\mathcal{F}_G \subseteq \mathcal{F}$ be a set of representatives containing exactly one element from each equivalence class. For each $F \in \mathcal{F}$, denote by $F^\star \in \mathcal{F}_G$ its representative. Then, on the same event,

$$\forall F \in \mathcal{F}, \qquad |P(F) - P(F^\star)| \le \varepsilon/2. \tag{3}$$

Using the second subsample $S^{(1)}$, define the empirical estimator

$$\widehat{P}^{(1)}_{\mathrm{diag}}(A) \;:=\; \frac{1}{m_1} \sum_{i=1}^{m_1} \mathbf{1}\{Y^{(i)} \in A\}, \qquad A \subseteq \mathcal{X}.$$

We define the final estimator by

$$\widehat{P}(S)(F) := \widehat{P}^{(1)}_{\mathrm{diag}}(F^\star), \qquad F \in \mathcal{F}.$$

Conditioning on $S^{(0)}$, the variables $\mathbf{1}\{Y^{(i)} \in F^\star\}$ are i.i.d., bounded in $[0,1]$, with mean $P(F^\star)$ and the family $\mathcal{F}_G$ is finite and fixed. For any fixed $F^\star \in \mathcal{F}_G$, Hoeffding's inequality gives

$$\Pr\Big(\big|\widehat{P}^{(1)}_{\mathrm{diag}}(F^\star) - P(F^\star)\big| > \varepsilon/2 \;\Big|\; S^{(0)} \Big) \;\le\; 2\exp\!\left(-\frac{m_1 \varepsilon^2}{2}\right).$$

Applying a union bound over all $F^\star \in \mathcal{F}_G$ yields

$$\Pr\Big(\sup_{F^\star \in \mathcal{F}_G} \big|\widehat{P}^{(1)}_{\mathrm{diag}}(F^\star) - P(F^\star)\big| > \varepsilon/2 \;\Big|\; S^{(0)} \Big) \;\le\; 2|\mathcal{F}_G| \exp\!\left(-\frac{m_1 \varepsilon^2}{2}\right).$$

Thus, if

$$m_1 \;\ge\; \frac{2}{\varepsilon^2} \log\!\left(\frac{4|\mathcal{F}_G|}{\delta}\right),$$

then

$$\Pr\Big(\sup_{F^\star \in \mathcal{F}_G} \big|\widehat{P}^{(1)}_{\mathrm{diag}}(F^\star) - P(F^\star)\big| \le \varepsilon/2 \;\Big|\; S^{(0)} \Big) \ge 1 - \delta/2. \tag{4}$$

On the intersection of the events in (3) and (4), for every $F \in \mathcal{F}$,

$$\big|P(F) - \widehat{P}(S)(F)\big| \le |P(F) - P(F^\star)| + |P(F^\star) - \widehat{P}^{(1)}_{\mathrm{diag}}(F^\star)| \le \varepsilon.$$

By a union bound, this event occurs with probability at least $1 - \delta$. This completes the proof. ∎

Lemma 12 (Sample size bound for Proposition 11) *Under the assumptions of Proposition 11, let β be the common modulus of uniform box–continuity of $\mathcal{P}$. Define*

$$M_0^{(1)} := C_0 \frac{d^2}{\beta(\varepsilon/2)^2}\left(g + \log\frac{1}{\delta}\right),$$

where $C_0 > 0$ is the universal constant from Lemma 22. Also define

$$M_0^{(2)} := C_0' \frac{d^2}{\beta(\varepsilon/2)}\left((g+d)\log\left(\frac{d^2(g+d)}{\beta(\varepsilon/2)}\right) + \log\frac{1}{\delta}\right),$$

where $C_0' > 0$ is the universal constant from Lemma 27.

Then there exists a universal constant $C > 0$ such that the estimator constructed in the proof of Proposition 11 succeeds whenever, for either choice $M_0 \in \{M_0^{(1)}, M_0^{(2)}\}$,

$$m \;\geq\; C\left[M_0 + \frac{1}{\varepsilon^2}\left(g\, M_0^{d-1}\log\left(\frac{eM_0}{g}\right) + \log\frac{1}{\delta}\right)\right].$$

In particular, using $M_0 = M_0^{(1)}$ gives

$$m = \tilde{O}\left(\frac{C_0^{d-1} d^{2(d-1)}}{\varepsilon^2\,\beta(\varepsilon/2)^{2(d-1)}}\, g\left(g + \log\frac{1}{\delta}\right)^{d-1}\right),$$

whereas using $M_0 = M_0^{(2)}$ gives

$$m = \tilde{O}\left(\frac{{C_0'}^{d-1} d^{2(d-1)}}{\varepsilon^2\,\beta(\varepsilon/2)^{d-1}}\, g\left((g+d)\log\left(\frac{d^2(g+d)}{\beta(\varepsilon/2)}\right) + \log\frac{1}{\delta}\right)^{d-1}\right).$$

Here $\tilde{O}(\cdot)$ hides polylogarithmic factors in $d, g, 1/\beta(\varepsilon/2), 1/\varepsilon$, and $1/\delta$, but not the displayed powers of C_0 and C_0'.

Proof We keep the notation from the proof of Proposition 11. Write $m = m_0 + m_1$ for the sample split, and let $G = G(S^{(0)}) = \prod_{i=1}^d A_i$ be the empirical grid formed from $S^{(0)}$.

Step 1: choice of m_0. The first-stage sample is used only to ensure that the empirical grid hits all large symmetric differences. This can be done in either of two ways. By Lemma 22, it suffices to take

$$m_0 \geq M_0^{(1)}.$$

Alternatively, by Lemma 27, it suffices to take

$$m_0 \geq M_0^{(2)}.$$

In the rest of the proof, fix either choice $M_0 \in \{M_0^{(1)}, M_0^{(2)}\}$ and take $m_0 = M_0$.

Step 2: bound on $|\mathcal{F}_G|$. Since $G = \prod_{i=1}^d A_i$ is induced by $S^{(0)}$, we have $|A_i| \leq m_0$ for every i. Hence, writing $n := \max_i |A_i|$, we have $n \leq m_0$. Assuming $m_0 \geq g$, Corollary 36 gives

$$\log|\mathcal{F}_G| = \log\big|\{F \cap G : F \in \mathcal{F}\}\big| \leq g\, n^{d-1}\log_2\left(\frac{en}{g}\right) \leq g\, m_0^{d-1}\log_2\left(\frac{em_0}{g}\right).$$

Step 3: choice of m_1. As shown in the proof of Proposition 11, conditioning on $S^{(0)}$, it suffices to choose

$$m_1 \geq \frac{2}{\varepsilon^2}\left(\log\bigl(4|\mathcal{F}_G|\bigr) + \log\frac{1}{\delta}\right).$$

Substituting the bound on $\log|\mathcal{F}_G|$, we see that it suffices to take

$$m_1 \geq C\frac{1}{\varepsilon^2}\left(g\, m_0^{d-1}\log\left(\frac{em_0}{g}\right) + \log\frac{1}{\delta}\right)$$

for a universal constant $C > 0$.

Taking $m_0 = M_0$ and $m = m_0 + m_1$ gives the stated bound. The two $\tilde{O}(\cdot)$ estimates follow by substituting respectively $M_0 = M_0^{(1)}$ and $M_0 = M_0^{(2)}$, and absorbing logarithmic factors. ■

Definition 13 (Sections and section classes) *Let $\mathcal{X} = \prod_{i=1}^d \mathcal{W}_i$. For $j \in [d]$ and*

$$x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) \in \prod_{i\neq j}\mathcal{W}_i,$$

we write (x_{-j}, x_j) for the point of $\mathcal{X}$ obtained by inserting x_j in the j-th coordinate.

For a set $F \subseteq \mathcal{X}$, the j-th section of F at x_{-j} is

$$F_{x_{-j}} := \{x_j \in \mathcal{W}_j : (x_{-j}, x_j) \in F\}.$$

For a family $\mathcal{F} \subseteq 2^{\mathcal{X}}$, we denote the corresponding section class by

$$\mathcal{F}|_{x_{-j}} := \{F_{x_{-j}} : F \in \mathcal{F}\} \subseteq 2^{\mathcal{W}_j}.$$

Proposition 14 (Lower bound via infinite linear VC dimension) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ and let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ be a family of measurable sets with $\mathrm{LVC}(\mathcal{F}) = \infty$. Let $\mathcal{P}_{\mathrm{prod}}$ be the family of all product distributions on $\mathcal{X}$. Then $(\mathcal{F}, \mathcal{P}_{\mathrm{prod}})$ is not uniformly estimable.*

More precisely, for every sample–size function $m = m(\varepsilon, \delta)$ there exist parameters $\varepsilon, \delta \in (0, 1)$ and a distribution $P \in \mathcal{P}_{\mathrm{prod}}$ such that for any estimator $\widehat{P}$ based on $m(\varepsilon, \delta)$ i.i.d. samples,

$$\Pr_{S\sim P^{\otimes m(\varepsilon,\delta)}}\Bigl[\sup_{F\in\mathcal{F}}\bigl|P(F) - \widehat{P}(F)\bigr| > \varepsilon\Bigr] \;\geq\; \delta.$$

Proof Fix an arbitrary sample–size function $m = m(\varepsilon, \delta)$. Let $\varepsilon := \varepsilon_0$ and $\delta := 1/4$, where $\varepsilon_0 > 0$ is the universal constant from a standard VC lower bound (See Shalev-Shwartz and Ben-David (2014); Blumer et al. (1989)).

Set $m_0 := m(\varepsilon, \delta)$. Since $\mathrm{LVC}(\mathcal{F}) = \infty$, there exist a coordinate $j \in [d]$ and a section $x_{-j} \in \prod_{i\neq j}\mathcal{W}_i$ such that the associated section class $\mathcal{F}|_{x_{-j}}$ has VC dimension at least g, for some $g > m_0$. In particular, $\mathcal{F}|_{x_{-j}}$ shatters a set $\{z_1, \ldots, z_g\} \subseteq \mathcal{W}_j$.

Let Q_j be the uniform distribution on $\{z_1, \ldots, z_g\}$, and define the product distribution P on $\mathcal{X}$ by

$$P = \delta_{x_1} \otimes \cdots \otimes \delta_{x_{j-1}} \;\otimes\; Q_j \;\otimes\; \delta_{x_{j+1}} \otimes \cdots \otimes \delta_{x_d}.$$

Under P, sampling $S \sim P^{\otimes m_0}$ is equivalent to sampling m_0 i.i.d. points from Q_j on $\mathcal{W}_j$, since all other coordinates are fixed. Moreover, for every $F \in \mathcal{F}$,

$$P(F) = Q_j\big(F_{x_{-j}}\big)\,.$$

Hence estimating the probabilities $\{P(F) : F \in \mathcal{F}\}$ is at least as hard as estimating the probabilities $\{Q_j(A) : A \in \mathcal{F}|_{x_{-j}}\}$ on the one–dimensional domain $\mathcal{W}_j$.

Since $\mathrm{VC}(\mathcal{F}|_{x_{-j}}) \geq g > m_0$, the classical VC lower bound implies that for any estimator $\widehat{P}$ based on m_0 samples,

$$\Pr_{S\sim P^{\otimes m_0}}\Big[\sup_{F\in\mathcal{F}}\big|P(F) - \widehat{P}(F)\big| > \varepsilon\Big] \;\geq\; \delta.$$

This shows that no sample–size function $m(\varepsilon,\delta)$ can guarantee uniform estimation over $\mathcal{P}_{\mathrm{prod}}$, and therefore $(\mathcal{F}, \mathcal{P}_{\mathrm{prod}})$ is not uniformly estimable. ∎

Appendix B. Upper Bound Lemmas

In this section, we prove the auxiliary lemmas used in the proof of Proposition 11 and in the derivation of the sample-size bound in Lemma 12.

B.1. Uniform control under product measures

B.1.1. THE EMPIRICAL PRODUCT ESTIMATOR

Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ be a product measurable space. For a sample

$$S = (X^{(1)}, \ldots, X^{(m)}) \sim P^{\otimes m},$$

define, for each coordinate $j \in [d]$, the empirical marginal distribution $\tilde{P}_j \in \Delta(\mathcal{W}_j)$ by

$$\tilde{P}_j(A) \;:=\; \frac{1}{m}\sum_{t=1}^{m} \mathbf{1}\{X_j^{(t)} \in A\}, \qquad A \subseteq \mathcal{W}_j \text{ measurable.}$$

The *empirical product estimator* associated with S is the product distribution

$$\tilde{P} \;:=\; \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d \quad \in \Delta(\mathcal{X}). \tag{5}$$

Relation to prior work and statistical structure. The empirical product estimator (5) coincides with the estimator used in the product empirical risk minimization (PERM) framework of Guo et al. (2020). Closely related procedures also appear in earlier work of Livni and Mansour (2019), where empirical products of coordinate-wise marginals are employed to decouple dependencies across dimensions.

From a statistical perspective, the empirical product estimator $\tilde{P}$ can be viewed as a *V-statistic* of order d, obtained by averaging over all d-tuples formed from the sample with replacement. While most of the classical literature focuses on *U-statistics*, which average over tuples without replacement, the same analytical machinery applies to V-statistics with only minor modifications (see, e.g., Serfling (2009); van der Vaart (1998)).

B.1.2. EXPECTED UNIFORM DEVIATION

Lemma 15 (One-coordinate section bound) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ be a product measurable space, let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ satisfy $\mathrm{LVC}(\mathcal{F}) = g < \infty$, and let P be a probability measure on $\mathcal{X}$ with marginals $P_1, \ldots, P_d$. Let $S = (X_1, \ldots, X_m) \sim P^m$ be an i.i.d. sample, and let $\tilde{P}_i$ be the empirical marginal on coordinate i.*

Fix $j \in [d]$, and define the mixed marginal measure on the coordinates outside j by

$$Q_{-j}^{(j-1)} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_{j-1} \otimes P_{j+1} \otimes \cdots \otimes P_d.$$

Then there exists a universal constant $C > 0$ such that

$$\mathbb{E}_S \left[\mathbb{E}_{x_{-j} \sim Q_{-j}^{(j-1)}} \left[\sup_{F \in \mathcal{F}} \big(P_j(F_{x_{-j}}) - \tilde{P}_j(F_{x_{-j}}) \big) \right] \right] \leq C \sqrt{\frac{g}{m}} + \frac{j-1}{m}.$$

Proof We use the notation of Definition 13. Since $\mathrm{LVC}(\mathcal{F}) = g$, every section class $\mathcal{F}|_{x_{-j}}$ has VC dimension at most g.

The goal is to compare P_j and $\tilde{P}_j$ uniformly over the sections $\mathcal{F}|_{x_{-j}}$, where

$$x_{-j} \sim Q_{-j}^{(j-1)} = \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_{j-1} \otimes P_{j+1} \otimes \cdots \otimes P_d.$$

The only point requiring care is that x_{-j} is partly constructed from the same sample S, so the empirical sample $(X_1^{(j)}, \ldots, X_m^{(j)})$ is not automatically independent of the chosen section. We now remove this dependence by replacing the few potentially exposed j-coordinates.

Generate $x_{-j} \sim Q_{-j}^{(j-1)}$ as follows. For each $k < j$, choose an index $I_k \in [m]$ uniformly and independently and set

$$x_k := X_{I_k}^{(k)}.$$

For each $k > j$, sample $x_k \sim P_k$ independently. Let

$$I := \{I_1, \ldots, I_{j-1}\} \subseteq [m].$$

Then $|I| \leq j - 1$.

Let $Z_1, \ldots, Z_m \sim P_j$ be independent fresh samples, independent of all previous randomness, and define

$$Y_r := \begin{cases} X_r^{(j)}, & r \notin I, \\ Z_r, & r \in I. \end{cases}$$

Let $\tilde{P}_j'$ be the empirical distribution of $Y_1, \ldots, Y_m$, i.e., for every measurable $A \subseteq \mathcal{W}_j$,

$$\tilde{P}_j'(A) := \frac{1}{m} \sum_{r=1}^{m} \mathbf{1}\{Y_r \in A\}.$$

We claim that, after the section has been chosen, $Y_1, \ldots, Y_m$ form an i.i.d. sample from P_j. Indeed, the construction of x_{-j} reveals only the indices $I_1, \ldots, I_{j-1}$, the coordinates $X_{I_k}^{(k)}$ for $k < j$, and the independently sampled coordinates x_k for $k > j$. Once this information is fixed, if $r \notin I$, then the whole sample point X_r has not been queried, and hence $X_r^{(j)}$ still has law P_j, independently

of the revealed information. If $r \in I$, then $Y_r = Z_r$, which has law P_j and is independent by construction. Therefore $Y_1, \dots, Y_m$ are independent with common law P_j, conditionally on the chosen section data.

Let

$$\mathcal{G} := \sigma\Big((I_1, \dots, I_{j-1}), (X_{I_k}^{(k)})_{k<j}, (x_k)_{k>j}\Big)$$

denote the information revealed in the construction of x_{-j}.

Conditionally on $\mathcal{G}$, the class $\mathcal{F}|_{x_{-j}}$ is fixed, and $Y_1, \dots, Y_m$ are i.i.d. with common law P_j. Since $\mathrm{VC}(\mathcal{F}|_{x_{-j}}) \le g$, the standard VC inequality yields

$$\mathbb{E}\left[\sup_{F\in\mathcal{F}} \big(P_j(F_{x_{-j}}) - \tilde{P}'_j(F_{x_{-j}})\big) \,\middle|\, \mathcal{G}\right] \le C\sqrt{\frac{g}{m}}$$

for a universal constant $C > 0$.

It remains to return from $\tilde{P}'_j$ to the original empirical marginal $\tilde{P}_j$. The two empirical measures differ only at indices in I, hence in at most $|I| \le j-1$ atoms of mass $1/m$. Therefore

$$\sup_{A\subseteq\mathcal{W}_j} \big|\tilde{P}_j(A) - \tilde{P}'_j(A)\big| \le \frac{j-1}{m}.$$

Consequently, for every chosen x_{-j},

$$\sup_{F\in\mathcal{F}} \big(P_j(F_{x_{-j}}) - \tilde{P}_j(F_{x_{-j}})\big) \le \sup_{F\in\mathcal{F}} \big(P_j(F_{x_{-j}}) - \tilde{P}'_j(F_{x_{-j}})\big) + \frac{j-1}{m}.$$

Combining the previous two displays, we obtain

$$\mathbb{E}\left[\sup_{F\in\mathcal{F}} \big(P_j(F_{x_{-j}}) - \tilde{P}_j(F_{x_{-j}})\big) \,\middle|\, \mathcal{G}\right] \le C\sqrt{\frac{g}{m}} + \frac{j-1}{m}.$$

Finally, taking expectations over all randomness defining $\mathcal{G}$ (in particular over the original sample S, the auxiliary indices $I_1, \dots, I_{j-1}$, the fresh samples $Z_1, \dots, Z_m$, and the independent coordinates $x_k \sim P_k$ for $k > j$), yields

$$\mathbb{E}_S\left[\mathbb{E}_{x_{-j}\sim Q_{-j}^{(j-1)}}\left[\sup_{F\in\mathcal{F}} \big(P_j(F_{x_{-j}}) - \tilde{P}_j(F_{x_{-j}})\big)\right]\right] \le C\sqrt{\frac{g}{m}} + \frac{j-1}{m}.$$

This proves the lemma. ∎

Lemma 16 (Expected deviation of the empirical product estimator) *Under the assumptions of Lemma 15, let*

$$P_\square := P_1 \otimes \cdots \otimes P_d$$

be the product of the marginals of P, and let

$$\tilde{P} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d$$

be the empirical product estimator associated with the sample S.

Then there exists a universal constant $C > 0$ such that

$$\mathbb{E}_S\Big[\sup_{F\in\mathcal{F}} \big(P_\square(F) - \tilde{P}(F)\big)\Big] \le C\, d\sqrt{\frac{g}{m}} + \frac{d(d-1)}{2m}.$$

Proof We interpolate between $P_\square$ and $\tilde{P}$ by replacing the coordinates one at a time. For $j = 0, 1, \ldots, d$, define

$$Q^{(0)} := P_\square,$$

and for $j \geq 1$,

$$Q^{(j)} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_j \otimes P_{j+1} \otimes \cdots \otimes P_d.$$

Thus

$$Q^{(d)} = \tilde{P}.$$

For every $F \in \mathcal{F}$,

$$P_\square(F) - \tilde{P}(F) = \sum_{j=1}^{d} \big(Q^{(j-1)}(F) - Q^{(j)}(F)\big).$$

Taking suprema and using subadditivity,

$$\sup_{F\in\mathcal{F}} \big(P_\square(F) - \tilde{P}(F)\big) \leq \sum_{j=1}^{d} \sup_{F\in\mathcal{F}} \big(Q^{(j-1)}(F) - Q^{(j)}(F)\big).$$

Hence,

$$\mathbb{E}_S\Big[\sup_{F\in\mathcal{F}} \big(P_\square(F) - \tilde{P}(F)\big)\Big] \leq \sum_{j=1}^{d} \mathbb{E}_S\Big[\sup_{F\in\mathcal{F}} \big(Q^{(j-1)}(F) - Q^{(j)}(F)\big)\Big].$$

Fix $j \in [d]$. Let

$$Q_{-j}^{(j-1)} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_{j-1} \otimes P_{j+1} \otimes \cdots \otimes P_d.$$

Then for every $F \in \mathcal{F}$,

$$Q^{(j-1)}(F) = \mathbb{E}_{x_{-j}\sim Q_{-j}^{(j-1)}}\big[P_j(F_{x_{-j}})\big],$$

and

$$Q^{(j)}(F) = \mathbb{E}_{x_{-j}\sim Q_{-j}^{(j-1)}}\big[\tilde{P}_j(F_{x_{-j}})\big].$$

Therefore,

$$\begin{aligned}\sup_{F\in\mathcal{F}} \big(Q^{(j-1)}(F) - Q^{(j)}(F)\big) &= \sup_{F\in\mathcal{F}} \mathbb{E}_{x_{-j}\sim Q_{-j}^{(j-1)}}\Big[P_j(F_{x_{-j}}) - \tilde{P}_j(F_{x_{-j}})\Big] \\ &\leq \mathbb{E}_{x_{-j}\sim Q_{-j}^{(j-1)}}\left[\sup_{F\in\mathcal{F}} \big(P_j(F_{x_{-j}}) - \tilde{P}_j(F_{x_{-j}})\big)\right].\end{aligned}$$

Taking expectation over S, and applying Lemma 15, yields

$$\mathbb{E}_S\Big[\sup_{F\in\mathcal{F}} \big(Q^{(j-1)}(F) - Q^{(j)}(F)\big)\Big] \leq C\sqrt{\frac{g}{m}} + \frac{j-1}{m}.$$

Summing over $j = 1, \ldots, d$, we obtain

$$\mathbb{E}_S\Big[\sup_{F\in\mathcal{F}} \big(P_\square(F) - \tilde{P}(F)\big)\Big] \leq \sum_{j=1}^{d} \left(C\sqrt{\frac{g}{m}} + \frac{j-1}{m}\right).$$

Since

$$\sum_{j=1}^{d} 1 = d, \qquad \sum_{j=1}^{d} (j-1) = \frac{d(d-1)}{2},$$

it follows that

$$\mathbb{E}_S\Big[\sup_{F\in\mathcal{F}}\big(P_{\square}(F) - \tilde{P}(F)\big)\Big] \le C\,d\sqrt{\frac{g}{m}} + \frac{d(d-1)}{2m}.$$

This completes the proof. ■

Remark 17 *We now specialize Lemma 16 to the case where $P = P_1 \otimes \cdots \otimes P_d$ is itself a product distribution. In this case, the proof of Lemma 16 gives the sharper bound*

$$\mathbb{E}_S\Big[\sup_{F\in\mathcal{F}}\big(P(F) - \tilde{P}(F)\big)\Big] \le C\,d\sqrt{\frac{g}{m}},$$

without the correction term (d(d-1)/(2m)). Indeed, that correction term comes from the ghost-replacement step, which is needed only to handle possible dependence between different coordinates of the same sample point. When P is a product measure, the coordinate samples used to form $\tilde{P}_1, \ldots, \tilde{P}_d$ are mutually independent, and hence in each coordinate-replacement step one may apply the usual one-dimensional VC bound directly, conditionally on the other coordinates.

Combining this expectation bound with the McDiarmid-type concentration lemma (Lemma 20) yields a high-probability guarantee and, consequently, a sharper sample-size requirement in the product case (corresponding to the box–continuity modulus $\beta(\alpha) = \alpha$).

To simplify the presentation, we assume throughout this remark that the class $\mathcal{F}$ is closed under complementation. This assumption is without loss of generality, since replacing $\mathcal{F}$ by $\mathcal{F} \cup \{\, \mathcal{X} \setminus F : F \in \mathcal{F}\}$ increases the linear VC dimension by at most a constant factor. Under this assumption,

$$\sup_{F\in\mathcal{F}}\big|P(F) - \tilde{P}(F)\big| = \sup_{F\in\mathcal{F}}\big(P(F) - \tilde{P}(F)\big),$$

so absolute values can be omitted.

Indeed, let

$$\Phi(S) := \sup_{F\in\mathcal{F}}\big(P(F) - \tilde{P}(F)\big),$$

where $\tilde{P} = \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d$ is the empirical product estimator built from $S \sim P^{\otimes m}$. By Lemma 16, there exists a universal constant $C_0 > 0$ such that

$$\mathbb{E}[\Phi(S)] \;\le\; C_0\,d\sqrt{\frac{g}{m}}.$$

Moreover, Lemma 20 implies that for some universal $c > 0$ and all $t > 0$,

$$\Pr[\Phi(S) - \mathbb{E}[\Phi(S)] \ge t] \;\le\; \exp\!\left(-c\,\frac{mt^2}{d^2}\right).$$

Choosing m so that

$$C_0\,d\sqrt{\frac{g}{m}} \le \frac{\varepsilon}{2}, \qquad \sqrt{\frac{d^2\log(1/\delta)}{c\,m}} \le \frac{\varepsilon}{2},$$

yields, with probability at least $1-\delta$,

$$\sup_{F\in\mathcal{F}} |P(F) - \tilde{P}(F)| \;\le\; \varepsilon.$$

In particular, there exists a universal constant $C>0$ *such that it suffices to take*

$$m \;\ge\; C\,\frac{d^2}{\varepsilon^2}\Big(g + \log\tfrac{1}{\delta}\Big).$$

This improves substantially over the general $\mathcal{P}_\beta$ *sample-size bound and captures the favorable behavior of the product estimator when* P *is a product measure.*

B.2. A finite reduction via a random grid

The goal of this subsection is to show that, with high probability, the random grid $G(S)$ from Definition 18 intersects every "large" symmetric difference in $\mathcal{F}\triangle\mathcal{F}$.

Definition 18 (Sample grid) *Let* $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ *and let* $S = (X^{(1)}, \dots, X^{(m)}) \in \mathcal{X}^m$ *be a finite sample. For each* $j \in [d]$*, define the coordinate projection set*

$$\pi_j(S) \;:=\; \{X_j^{(1)}, \dots, X_j^{(m)}\} \subseteq \mathcal{W}_j,$$

and define the associated (empirical) product grid

$$G(S) \;:=\; \pi_1(S) \times \cdots \times \pi_d(S) \subseteq \mathcal{X}.$$

We next combine the bound on the linear VC dimension of the symmetric–difference class $\mathcal{H} = \mathcal{F}\triangle\mathcal{F}$ given by Lemma 34 with the expected deviation bound for the empirical product estimator. This allows us to control deviations of the form $P(F) - \tilde{P}(F')$ uniformly over pairs $F, F' \in \mathcal{F}$ by reducing them to deviations over $\mathcal{H}$.

Lemma 19 (Expected product deviation for class of symmetric differences) *Let* $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ *and* $\mathcal{F} \subseteq 2^{\mathcal{X}}$ *with* $\mathrm{LVC}(\mathcal{F}) = g < \infty$*, and set* $\mathcal{H} := \mathcal{F}\triangle\mathcal{F}$*. Let* P *be any distribution on* $\mathcal{X}$*, write* $P_\square := P_1 \otimes \cdots \otimes P_d$*, and let* $\tilde{P}$ *be the empirical product estimator built from* m *samples as in Lemma 16. Then there exists a universal constant* $C>0$ *such that*

$$\mathbb{E}_{S\sim P^m}\Big[\sup_{H\in\mathcal{H}}\big(P_\square(H) - \tilde{P}(H)\big)\Big] \;\le\; C\,d\sqrt{\frac{\mathrm{LVC}(\mathcal{H})}{m}} \;+\; \frac{d(d-1)}{2m} \;\le\; C\,d\sqrt{\frac{20g}{m}} \;+\; \frac{d(d-1)}{2m}.$$

Proof Apply Lemma 16 to the class $\mathcal{H}$ in place of $\mathcal{F}$. By definition of the empirical product estimator and the bound established in Lemma 16, we have:

$$\mathbb{E}_{S\sim P^m}\Big[\sup_{H\in\mathcal{H}}\big(P_\square(H) - \tilde{P}(H)\big)\Big] \;\le\; C\,d\sqrt{\frac{\mathrm{LVC}(\mathcal{H})}{m}} \;+\; \frac{d(d-1)}{2m}.$$

Finally, Lemma 34 gives $\mathrm{LVC}(\mathcal{H}) \le 20\,\mathrm{LVC}(\mathcal{F}) = 20g$, and the result follows. ∎

Lemma 20 (McDiarmid for the product supremum) *Let P be any distribution on $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, and write $P_\square := P_1 \otimes \cdots \otimes P_d$ for the product of its marginals. Let $\mathcal{H} \subseteq 2^{\mathcal{X}}$ be any class. For $S = (X_1, \ldots, X_m) \sim P^m$, let $\tilde{P}$ be the empirical product estimator based on S, and define*

$$\Phi(S) := \sup_{H \in \mathcal{H}} \big(P_\square(H) - \tilde{P}(H)\big).$$

Then for all $t > 0$,

$$\Pr\big[\Phi(S) - \mathbb{E}[\Phi(S)] \geq t\big] \;\leq\; \exp\Big(-\frac{2mt^2}{d^2}\Big).$$

Proof Let S and S' be two samples that differ only in the k-th point, and let $\tilde{P} = \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d$ and $\tilde{P}' = \tilde{P}'_1 \otimes \cdots \otimes \tilde{P}'_d$ be the corresponding product estimators. Changing a single observation affects each empirical marginal by at most $1/m$ in total variation:

$$\|\tilde{P}_j - \tilde{P}'_j\|_{\mathrm{TV}} \leq \frac{1}{m}, \qquad j = 1, \ldots, d.$$

Using the standard product bound

$$\big\|\bigotimes_{j=1}^d \mu_j - \bigotimes_{j=1}^d \nu_j\big\|_{\mathrm{TV}} \leq \textstyle\sum_{j=1}^d \|\mu_j - \nu_j\|_{\mathrm{TV}},$$

we obtain

$$\|\tilde{P} - \tilde{P}'\|_{\mathrm{TV}} \leq \sum_{j=1}^d \|\tilde{P}_j - \tilde{P}'_j\|_{\mathrm{TV}} \leq \frac{d}{m}.$$

Therefore, for every measurable $H \subseteq \mathcal{X}$,

$$|\tilde{P}(H) - \tilde{P}'(H)| \leq \|\tilde{P} - \tilde{P}'\|_{\mathrm{TV}} \leq \frac{d}{m}.$$

Since $P_\square$ does not depend on the sample, it follows that

$$\begin{aligned} |\Phi(S) - \Phi(S')| &\leq \sup_{H \in \mathcal{H}} \Big|\big(P_\square(H) - \tilde{P}(H)\big) - \big(P_\square(H) - \tilde{P}'(H)\big)\Big| \\ &= \sup_{H \in \mathcal{H}} \big|\tilde{P}(H) - \tilde{P}'(H)\big| \leq \frac{d}{m}. \end{aligned}$$

Thus Φ has bounded differences with constants $c_k = d/m$ for $k = 1, \ldots, m$. McDiarmid's inequality yields, for all $t > 0$,

$$\Pr\big[\Phi(S) - \mathbb{E}\Phi(S) \geq t\big] \leq \exp\Big(-\frac{2t^2}{\sum_{k=1}^m (d/m)^2}\Big) = \exp\Big(-\frac{2mt^2}{d^2}\Big),$$

as claimed. ∎

Corollary 21 (High-probability bound for $\mathcal{H} = \mathcal{F} \triangle \mathcal{F}$) *There exists a universal constant $C > 0$ such that the following holds. Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ and let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ satisfy $\mathrm{LVC}(\mathcal{F}) = g < \infty$,*

and set $\mathcal{H} := \mathcal{F} \triangle \mathcal{F}$. *Then for every distribution* P *on* $\mathcal{X}$, *writing* $P_\square := P_1 \otimes \cdots \otimes P_d$, *and every* $\varepsilon, \delta \in (0,1)$, *if*

$$m \;\geq\; C\,\frac{d^2}{\varepsilon^2}\Big(g + \log\tfrac{1}{\delta}\Big),$$

then with probability at least $1-\delta$ *over* $S \sim P^m$,

$$\sup_{H \in \mathcal{H}} \big(P_\square(H) - \tilde{P}(H)\big) \;\leq\; \varepsilon.$$

Proof Define

$$\Phi(S) := \sup_{H \in \mathcal{H}} \big(P_\square(H) - \tilde{P}(H)\big).$$

By Lemma 19, we have the expectation bound:

$$\mathbb{E}[\Phi(S)] \;\leq\; C_0\, d\,\sqrt{\frac{\mathrm{LVC}(\mathcal{H})}{m}} \;+\; \frac{d(d-1)}{2m} \;\leq\; C_0\, d\,\sqrt{\frac{20g}{m}} \;+\; \frac{d(d-1)}{2m},$$

for a universal constant $C_0 > 0$. Moreover, by Lemma 20, for all $t > 0$,

$$\Pr\big[\Phi(S) \geq \mathbb{E}[\Phi(S)] + t\big] \;\leq\; \exp\Big(-\frac{2mt^2}{d^2}\Big).$$

Taking $t = \varepsilon/2$, we get

$$\Pr\big[\Phi(S) > \varepsilon\big] \;\leq\; \Pr\big[\Phi(S) > \mathbb{E}[\Phi(S)] + \tfrac{\varepsilon}{2}\big] \;\leq\; \exp\Big(-\frac{m\varepsilon^2}{2d^2}\Big),$$

provided that $\mathbb{E}[\Phi(S)] \leq \varepsilon/2$. Thus it suffices that

$$C_0\, d\,\sqrt{\frac{20g}{m}} \;+\; \frac{d(d-1)}{2m} \leq \frac{\varepsilon}{2} \qquad \text{and} \qquad \exp\Big(-\frac{m\varepsilon^2}{2d^2}\Big) \leq \delta.$$

To satisfy the expectation condition, it is sufficient to bound each term by $\varepsilon/4$:

$$C_0\, d\,\sqrt{\frac{20g}{m}} \leq \frac{\varepsilon}{4} \qquad \text{and} \qquad \frac{d^2}{2m} \geq \frac{d(d-1)}{2m} \leq \frac{\varepsilon}{4},$$

i.e., we require

$$m \;\geq\; 320\, C_0^2\,\frac{d^2 g}{\varepsilon^2}, \qquad m \;\geq\; 2\,\frac{d^2}{\varepsilon}, \qquad \text{and} \qquad m \;\geq\; 2\,\frac{d^2}{\varepsilon^2}\log\tfrac{1}{\delta}.$$

Since $\varepsilon \in (0,1)$ implies $1/\varepsilon^2 > 1/\varepsilon$, all three conditions are implied by

$$m \;\geq\; C\,\frac{d^2}{\varepsilon^2}\Big(g + \log\tfrac{1}{\delta}\Big)$$

for a sufficiently large universal constant $C > 0$, completing the proof. ∎

Lemma 22 (Hitting large symmetric differences for $P \in \mathcal{P}_\beta$) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ be a product space, and let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ satisfy $\mathrm{LVC}(\mathcal{F}) = g < \infty$. Set $\mathcal{H} := \mathcal{F} \triangle \mathcal{F}$.*

There exists a universal constant $C_0 > 0$ such that the following holds. For every $\varepsilon, \delta \in (0,1)$, every $P \in \mathcal{P}_\beta$, and every i.i.d. sample $S \sim P^m$, if

$$m \;\geq\; C_0 \frac{d^2}{\beta(\varepsilon)^2}\Big(g + \log \tfrac{1}{\delta}\Big),$$

then with probability at least $1-\delta$ over S,

$$\forall H \in \mathcal{H} \text{ with } P(H) \geq \varepsilon : \qquad H \cap G(S) \neq \emptyset,$$

where $G(S)$ denotes the empirical product grid induced by S. Equivalently, with probability at least $1-\delta$, for all $F, F' \in \mathcal{F}$,

$$P(F \triangle F') \geq \varepsilon \quad \Longrightarrow \quad (F \triangle F') \cap G(S) \neq \emptyset.$$

Proof Let $P_\square := P_1 \otimes \cdots \otimes P_d$ be the product of the marginals of P, and set $\eta := \beta(\varepsilon)$.

Applying Corollary 21 to the class

$$\mathcal{H} = \mathcal{F} \triangle \mathcal{F}$$

with accuracy $\eta/2$ and confidence δ, we obtain that there exists a universal constant $C_0 > 0$ such that, if

$$m \;\geq\; C_0 \frac{d^2}{\eta^2}\Big(g + \log \tfrac{1}{\delta}\Big),$$

then with probability at least $1-\delta$ over $S \sim P^m$,

$$\sup_{H \in \mathcal{H}} \big(P_\square(H) - \tilde{P}(H)\big) \;\leq\; \frac{\eta}{2}, \tag{6}$$

where

$$\tilde{P} = \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d$$

is the empirical product estimator associated with S.

Now work on the event (6), and fix any $H \in \mathcal{H}$ satisfying

$$P(H) \geq \varepsilon.$$

Since $P \in \mathcal{P}_\beta$, by the definition of $\mathcal{P}_\beta$ we have

$$P_\square(H) \geq \beta(\varepsilon) = \eta.$$

Combining this with (6) yields

$$\tilde{P}(H) \;\geq\; P_\square(H) - \sup_{K \in \mathcal{H}} \big(P_\square(K) - \tilde{P}(K)\big) \;\geq\; \eta - \frac{\eta}{2} \;=\; \frac{\eta}{2} \;>\; 0.$$

Since $\tilde{P}$ is supported on the empirical product grid $G(S)$, the inequality $\tilde{P}(H) > 0$ implies that

$$H \cap G(S) \neq \emptyset.$$

Thus, with probability at least $1-\delta$, every $H \in \mathcal{H}$ satisfying $P(H) \geq \varepsilon$ intersects the grid $G(S)$.

Finally, since

$$\mathcal{H} = \mathcal{F} \triangle \mathcal{F} = \{F \triangle F' : F, F' \in \mathcal{F}\},$$

this is equivalent to saying that for all $F, F' \in \mathcal{F}$,

$$P(F \triangle F') \geq \varepsilon \quad \Longrightarrow \quad (F \triangle F') \cap G(S) \neq \emptyset.$$

This completes the proof. ■

Appendix C. Hitting large symmetric differences, alternative bound

This section proves an alternative version of Lemma 22, which is stated as Lemma 27. The main advantage of this alternative argument is that it improves the dependence on ε: instead of a quadratic dependence through $1/\beta(\varepsilon)^2$, it yields a linear dependence through $1/\beta(\varepsilon)$.

The proof is based on a relative VC inequality. This inequality allows us to control the mass of sections more efficiently than the standard uniform VC bound, and this improvement is what ultimately leads to the better dependence on ε.

For a finite sequence $x_1^N = (x_1, \ldots, x_N)$ and a class $\mathcal{F}$ of $\{0,1\}$-valued functions, define

$$\mathbb{S}_{\mathcal{F}}(x_1^N) := \left|\{(f(x_1), \ldots, f(x_N)) : f \in \mathcal{F}\}\right|.$$

Thus $\mathbb{S}_{\mathcal{F}}(x_1^N)$ is the number of distinct labelings of $x_1, \ldots, x_N$ induced by $\mathcal{F}$.

Theorem 23 (Relative VC inequality (Boucheron et al., 2005, Theorem 5.1)) *Let $X_1, \ldots, X_{2n}$ be i.i.d. random variables with law P, and let $\mathcal{F}$ be a class of $\{0,1\}$-valued measurable functions. We call*

$$X_1^{2n} := (X_1, \ldots, X_n, X_{n+1}, \ldots, X_{2n})$$

the double sample, where $X_{n+1}, \ldots, X_{2n}$ play the role of an independent sample. For $f \in \mathcal{F}$, write

$$Pf := \mathbb{E}[f(X_1)], \qquad P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

so P_n is formed only from the first n sample points.

Then for every $\delta \in (0,1)$, with probability at least $1-\delta$, every $f \in \mathcal{F}$ satisfies

$$\frac{Pf - P_n f}{\sqrt{Pf}} \leq 2\sqrt{\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log(4/\delta)}{n}}.$$

Equivalently, with the same probability,

$$Pf \leq P_n f + 2\sqrt{Pf\,\frac{\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) + \log(4/\delta)}{n}}.$$

Corollary 24 (Linearized relative VC inequality) *Under the assumptions of Theorem 23, assume in addition that* $\mathrm{VC}(\mathcal{F}) \le g$. *Then, for every* $r > 0$, *with probability at least* $1-\delta$, *every* $f \in \mathcal{F}$ *satisfies*

$$Pf \le (1+r)P_n f + \left(4 + \frac{1}{r}\right)\frac{g\log(2n+1) + \log(4/\delta)}{n}.$$

Proof By Lemma 30, applied to the set system induced by the supports of functions in $\mathcal{F}$, we have for every realization of the double sample

$$\mathbb{S}_{\mathcal{F}}(X_1^{2n}) \le \Pi_{\mathcal{F}}(2n) \le (2n+1)^g.$$

Hence

$$\log \mathbb{S}_{\mathcal{F}}(X_1^{2n}) \le g\log(2n+1).$$

Set

$$a := \frac{g\log(2n+1) + \log(4/\delta)}{n}.$$

By Theorem 23, with probability at least $1-\delta$, every $f \in \mathcal{F}$ satisfies

$$Pf \le P_n f + 2\sqrt{a\,Pf}.$$

We use the elementary implication

$$A \le C + B\sqrt{A} \quad \Longrightarrow \quad A \le C + B\sqrt{C} + B^2.$$

Applying it with $A = Pf$, $C = P_n f$, and $B = 2\sqrt{a}$, we get

$$Pf \le P_n f + 2\sqrt{aP_n f} + 4a.$$

By Cauchy's inequality with parameter $r > 0$,

$$2\sqrt{aP_n f} \le rP_n f + \frac{a}{r}.$$

Therefore

$$Pf \le (1+r)P_n f + \left(4 + \frac{1}{r}\right)a,$$

and substituting the definition of a completes the proof. ■

For $j \in [d]$, define

$$Q^{(j)} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_j \otimes P_{j+1} \otimes \cdots \otimes P_d.$$

Thus

$$Q^{(0)} = P_\square := P_1 \otimes \cdots \otimes P_d, \qquad Q^{(d)} = \tilde{P} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d.$$

The next two results are high-probability analogues of Lemma 15 and Lemma 16. In the earlier argument, Lemma 15 controlled a single coordinate replacement in expectation, and Lemma 16 obtained the full product bound by telescoping over the coordinates. We follow the same structure here, but use the linearized relative VC inequality in place of the standard VC bound. This is the step that will later improve the dependence on $\beta(\varepsilon)$.

Lemma 25 (High-probability one-coordinate transition) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ satisfy $\mathrm{LVC}(\mathcal{F}) = g < \infty$, and let P be a probability measure on $\mathcal{X}$ with marginals $P_1, \dots, P_d$. Let $S = (X_1, \dots, X_m) \sim P^m$, and let $\tilde{P}_i$ be the empirical marginal on coordinate i.*

For $0 \le j \le d$, define

$$Z_j(S) := \sup_{F \in \mathcal{F}:\, \tilde{P}(F)=0} Q^{(j)}(F).$$

Fix $j \in [d]$. Then there exists a universal constant $C > 0$ such that, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$Z_{j-1}(S) \le \left(1 + \frac{1}{d}\right) Z_j(S) + Cd\,\Lambda_{m,j}(g,\delta),$$

where

$$\Lambda_{m,j}(g,\delta) := \frac{g\log(2m+1) + j\log m + \log(4/\delta)}{m}.$$

Proof We may assume $m \ge 2j$, since otherwise the claim is trivial after increasing the universal constant C.

Fix $j \in [d]$. We use the notation of Definition 13. Since $\mathrm{LVC}(\mathcal{F}) = g$, every section class $\mathcal{F}|_{x_{-j}}$ has VC dimension at most g. The goal is to compare $Q^{(j-1)}(F)$ and $Q^{(j)}(F)$ uniformly over those $F \in \mathcal{F}$ satisfying $\tilde{P}(F) = 0$. Equivalently, we need to compare P_j and $\tilde{P}_j$ uniformly over sections $\mathcal{F}|_{x_{-j}}$, where

$$x_{-j} \sim \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_{j-1} \otimes P_{j+1} \otimes \cdots \otimes P_d.$$

The only point requiring care is that the first $j-1$ coordinates of x_{-j} are chosen from the same sample S. We remove this dependence by discarding the few j-coordinates whose sample points were used to construct the prefix.

Generate x_{-j} as follows. Choose a prefix tuple

$$I = (I_1, \dots, I_{j-1}) \in [m]^{j-1}$$

uniformly, set

$$U_I := (X^{(1)}_{I_1}, \dots, X^{(j-1)}_{I_{j-1}}),$$

and independently sample

$$z = (z_{j+1}, \dots, z_d) \sim P_{j+1} \otimes \cdots \otimes P_d.$$

Condition on the chosen prefix tuple I and on the values of the prefix coordinates U_I. After this conditioning, the prefix U_I is fixed. Now fix a suffix z. For each $F \in \mathcal{F}$, consider the corresponding section

$$\{x_j \in \mathcal{W}_j : (U_I, x_j, z) \in F\}.$$

As F ranges over $\mathcal{F}$, these sections form a class of subsets of $\mathcal{W}_j$ with VC dimension at most g.

The only sample points whose non-j coordinates were revealed are those with indices in $R_I = \{I_1, \dots, I_{j-1}\}$. Therefore, after removing these indices, the coordinates $(X^{(j)}_\ell)_{\ell \notin R_I}$ are still i.i.d.

with law P_j, independently of the above fixed class of sections. Define the empirical marginal on coordinate j, with these exposed sample points removed, by

$$\tilde{P}_{j,-I}(A) := \frac{1}{m - |R_I|} \sum_{\ell \notin R_I} \mathbf{1}\{X_\ell^{(j)} \in A\}, \qquad A \subseteq \mathcal{W}_j.$$

Thus Corollary 24 applies conditionally to this fixed class of sections, with distribution P_j and empirical measure $\tilde{P}_{j,-I}$.

Set

$$\eta := \frac{\delta}{m^j}, \qquad P_{>j} := P_{j+1} \otimes \cdots \otimes P_d.$$

Fix a prefix tuple I, and condition on the values of U_I. Now fix a suffix z. As explained above, after removing the indices in R_I, the points

$$(X_\ell^{(j)})_{\ell \notin R_I}$$

form an i.i.d. sample from P_j, independent of the fixed family of sections

$$\big\{\{x_j : (U_I, x_j, z) \in F\} : F \in \mathcal{F}\big\}.$$

This family has VC dimension at most g. Therefore, applying Corollary 24 with $r = 1/d$, we get that, with probability at least $1 - \eta$, the bound

$$P_j\big(\{x_j : (U_I, x_j, z) \in F\}\big) \le \left(1 + \frac{1}{d}\right) \tilde{P}_{j,-I}\big(\{x_j : (U_I, x_j, z) \in F\}\big) + Cd\,\Lambda_{m,j}(g, \delta)$$

holds simultaneously for all $F \in \mathcal{F}$. Here we used $r = 1/d$ and increased the universal constant C, since

$$d\,\frac{g\log(2m+1) + \log(4m^j/\delta)}{m} \le d\,\Lambda_{m,j}(g, \delta).$$

Let $B_I(S)$ be the set of suffixes z for which this simultaneous bound fails. Since each fixed suffix is bad with probability at most η, averaging over the suffix $z \sim P_{>j}$ gives

$$\mathbb{E}\big[P_{>j}(B_I(S))\big] = \int \Pr(z \in B_I(S))\, dP_{>j}(z) \le \eta.$$

Hence, by Markov's inequality,

$$\Pr\left(P_{>j}(B_I(S)) > \frac{1}{m}\right) \le m\eta.$$

Taking a union bound over all $I \in [m]^{j-1}$, we obtain that with probability at least

$$1 - m^{j-1} \cdot m\eta \ge 1 - \delta,$$

for every prefix tuple I, the set of bad suffixes has $P_{>j}$-measure at most $1/m$. We work on this event.

Fix $F \in \mathcal{F}$ and a prefix tuple I. Integrating over the good suffixes and paying at most $1/m$ for the bad suffixes gives

$$\int P_j\big(\{x_j : (U_I, x_j, z) \in F\}\big)\, dP_{>j}(z) \\ \leq \left(1 + \frac{1}{d}\right) \int \tilde{P}_{j,-I}\big(\{x_j : (U_I, x_j, z) \in F\}\big)\, dP_{>j}(z) + Cd\,\Lambda_{m,j}(g,\delta) + \frac{1}{m}.$$

It remains to replace $\tilde{P}_{j,-I}$ by $\tilde{P}_j$. Since $\tilde{P}_{j,-I}$ is obtained from $\tilde{P}_j$ by removing at most $|R_I| \leq j-1$ atoms, for every measurable $A \subseteq \mathcal{W}_j$,

$$\big|\tilde{P}_{j,-I}(A) - \tilde{P}_j(A)\big| \leq \frac{|R_I|}{m} \leq \frac{j-1}{m}.$$

Therefore,

$$\int P_j\big(\{x_j : (U_I, x_j, z) \in F\}\big)\, dP_{>j}(z) \\ \leq \left(1 + \frac{1}{d}\right) \int \tilde{P}_j\big(\{x_j : (U_I, x_j, z) \in F\}\big)\, dP_{>j}(z) \\ + Cd\,\Lambda_{m,j}(g,\delta) + \frac{1}{m} + \left(1 + \frac{1}{d}\right)\frac{j-1}{m}.$$

Since $d \geq 1$, the last two terms are bounded by Cj/m, after increasing the universal constant C.

Now average over $I \in [m]^{j-1}$. Averaging over I is exactly integration with respect to

$$\tilde{P}_1 \otimes \cdots \otimes \tilde{P}_{j-1}.$$

Thus, simultaneously for every $F \in \mathcal{F}$,

$$Q^{(j-1)}(F) \leq \left(1 + \frac{1}{d}\right) Q^{(j)}(F) + Cd\,\Lambda_{m,j}(g,\delta) + C\frac{j}{m}.$$

Since $m \geq 2$,

$$\frac{j}{m} \leq C\frac{j \log m}{m} \leq C\Lambda_{m,j}(g,\delta),$$

and hence this term is absorbed into $Cd\,\Lambda_{m,j}(g,\delta)$, after another adjustment of the universal constant.

Finally, taking the supremum over all $F \in \mathcal{F}$ such that $\tilde{P}(F) = 0$ gives

$$Z_{j-1}(S) \leq \left(1 + \frac{1}{d}\right) Z_j(S) + Cd\,\Lambda_{m,j}(g,\delta).$$

This proves the lemma.

■

Corollary 26 (High-probability empirical product bound) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ satisfy $\mathrm{LVC}(\mathcal{F}) = g < \infty$, and let P be a probability measure on $\mathcal{X}$ with marginals $P_1, \ldots, P_d$. Let $S = (X_1, \ldots, X_m) \sim P^m$, and define*

$$P_\square := P_1 \otimes \cdots \otimes P_d, \qquad \tilde{P} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d.$$

Assume $m \geq 2$. Then there exists a universal constant $C > 0$ such that for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{F \in \mathcal{F}:\, \tilde{P}(F)=0} P_{\square}(F) \leq C d^2 \frac{(g+d)\log m + \log(1/\delta)}{m}.$$

Proof For $j = 0, \dots, d$, define

$$Q^{(j)} := \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_j \otimes P_{j+1} \otimes \cdots \otimes P_d,$$

and

$$Z_j(S) := \sup_{F \in \mathcal{F}:\, \tilde{P}(F)=0} Q^{(j)}(F).$$

Then

$$Z_0(S) = \sup_{F \in \mathcal{F}:\, \tilde{P}(F)=0} P_{\square}(F), \qquad Z_d(S) = 0.$$

Apply Lemma 25 with failure probability δ/d, and take a union bound over $j = 1, \dots, d$. With probability at least $1-\delta$, for every $j \in [d]$,

$$Z_{j-1}(S) \leq \left(1 + \frac{1}{d}\right) Z_j(S) + C d\, \Lambda_{m,j}(g, \delta/d).$$

Since $\Lambda_{m,j}(g, \delta/d)$ is increasing in j,

$$\Lambda_{m,j}(g, \delta/d) \leq \Lambda_{m,d}(g, \delta/d) \qquad \forall j \in [d].$$

Thus, on the same event,

$$Z_{j-1}(S) \leq \left(1 + \frac{1}{d}\right) Z_j(S) + C d\, \Lambda_*, \qquad \Lambda_* := \Lambda_{m,d}(g, \delta/d).$$

By Lemma 28, using $Z_d(S) = 0$,

$$Z_0(S) \leq C d^2 \Lambda_*.$$

Finally,

$$\Lambda_* = \frac{g \log(2m+1) + d \log m + \log(4d/\delta)}{m}.$$

Since $m \geq 2$,

$$\log(2m+1) \leq C \log m, \qquad \log(4d/\delta) \leq C\big(\log d + \log(1/\delta)\big).$$

Also,

$$\log d \leq d \log m.$$

Therefore,

$$\Lambda_* \leq C \frac{(g+d)\log m + \log(1/\delta)}{m}.$$

Hence

$$Z_0(S) \leq C d^2 \frac{(g+d)\log m + \log(1/\delta)}{m}.$$

This proves the corollary. ∎

Lemma 27 (Hitting large symmetric differences for $P \in \mathcal{P}_\beta$) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, and let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ satisfy $\mathrm{LVC}(\mathcal{F}) = g < \infty$. Define*

$$\mathcal{H} := \mathcal{F} \triangle \mathcal{F}.$$

There exists a universal constant $C_0 > 0$ such that the following holds. For every $\varepsilon, \delta \in (0,1)$, every $P \in \mathcal{P}_\beta$, and every i.i.d. sample $S \sim P^m$, if

$$m \geq C_0 \frac{d^2}{\beta(\varepsilon)} \left((g+d) \log\left(\frac{d^2(g+d)}{\beta(\varepsilon)} \right) + \log \frac{1}{\delta} \right),$$

then with probability at least $1-\delta$,

$$\forall H \in \mathcal{H} \text{ with } P(H) \geq \varepsilon, \qquad H \cap G(S) \neq \emptyset.$$

Equivalently, with probability at least $1-\delta$, for all $F, F' \in \mathcal{F}$,

$$P(F \triangle F') \geq \varepsilon \quad \Longrightarrow \quad (F \triangle F') \cap G(S) \neq \emptyset.$$

Proof Let

$$P_\square := P_1 \otimes \cdots \otimes P_d, \qquad \eta := \beta(\varepsilon).$$

Since $P \in \mathcal{P}_\beta$, every measurable $H \subseteq \mathcal{X}$ satisfies

$$P(H) \geq \varepsilon \quad \Longrightarrow \quad P_\square(H) \geq \eta.$$

Let $\tilde{P} = \tilde{P}_1 \otimes \cdots \otimes \tilde{P}_d$. Since $\tilde{P}$ is supported exactly on the empirical product grid $G(S)$,

$$\tilde{P}(H) = 0 \quad \Longleftrightarrow \quad H \cap G(S) = \emptyset.$$

Therefore,

$$\left\{ \exists H \in \mathcal{H} : \; P(H) \geq \varepsilon, \; H \cap G(S) = \emptyset \right\} \subseteq \left\{ \sup_{H \in \mathcal{H} : \, \tilde{P}(H) = 0} P_\square(H) \geq \eta \right\}.$$

By Lemma 34,

$$\mathrm{LVC}(\mathcal{H}) \leq Cg$$

for a universal constant $C > 0$. Applying Corollary 26 to $\mathcal{H}$, we get that with probability at least $1-\delta$,

$$\sup_{H \in \mathcal{H} : \, \tilde{P}(H) = 0} P_\square(H) \leq C_1 d^2 \frac{(g+d)\log m + \log(1/\delta)}{m}.$$

Thus it suffices to ensure

$$C_1 d^2 \frac{(g+d)\log m + \log(1/\delta)}{m} \leq \eta.$$

Equivalently, it suffices that

$$m \geq K\big((g+d)\log m + \log(1/\delta)\big), \qquad K := \frac{C_1 d^2}{\eta}.$$

We claim that this is guaranteed by the displayed lower bound on m, after increasing the universal constant. Indeed, if the desired inequality failed, then

$$m \le K \log(1/\delta) + K(g+d)\log m.$$

Applying Lemma 29 with

$$x = m, \qquad a = K\log(1/\delta), \qquad b = K(g+d),$$

and increasing constants so that $b \ge 1$, gives

$$m \le 2K\log(1/\delta) + 4K(g+d)\log\big(4K(g+d)\big).$$

Since $K = C_1 d^2/\eta$, the right-hand side is at most

$$C_2 \frac{d^2}{\eta}\left((g+d)\log\left(\frac{d^2(g+d)}{\eta}\right) + \log\frac{1}{\delta}\right).$$

Therefore, choosing $C_0 > C_2$ contradicts the assumed lower bound on m. Hence

$$C_1 d^2 \frac{(g+d)\log m + \log(1/\delta)}{m} < \eta.$$

On this event,

$$\sup_{H\in\mathcal{H}:\,\tilde{P}(H)=0} P_{\square}(H) < \eta,$$

so no $H \in \mathcal{H}$ with $P(H) \ge \varepsilon$ can satisfy $\tilde{P}(H) = 0$. Equivalently, every such H intersects $G(S)$.

Since

$$\mathcal{H} = \mathcal{F}\triangle\mathcal{F} = \{F\triangle F' : F, F' \in \mathcal{F}\},$$

the final formulation follows. ■

Appendix D. Technical proofs

Lemma 28 (Iterating the one-step recurrence) *Let $B_0, \dots, B_d \ge 0$ satisfy $B_d = 0$, and suppose that for some $\Gamma \ge 0$,*

$$B_{j-1} \le \left(1 + \frac{1}{d}\right) B_j + \Gamma \qquad \forall j \in [d].$$

Then

$$B_0 \le ed\,\Gamma.$$

Proof Iterating the recurrence gives

$$B_0 \le \Gamma \sum_{j=0}^{d-1} \left(1 + \frac{1}{d}\right)^j.$$

Since

$$\left(1 + \frac{1}{d}\right)^j \le \left(1 + \frac{1}{d}\right)^d < e,$$

we obtain

$$B_0 \leq ed\,\Gamma.$$

■

Lemma 29 (Logarithmic self-bound) *Let $x \geq 1$, $a \geq 0$, and $b \geq 1$. If*

$$x \leq a + b \log x,$$

then

$$x \leq 2a + 4b \log(4b).$$

Proof If $x \leq 2a$, then immediately

$$x \leq 2a \leq 2a + 4b \log(4b).$$

It remains to consider the case $x > 2a$. Then $a < x/2$, and therefore

$$x \leq a + b \log x < \frac{x}{2} + b \log x.$$

Hence $x < 2b \log x$. Set $B := 2b$. Since $b \geq 1$, we have $B \geq 2$, and the last inequality becomes

$$x < B \log x.$$

We now claim that $x \leq 2B \log(2B)$. Suppose, toward a contradiction, that $x > 2B \log(2B)$. Since $B \geq 2$, we have $2B \log(2B) \geq 4 \log 4 > e$, and hence $x > e$. The function $t \mapsto \log t/t$ is decreasing for all $t \geq e$, so

$$\frac{\log x}{x} < \frac{\log(2B \log(2B))}{2B \log(2B)}.$$

Also,

$$\log(2B \log(2B)) \leq 2 \log(2B),$$

because this is equivalent to

$$2B \log(2B) \leq (2B)^2,$$

or equivalently $\log(2B) \leq 2B$, which follows from the elementary inequality $\log u \leq u$ for $u > 0$. Therefore

$$\frac{\log x}{x} < \frac{2 \log(2B)}{2B \log(2B)} = \frac{1}{B}.$$

Multiplying by Bx, we get $B \log x < x$, contradicting $x < B \log x$. Hence $x \leq 2B \log(2B)$.

Substituting $B = 2b$, we obtain

$$x \leq 4b \log(4b).$$

Together with the first case $x \leq 2a$, this gives

$$x \leq 2a + 4b \log(4b).$$

■

Lemma 30 (VC growth bound) *Let $\mathcal{H} \subseteq 2^{\mathcal{X}}$ satisfy $\mathrm{VC}(\mathcal{H}) \leq g < \infty$. For $m \geq 1$, define*

$$\Pi_{\mathcal{H}}(m) := \sup_{x_1,\ldots,x_m \in \mathcal{X}} \left| \left\{ \left(\mathbf{1}\{x_1 \in H\}, \ldots, \mathbf{1}\{x_m \in H\} \right) : H \in \mathcal{H} \right\} \right|.$$

Then

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{\min\{g,m\}} \binom{m}{i}.$$

In particular,

$$\Pi_{\mathcal{H}}(m) \leq (m+1)^g.$$

Moreover, if $1 \leq g \leq m$, then

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{g} \right)^g.$$

Proof Fix $x_1, \ldots, x_m \in \mathcal{X}$. By the Sauer–Shelah lemma, the number of distinct labelings of $x_1, \ldots, x_m$ induced by $\mathcal{H}$ is at most

$$\sum_{i=0}^{\min\{g,m\}} \binom{m}{i}.$$

Taking the supremum over $x_1, \ldots, x_m$ gives

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{\min\{g,m\}} \binom{m}{i}.$$

We next prove the bound $\Pi_{\mathcal{H}}(m) \leq (m+1)^g$. If $g = 0$, this is immediate. Assume $g \geq 1$. The sum

$$\sum_{i=0}^{\min\{g,m\}} \binom{m}{i}$$

counts the number of subsets of $[m]$ of size at most g. Each such subset $\{a_1 < \cdots < a_k\}$, with $k \leq g$, can be encoded injectively by the g-tuple

$$(a_1, \ldots, a_k, 0, \ldots, 0) \in \{0, 1, \ldots, m\}^g.$$

Therefore

$$\sum_{i=0}^{\min\{g,m\}} \binom{m}{i} \leq (m+1)^g.$$

Finally, when $1 \leq g \leq m$, Lemma 37 gives

$$\sum_{i=0}^{g} \binom{m}{i} = \binom{m}{\leq g} \leq \left(\frac{em}{g} \right)^g.$$

This completes the proof. ∎

D.1. VC dimension of symmetric differeces

Definition 31 (Symmetric Difference) *For two sets $A, B \subseteq \mathcal{X}$, their* symmetric difference *is defined by*

$$A \Delta B \;:=\; (A \setminus B) \cup (B \setminus A) \;=\; \{\, x \in \mathcal{X} : \mathbf{1}_A(x) \neq \mathbf{1}_B(x) \,\}.$$

For a family of sets $\mathcal{F} \subseteq 2^{\mathcal{X}}$ we denote

$$\mathcal{F} \Delta \mathcal{F} \;:=\; \{\, A \Delta B : A, B \in \mathcal{F} \,\},$$

that is, the family of all pairwise symmetric differences between members of $\mathcal{F}$.

Theorem 32 (Aggregation bound, (Alon et al., 2023)) *Let $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a base class, and let $\mathcal{G}$ be a class of aggregation rules $g : \{\pm 1\}^T \to \{\pm 1\}$. Then*

$$\mathrm{VC}\big(\{x \mapsto g(b_1(x), \ldots, b_T(x)) : b_i \in \mathcal{B},\, g \in \mathcal{G}\}\big) \;\leq\; c_T\big(T \cdot \mathrm{VC}(\mathcal{B}) + \mathrm{VC}(\mathcal{G})\big),$$

where

$$c_T \;=\; \frac{1}{T \cdot \eta},$$

and $\eta \in (0, 1/2)$ is the unique solution to

$$H(\eta) \;=\; \frac{1}{T+1},$$

with $H(\cdot)$ denoting the binary entropy function.

Corollary 33 (VC of symmetric differences) *For every class $\mathcal{F} \subseteq 2^{\mathcal{X}}$,*

$$\mathrm{VC}(\mathcal{F} \triangle \mathcal{F}) \;\leq\; 20\,\mathrm{VC}(\mathcal{F}).$$

Proof Identify sets with $\{\pm 1\}$ indicators and let $g(a, b) = a \oplus b$ (XOR), so $T = 2$. Apply Theorem 32 with $\mathcal{B} = \mathcal{F}$ and $\mathcal{G} = \{g\}$ (so $\mathrm{VC}(\mathcal{G}) = 0$) to get

$$\mathrm{VC}(\mathcal{F} \triangle \mathcal{F}) \;\leq\; 2c_2\,\mathrm{VC}(\mathcal{F}), \qquad c_2 = \frac{1}{2\eta},$$

where $\eta \in (0, 1/2)$ satisfies $H(\eta) = 1/3$. Since H is strictly increasing on $(0, 1/2)$ and $H(0.05) \approx 0.286 < 1/3$, we have $\eta > 0.05$, hence $2c_2 = 1/\eta < 20$. Thus, for concreteness, we fix the explicit bound $\mathrm{VC}(\mathcal{F} \triangle \mathcal{F}) \leq 20\,\mathrm{VC}(\mathcal{F})$. ∎

Lemma 34 (Linear VC dimension of symmetric differences) *Let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ be a set system on the product space $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ with $\mathrm{LVC}(\mathcal{F}) = g < \infty$. Then for $\mathcal{H} := \mathcal{F} \triangle \mathcal{F}$,*

$$\mathrm{LVC}(\mathcal{H}) \;\leq\; 20g.$$

Proof Fix $j \in [d]$ and $x_{-j} \in \prod_{i \neq j} \mathcal{W}_i$. Then $\mathrm{VC}(\mathcal{F}|_{x_{-j}}) \leq g$ by definition of $\mathrm{LVC}(\mathcal{F})$. Moreover, for $H = F \triangle F'$ we have $H_{x_{-j}} = F_{x_{-j}} \triangle F'_{x_{-j}}$, hence

$$\mathcal{H}|_{x_{-j}} \subseteq \mathcal{F}|_{x_{-j}} \triangle \mathcal{F}|_{x_{-j}}.$$

Since $\mathcal{H}|_{x_{-j}} \subseteq \mathcal{F}|_{x_{-j}} \triangle \mathcal{F}|_{x_{-j}}$, monotonicity of VC dimension and Corollary 33 give

$$\mathrm{VC}(\mathcal{H}|_{x_{-j}}) \leq 20\,\mathrm{VC}(\mathcal{F}|_{x_{-j}}) \leq 20g.$$

Taking the supremum over j and x_{-j} yields $\mathrm{LVC}(\mathcal{H}) \leq 20g$. ∎

Appendix E. Sauer–Shelah–Perles on Grids via Linear VC

For convenience, we use the shorthand

$$\binom{n}{\leq g} := \sum_{j=0}^{g} \binom{n}{j}, \qquad \binom{n}{\leq \infty} := 2^n.$$

Lemma 35 (Grid SSP bound from Linear VC) *Let $\mathcal{F} \subseteq 2^{\mathcal{W}_1 \times \cdots \times \mathcal{W}_d}$ and $N = A_1 \times \cdots \times A_d \subseteq \mathcal{X}$ be a finite grid with side lengths $n_i := |A_i|$. Fix an index $i \in [d]$.*

If $\mathrm{LVC}(\mathcal{F}) \leq g < \infty$, then

$$\left|\{F \cap N : F \in \mathcal{F}\}\right| \;\leq\; \left(\binom{n_i}{\leq g}\right)^{\prod_{j \neq i} n_j}. \tag{7}$$

In particular, letting $n := \max_{k \in [d]} n_k$ and choosing i such that $n_i = n$,

$$\left|\{F \cap N : F \in \mathcal{F}\}\right| \;\leq\; \left(\binom{n}{\leq g}\right)^{|N|/n}. \tag{8}$$

Proof Fix $i \in [d]$. For each choice of fixed coordinates $\mathbf{a}_{-i} \in \prod_{j \neq i} A_j$, define the axis–parallel line in direction i

$$L(\mathbf{a}_{-i}) \;:=\; \{(x_1, \ldots, x_d) \in N :\; x_j = (\mathbf{a}_{-i})_j \text{ for all } j \neq i,\;\; x_i \in A_i\}.$$

The family $\{L(\mathbf{a}_{-i}) : \mathbf{a}_{-i} \in \prod_{j \neq i} A_j\}$ partitions N into exactly $\prod_{j \neq i} n_j$ disjoint lines, each of size n_i.

Since $\mathrm{LVC}(\mathcal{F}) \leq g$, for every $\mathbf{a}_{-i}$ the restriction $\mathcal{F}|_{L(\mathbf{a}_{-i})}$ has VC dimension at most g. Therefore, by the classical Sauer–Shelah–Perles lemma on a ground set of size n_i,

$$\left|\{F \cap L(\mathbf{a}_{-i}) : F \in \mathcal{F}\}\right| \;\leq\; \sum_{j=0}^{g} \binom{n_i}{j} \;=\; \binom{n_i}{\leq g}.$$

For each $\mathbf{a}_{-i} \in \prod_{j \neq i} A_j$, the restriction of a set $F \in \mathcal{F}$ to the line $L(\mathbf{a}_{-i})$ is one of at most $\binom{n_i}{\leq g}$ possible traces. Since the lines $L(\mathbf{a}_{-i})$ are pairwise disjoint and together partition N, the collection of line–wise traces $\{F \cap L(\mathbf{a}_{-i})\}_{\mathbf{a}_{-i}}$ uniquely determines the global trace $F \cap N$. It follows that the total number of distinct traces on N is at most the product of the numbers of possible traces on each line, namely

$$\left|\{F \cap N : F \in \mathcal{F}\}\right| \;\leq \prod_{\mathbf{a}_{-i} \in \prod_{j \neq i} A_j} \left|\{F \cap L(\mathbf{a}_{-i}) : F \in \mathcal{F}\}\right| \;\leq\; \left(\binom{n_i}{\leq g}\right)^{\prod_{j \neq i} n_j},$$

which proves (7).

To obtain (8), let $n := \max_{k \in [d]} n_k$ and choose $i \in [d]$ such that $n_i = n$. Then $\prod_{j \neq i} n_j = |N|/n$, and substituting this into (7) yields

$$\left|\{F \cap N : F \in \mathcal{F}\}\right| \;\leq\; \left(\binom{n}{\leq g}\right)^{|N|/n},$$

as claimed. ∎

Corollary 36 (Grid Sauer–Shelah–Perles bound (rate form)) *With the notation of Lemma 35, let* $n := \max_i |A_i|$. *If* $1 \le g \le n$, *then*

$$\log_2 \big|\{F \cap N : F \in \mathcal{F}\}\big| \;\le\; \frac{|N|}{n}\, g\, \log_2\!\Big(\tfrac{en}{g}\Big) \;\le\; g\, n^{d-1} \log_2\!\Big(\tfrac{en}{g}\Big) \;=\; O\big(g\, n^{d-1} \log(n/g)\big). \quad (9)$$

Equivalently,

$$\big|\{F \cap N : F \in \mathcal{F}\}\big| \;\le\; 2^{O\left(g\, n^{d-1} \log(n/g)\right)}.$$

Proof Choose i with $n_i = n$. By (7) and Lemma 37,

$$\big|\{F \cap N : F \in \mathcal{F}\}\big| \;\le\; \left(\binom{n}{\le g}\right)^{|N|/n} \;\le\; \left(\tfrac{en}{g}\right)^{g|N|/n}.$$

Taking $\log_2$ yields the first inequality in (9). Since $|N|/n = \prod_{j \ne i} n_j \le n^{d-1}$, the second inequality follows. ∎

Lemma 37 *For all integers* $n \ge 1$ *and* $1 \le g \le n$,

$$\binom{n}{\le g} \;=\; \sum_{j=0}^{g} \binom{n}{j} \;\le\; \left(\tfrac{en}{g}\right)^{g}.$$

Moreover, if $g \ge n$ *then* $\binom{n}{\le g} = 2^n$.

Proof Fix integers $n \ge 1$ and $1 \le g \le n$. For every $0 \le j \le g$,

$$\binom{n}{j} = \frac{n(n-1)\cdots(n-j+1)}{j!} \;\le\; \frac{n^j}{j!},$$

and therefore

$$\sum_{j=0}^{g} \binom{n}{j} \;\le\; \sum_{j=0}^{g} \frac{n^j}{j!}.$$

Rewrite each term as

$$\frac{n^j}{j!} = \left(\tfrac{n}{g}\right)^j \cdot \frac{g^j}{j!},$$

so

$$\sum_{j=0}^{g} \frac{n^j}{j!} = \sum_{j=0}^{g} \left(\tfrac{n}{g}\right)^j \frac{g^j}{j!}.$$

Since $1 \le g \le n$, we have $\frac{n}{g} \ge 1$, hence $\left(\frac{n}{g}\right)^j \le \left(\frac{n}{g}\right)^g$ for all $0 \le j \le g$. Thus,

$$\sum_{j=0}^{g} \left(\tfrac{n}{g}\right)^j \frac{g^j}{j!} \;\le\; \left(\tfrac{n}{g}\right)^g \sum_{j=0}^{g} \frac{g^j}{j!} \;\le\; \left(\tfrac{n}{g}\right)^g \sum_{j=0}^{\infty} \frac{g^j}{j!} = \left(\tfrac{n}{g}\right)^g e^g = \left(\tfrac{en}{g}\right)^g.$$

Finally, if $g \ge n$, then $\sum_{j=0}^{g} \binom{n}{j} = \sum_{j=0}^{n} \binom{n}{j} = 2^n$. ∎

E.1. A matching lower bound for the grid SSP lemma

We show that the upper bound in Lemma 10 is essentially tight, up to constant factors, for every fixed dimension d and linear VC dimension g. In particular, the dependence on $n^{d-1}\log n$ in the exponent is unavoidable.

Throughout this subsection, let $N = [n]^d$. Recall that a $(d-1)$*–dimensional permutation* is a subset $F \subseteq [n]^d$ that contains exactly one point on every axis–parallel line. Equivalently, for each choice of a coordinate $i \in [d]$ and every fixing of the remaining $d-1$ coordinates, F contains exactly one point on the corresponding line. Let $\mathcal{F}$ denote the family of all $(d-1)$–dimensional permutations of $[n]^d$.

By a result of Keevash (2018) (Theorem 1.8), the cardinality of $\mathcal{F}$ satisfies

$$|\mathcal{F}| = \left(\tfrac{n}{e^{d-1}} + o(n)\right)^{n^{d-1}}. \tag{10}$$

Lemma 38 (A lower bound for the grid SSP rate) *Fix $d \geq 2$. Let $N = [n]^d$ and let $\mathcal{F}$ denote the family of all $(d-1)$–dimensional permutations of $[n]^d$, i.e. subsets $F \subseteq [n]^d$ that intersect every axis–parallel line in exactly one point. Fix an integer $g \geq 1$ and define*

$$\mathcal{G} \;:=\; \Big\{ \bigcup_{i=1}^{r} F_i \;:\; 0 \leq r \leq g,\;\; F_1, \ldots, F_r \in \mathcal{F} \Big\}.$$

Then $\mathrm{LVC}(\mathcal{G}) = g$. Moreover,

$$\log_2 |\mathcal{G}| \;\geq\; g\, n^{d-1} \Big(\log_2 \tfrac{n}{g\, e^{d-1}} + o(1) \Big). \tag{11}$$

In particular, for fixed d and any $g = o(n)$,

$$\log_2 |\mathcal{G}| \;=\; \Omega\big(g\, n^{d-1} \log(n/g)\big).$$

Proof

Proof of $\mathrm{LVC}(\mathcal{G}) = g$. Every $F \in \mathcal{F}$ intersects each axis–parallel line in exactly one point. Hence any union of $r \leq g$ members of $\mathcal{F}$ intersects every such line in at most $r \leq g$ points, so $\mathrm{LVC}(\mathcal{G}) \leq g$.

For the reverse inequality, fix any axis–parallel line $L \subseteq [n]^d$ and choose g distinct points $x^{(1)}, \ldots, x^{(g)} \in L$. We claim that $\mathcal{G}|_L$ shatters $\{x^{(1)}, \ldots, x^{(g)}\}$. Indeed, for each $j \in [g]$ choose a set $F^{(j)} \in \mathcal{F}$ whose unique point on L is $x^{(j)}$ (this is possible since $\mathcal{F}$ is invariant under independent permutations of the coordinates, hence contains a permutation through any prescribed point on L). Then for any subset $S \subseteq \{x^{(1)}, \ldots, x^{(g)}\}$, letting $U_S := \bigcup_{x^{(j)} \in S} F^{(j)}$ gives $U_S \in \mathcal{G}$ and $U_S \cap L = S$ (each $F^{(j)}$ contributes exactly one point on L). Thus $\mathrm{VC}(\mathcal{G}|_L) \geq g$, and taking the supremum over L yields $\mathrm{LVC}(\mathcal{G}) \geq g$.

A counting lower bound on $|\mathcal{G}|$. There are $|\mathcal{F}|^g$ ordered g–tuples $(F_1, \ldots, F_g) \in \mathcal{F}^g$, each producing a union $U = \bigcup_{i=1}^{g} F_i \in \mathcal{G}$. To lower bound $|\mathcal{G}|$, we upper bound how many g–tuples can generate the same union U.

Fix $U \in \mathcal{G}$. Partition $[n]^d$ into the n^{d-1} axis–parallel lines in (say) direction 1. Since U is a union of g members of $\mathcal{F}$, and each member of $\mathcal{F}$ contributes exactly one point to each such line,

the set U contains at most g points on every direction 1 line. Therefore, the number of possibilities for a set $F \in \mathcal{F}$ with $F \subseteq U$ is at most $g^{n^{d-1}}$ (on each of the n^{d-1} lines we have at most g choices for the unique point of F). Consequently, U can be represented as a union of an ordered g–tuple from $\mathcal{F}^g$ in at most $(g^{n^{d-1}})^g = g^{gn^{d-1}}$ ways. It follows that

$$|\mathcal{G}| \;\geq\; \frac{|\mathcal{F}|^g}{g^{gn^{d-1}}}. \tag{12}$$

Substituting Keevash and taking logs. Taking $\log_2$ in (12) gives

$$\log_2 |\mathcal{G}| \;\geq\; g \log_2 |\mathcal{F}| \;-\; gn^{d-1} \log_2 g.$$

Using (10),

$$\log_2 |\mathcal{F}| = n^{d-1} \log_2\Big(\tfrac{n}{e^{d-1}} + o(n)\Big) = n^{d-1}\Big(\log_2 n - (d-1)\log_2 e + o(1)\Big).$$

Substituting this and regrouping terms yields

$$\log_2 |\mathcal{G}| \;\geq\; gn^{d-1}\Big(\log_2 n - \log_2 g - (d-1)\log_2 e + o(1)\Big) = gn^{d-1}\Big(\log_2 \tfrac{n}{ge^{d-1}} + o(1)\Big),$$

which is (11). The final $\Omega(gn^{d-1}\log(n/g))$ statement follows whenever $g = o(n)$. ∎

Appendix F. Examples

F.1. Mixtures of product distributions

Proposition 5. *Let P be a probability distribution on $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ that can be written as a mixture of at most k product distributions,*

$$P = \sum_{t=1}^{k} \lambda_t \left(\mu_1^{(t)} \otimes \cdots \otimes \mu_d^{(t)}\right), \qquad \lambda_t \geq 0,\ \sum_{t=1}^{k} \lambda_t = 1.$$

Then P is uniformly box–continuous with modulus

$$\beta(\alpha) \;=\; \frac{\alpha^d}{(k-1+\alpha)^{\,d-1}}.$$

Proof Fix a measurable set $E \subseteq \mathcal{X}$ and define

$$\phi_t(E) \;:=\; (\mu_1^{(t)} \otimes \cdots \otimes \mu_d^{(t)})(E) \in [0,1], \qquad P(E) = \sum_{t=1}^{k} \lambda_t \phi_t(E).$$

Let $P_i = \sum_{t=1}^{k} \lambda_t \mu_i^{(t)}$ denote the i-th marginal of P, and let $P_\square := P_1 \otimes \cdots \otimes P_d$ be the product of marginals. Expanding the product of marginals yields

$$P_\square(E) = \sum_{t_1,\ldots,t_d=1}^{k} \Big(\prod_{j=1}^{d} \lambda_{t_j}\Big)(\mu_1^{(t_1)} \otimes \cdots \otimes \mu_d^{(t_d)})(E).$$

All terms are nonnegative, so keeping only the diagonal terms $t_1 = \cdots = t_d = t$ gives

$$P_{\square}(E) \;\geq\; \sum_{t=1}^{k} \lambda_t^d \, \phi_t(E). \tag{13}$$

Applying Hölder's inequality with conjugate exponents d and $d/(d-1)$ yields

$$P(E) = \sum_{t=1}^{k} \lambda_t \phi_t(E) \;\leq\; \Big(\sum_{t=1}^{k} \lambda_t^d \phi_t(E)\Big)^{1/d} \Big(\sum_{t=1}^{k} \phi_t(E)\Big)^{(d-1)/d}.$$

Rearranging gives

$$\sum_{t=1}^{k} \lambda_t^d \phi_t(E) \;\geq\; \frac{P(E)^d}{\big(\sum_{t=1}^{k} \phi_t(E)\big)^{d-1}}. \tag{14}$$

Since $\sum_t \lambda_t = 1$ and $0 \leq \phi_t(E) \leq 1$,

$$\sum_{t=1}^{k} \phi_t(E) = \sum_{t=1}^{k} \lambda_t \phi_t(E) + \sum_{t=1}^{k} (1-\lambda_t)\phi_t(E) \;\leq\; P(E) + (k-1). \tag{15}$$

Combining (13), (14), and (15) we obtain

$$P_{\square}(E) \;\geq\; \frac{P(E)^d}{\big(k - 1 + P(E)\big)^{d-1}}.$$

Since the function $x \mapsto x^d/(k-1+x)^{d-1}$ is increasing on $(0,\infty)$, replacing $P(E)$ by α completes the proof. ∎

Lemma 39 (Lower bound for mixtures) *Fix integers $d \geq 2$ and $k \geq 2$. For every $\alpha \in (0,1]$ there exist a finite product space $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$, a distribution P on $\mathcal{X}$ that is a mixture of k product distributions, and a measurable set $E \subseteq \mathcal{X}$ such that*

$$P(E) = \alpha \qquad \textit{and} \qquad P_{\square}(E) = \frac{\alpha^d}{(k-1)^{d-1}}.$$

Proof Let $\mathcal{W}_1 = \cdots = \mathcal{W}_d = [k]$ and $\mathcal{X} = [k]^d$. For each $t \in [k]$, let $P^{(t)}$ be the product distribution supported on the single point $(t, \ldots, t)$.

Define mixture weights by

$$\lambda_t := \frac{\alpha}{k-1} \quad \text{for } t = 1, \ldots, k-1, \qquad \lambda_k := 1 - \alpha,$$

and set $P := \sum_{t=1}^{k} \lambda_t P^{(t)}$. Let

$$E := \{(t, \ldots, t) : \; t = 1, \ldots, k-1\}.$$

Then, clearly $P(E) = \alpha$.

Each marginal P_i equals the distribution on $[k]$ given by $P_i(t) = \lambda_t$, and hence

$$P_{\square}(E) = \sum_{t=1}^{k-1} \prod_{i=1}^{d} P_i(t) = \sum_{t=1}^{k-1} \lambda_t^d = (k-1)\left(\frac{\alpha}{k-1}\right)^d = \frac{\alpha^d}{(k-1)^{d-1}}.$$

∎

F.2. Bounded total correlation

Definition 40 (Total correlation) *Let $\mathcal{X} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_d$ be a product measurable space, and let $P \in \Delta(\mathcal{X})$ be a probability distribution. For each $i \in [d]$, let P_i denote the marginal distribution of P on $\mathcal{W}_i$, and define the* product of marginals *by*

$$P_{\square} \;:=\; P_1 \otimes \cdots \otimes P_d.$$

The total correlation *(also known as* multi–information*) of P is defined as*

$$\mathrm{TC}(P) \;:=\; \mathrm{KL}(P \,\|\, P_{\square})\,,$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. For $d = 2$, the total correlation coincides with the mutual information.

Proposition 6. *Fix $C \geq 0$ and let*

$$\mathcal{P}_C \;:=\; \{\, P \in \Delta(\mathcal{X}) :\; \mathrm{TC}(P) \leq C \,\}.$$

Then the family $\mathcal{P}_C$ is uniformly box–continuous with modulus

$$\beta(\alpha) \;=\; \exp\!\left(\frac{-H(\alpha) - C}{\alpha}\right),$$

where $H(\alpha) := -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ is the binary entropy function. In particular, for every $P \in \mathcal{P}_C$, every measurable set $E \subseteq \mathcal{X}$, and every $\alpha \in (0, 1]$,

$$P(E) \geq \alpha \quad\Longrightarrow\quad P_{\square}(E) \;\geq\; \beta(\alpha).$$

Proof Fix $P \in \mathcal{P}_C$, a measurable set $E \subseteq \mathcal{X}$, and $\alpha \in (0, 1]$ such that $P(E) \geq \alpha$. Write

$$a := P(E), \qquad b := P_{\square}(E).$$

Define the measurable map $T : \mathcal{X} \to \{0, 1\}$ by $T(x) = \mathbf{1}_E(x)$. By the data–processing inequality for KL divergence,

$$\mathrm{TC}(P) = \mathrm{KL}(P \,\|\, P_{\square}) \;\geq\; \mathrm{KL}(P \circ T^{-1} \,\|\, P_{\square} \circ T^{-1}).$$

The pushforward measures are Bernoulli distributions,

$$P \circ T^{-1} = \mathrm{Bern}(a), \qquad P_{\square} \circ T^{-1} = \mathrm{Bern}(b),$$

and hence

$$\mathrm{TC}(P) \;\geq\; d(a\|b),$$

where

$$d(a\|b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$$

is the binary KL divergence.

Since $\mathrm{TC}(P) \leq C$, we have $d(a\|b) \leq C$. Expanding,

$$d(a\|b) = -H(a) - a \log b - (1-a)\log(1-b),$$

where $H(a) = -a\log a - (1-a)\log(1-a)$. Using $\log(1-b) \le 0$, we obtain

$$d(a\|b) \;\ge\; -H(a) - a\log b.$$

Combining with $d(a\|b) \le C$ yields

$$\log b \;\ge\; -\frac{H(a)+C}{a}, \qquad \text{hence} \qquad b \;\ge\; \exp\!\left(-\frac{H(a)+C}{a}\right).$$

Define $\phi(x) := \frac{H(x)+C}{x}$ for $x \in (0,1]$. A direct computation using $H'(x) = \log\!\left(\frac{1-x}{x}\right)$ shows that

$$\phi'(x) = \frac{\log(1-x) - C}{x^2} \;\le\; 0,$$

so ϕ is nonincreasing. Since $a \ge \alpha$, it follows that

$$\exp\!\left(-\frac{H(a)+C}{a}\right) \;\ge\; \exp\!\left(-\frac{H(\alpha)+C}{\alpha}\right) = \beta(\alpha).$$

Therefore $P_{\square}(E) \ge \beta(\alpha)$, as claimed. ■

F.3. Proofs for the permutation-matrix example

We collect here the proofs of the claims stated in Example 4. Throughout, let $n \ge 4$, let $\mathcal{X} = [n] \times [n]$, and let $\mathcal{F} = \{F_\pi : \pi \in S_n\}$ denote the class of permutation graphs, where

$$F_\pi := \{(i, \pi(i)) : \; i \in [n]\}.$$

Linear VC dimension. We verify that $\mathrm{LVC}(\mathcal{F}) = 1$. Fix an axis–parallel line L.

If L is of the form $\{i\} \times [n]$, then for every permutation $\pi \in S_n$,

$$F_\pi \cap L = \{(i, \pi(i))\},$$

which consists of exactly one point. Hence the induced one–dimensional class $\mathcal{F}|_L$ contains exactly one point from L for each π and therefore has VC dimension equal to 1.

The same argument applies to lines of the form $[n] \times \{j\}$. Taking the supremum over all axis–parallel lines yields

$$\mathrm{LVC}(\mathcal{F}) = 1.$$

Lemma 41 (Permutation graphs: empirical vs. product estimators) *Let $\mathcal{X} = [n] \times [n]$ and $\mathcal{F} = \{F_\pi : \pi \in S_n\}$, and let $P := \mathrm{Unif}([n]) \otimes \mathrm{Unif}([n])$. Let $S \sim P^m$, let $\widehat{P}_{\mathrm{emp}}$ denote the empirical distribution on $\mathcal{X}$, and let $\tilde{P}_{\prod}$ be the empirical product-of-marginals estimator. Then:*

1. *If $m \le \frac{1}{2}\sqrt{n}$, then*

$$\Pr\Big[\sup_{F \in \mathcal{F}} \big|\widehat{P}_{\mathrm{emp}}(F) - P(F)\big| \ge \tfrac{3}{4}\Big] \;\ge\; \tfrac{3}{4}.$$

2. *There exists a universal constant $C > 0$ such that for every $\varepsilon, \delta \in (0,1)$, if*

$$m \;\geq\; C\,\frac{1}{\varepsilon^2}\Big(1 + \log\tfrac{1}{\delta}\Big),$$

then

$$\Pr\Big[\sup_{F\in\mathcal{F}} \big|\tilde{P}_{\prod}(F) - P(F)\big| \leq \varepsilon\Big] \;\geq\; 1-\delta.$$

Proof Let $S = ((I_1, J_1), \ldots, (I_m, J_m)) \sim P^m$ and denote by $\widehat{P}_{\mathrm{emp}}$ the empirical distribution on $\mathcal{X}$.

Define the event

$$\mathcal{E} := \Big\{ I_1, \ldots, I_m \text{ are all distinct and } J_1, \ldots, J_m \text{ are all distinct} \Big\}.$$

By a standard birthday bound and a union bound,

$$\Pr(\mathcal{E}^c) \leq 2\binom{m}{2}\frac{1}{n} \leq \frac{m^2}{n}.$$

Hence, if $m \leq \frac{1}{2}\sqrt{n}$, then $\Pr(\mathcal{E}) \geq \frac{3}{4}$.

On the event $\mathcal{E}$, the sample points form a partial matching in the complete bipartite graph $[n] \times [n]$. By Hall's theorem, such a partial matching can always be extended to a perfect matching. Equivalently, there exists a permutation $\pi^* \in S_n$ such that $(I_t, J_t) \in F_{\pi^*}$ for all $t = 1, \ldots, m$. Consequently,

$$\widehat{P}_{\mathrm{emp}}(F_{\pi^*}) = 1.$$

On the other hand, for every $\pi \in S_n$,

$$P(F_\pi) = \sum_{i=1}^{n} P\big((i, \pi(i))\big) = \frac{1}{n}.$$

Therefore, on the event $\mathcal{E}$,

$$\sup_{F\in\mathcal{F}} \big|\widehat{P}_{\mathrm{emp}}(F) - P(F)\big| \geq 1 - \frac{1}{n} \geq \frac{3}{4},$$

where we used $n \geq 4$. This proves the first claim.

Let $\tilde{P}_1$ and $\tilde{P}_2$ denote the empirical marginals of the sample on $[n]$, and define $\tilde{P}_{\prod} := \tilde{P}_1 \otimes \tilde{P}_2$. Since P is a product distribution and $\mathrm{LVC}(\mathcal{F}) = 1$, Remark 17 yields the stated bound for $\tilde{P}_{\prod}$. ■

F.4. Infinite product spaces

This section provides the technical details underlying the discussion in Section 2.2.2.

Lemma 42 (Failure of uniform estimability in infinite products) *Let $\mathcal{X} := \{0,1\}^{\mathbb{N}}$ equipped with the product σ–algebra, and let $\mathcal{P}_{\mathrm{prod}}$ be the family of all product probability measures on $\mathcal{X}$. Define the class of finite–cylinder sets*

$$\mathcal{F}_\infty \;:=\; \bigcup_{d\geq 1} \Big\{ \{x \in \mathcal{X} :\; (x_1, \ldots, x_d) \in A\} \;:\; A \subseteq \{0,1\}^d \Big\}.$$

Then:

1. $\mathrm{LVC}(\mathcal{F}_\infty) = 2$.
2. *The pair* $(\mathcal{F}_\infty, \mathcal{P}_{\mathrm{prod}})$ *is not uniformly estimable. More precisely, there exists a universal constant* $c > 0$ *such that for every* $\varepsilon \in (0, 1/10)$ *and* $\delta \in (0, 1/3)$*, any estimator* $\widehat{P}$ *satisfying*

$$\Pr_{S \sim P^{\otimes n}}\Big[\sup_{F \in \mathcal{F}_\infty} |\widehat{P}(F) - P(F)| \le \varepsilon\Big] \;\ge\; 1 - \delta \qquad \textit{for all } P \in \mathcal{P}_{\mathrm{prod}}$$

must use

$$n \;\ge\; c\,\frac{d}{\varepsilon}$$

for arbitrarily large d*.*

Proof *Item 1.* Every axis–parallel line in $\mathcal{X} = \{0,1\}^{\mathbb{N}}$ consists of exactly two points, hence $\mathrm{VC}(\mathcal{G}|_L) \le 2$ for any class $\mathcal{G} \subseteq 2^{\mathcal{X}}$. Since $\mathcal{F}_\infty$ contains the coordinate cylinder $\{x : x_j = 1\}$ for every j, it shatters every such line. Therefore $\mathrm{LVC}(\mathcal{F}_\infty) = 2$.

Item 2. Fix $\varepsilon \in (0, 1/10)$ and $\delta \in (0, 1/3)$, and suppose for contradiction that there exist an estimator $\widehat{P}$ and a finite n such that for every $P \in \mathcal{P}_{\mathrm{prod}}$,

$$\Pr_{S \sim P^{\otimes n}}\Big[\sup_{F \in \mathcal{F}_\infty} |\widehat{P}(F) - P(F)| \le \varepsilon\Big] \;\ge\; 1 - \delta.$$

Let $d \ge 1$ be arbitrary. For $\theta \in \{\pm 1\}^d$, define the product measure

$$\overline{P}_\theta \;:=\; P_\theta \otimes \bigotimes_{i>d} Ber(1/2),$$

where P_θ is the d–dimensional product distribution from Proposition 49. Then $\overline{P}_\theta \in \mathcal{P}_{\mathrm{prod}}$.

Consider the subclass of $\mathcal{F}_\infty$ consisting of cylinder sets depending only on the first d coordinates: for each $A \subseteq \{0,1\}^d$, define

$$F_A := \{x \in \mathcal{X} :\; (x_1, \ldots, x_d) \in A\}.$$

For such sets,

$$\overline{P}_\theta(F_A) = P_\theta(A).$$

Hence the assumed guarantee implies that, for every θ, with probability at least $1 - \delta$,

$$\sup_{A \subseteq \{0,1\}^d} \big|\widehat{P}(F_A) - P_\theta(A)\big| \;\le\; \varepsilon.$$

This yields an estimator achieving uniform ε–accuracy over $\{0,1\}^d$ under the family $\{P_\theta\}$ with n samples. By Proposition 49, this requires

$$n \;\ge\; c\,\frac{d}{\varepsilon}.$$

Since d is arbitrary, no finite sample size can satisfy the assumed uniform guarantee. This contradiction completes the proof. ∎

F.4.1. TRIVIAL UNIFORM ESTIMATION FOR SOME VC CLASSES

We exhibit a class of events with infinite VC dimension for which *uniform estimation is trivial* under product measures. While infinite VC dimension can coexist with uniform estimability even in finite settings, the triviality of estimation (a single sample suffices with zero error) is a phenomenon specific to infinite product spaces.

Let

$$\mathcal{X} := \{0,1\}^{\mathbb{N}}$$

be the infinite Boolean cube equipped with its canonical product σ–algebra. Let $\mathcal{P}_{\mathrm{prod}}$ denote the family of all product probability measures on $\mathcal{X}$,

$$\mathcal{P}_{\mathrm{prod}} = \left\{ P = \bigotimes_{i=1}^{\infty} P_i \ : \ P_i \text{ is a probability measure on } \{0,1\} \right\}.$$

For $x \in \mathcal{X}$, write x_i for its i-th coordinate.

For $\alpha \in [0,1]$, define the tail event

$$E_\alpha := \left\{ x \in \mathcal{X} : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \alpha \right\},$$

where the limit is understood whenever it exists. Each E_α is measurable and invariant under modifications of finitely many coordinates, and hence is a tail event in the sense of Kolmogorov's 0-1 law.

We now define a class of sets built from these tail events. Let

$$\mathcal{F} := \Big\{ F_A := \bigcup_{\alpha\in A} E_\alpha \ : \ A \subseteq [0,1] \cap \mathbb{Q} \text{ finite} \Big\}.$$

Note that $\mathcal{F}$ is countable.

Lemma 43 *The class $\mathcal{F}$ has infinite VC dimension.*

Proof Fix $n \in \mathbb{N}$ and choose distinct rationals $\alpha_1, \dots, \alpha_n \in (0,1)$. For each $i \in [n]$, select an arbitrary point $x^{(i)} \in E_{\alpha_i}$, which is possible since each E_{α_i} is nonempty.

Let $X := \{x^{(1)}, \dots, x^{(n)}\} \subseteq \mathcal{X}$. For every subset $S \subseteq [n]$, define $A_S := \{\alpha_i : i \in S\}$ and consider the set $F_{A_S} \in \mathcal{F}$.

Since the events $\{E_{\alpha_i}\}_{i=1}^n$ are pairwise disjoint, we have

$$x^{(i)} \in F_{A_S} \quad \Longleftrightarrow \quad \alpha_i \in A_S \quad \Longleftrightarrow \quad i \in S.$$

Thus, for every labeling of the points in X, there exists a set $F_{A_S} \in \mathcal{F}$ that realizes it. Hence X is shattered by $\mathcal{F}$. Since n was arbitrary, $\mathrm{VC}(\mathcal{F}) = \infty$. ∎

Lemma 44 *For every $P \in \mathcal{P}_{\mathrm{prod}}$ and every $\alpha \in [0,1]$,*

$$P(E_\alpha) \in \{0,1\}.$$

Proof Recall that a measurable set $A \subseteq \mathcal{X} = \{0,1\}^{\mathbb{N}}$ is called a *tail event* if its membership is invariant under modifications of finitely many coordinates, i.e., if for every $x, y \in \mathcal{X}$,

$$x_i = y_i \text{ for all but finitely many } i \quad \Longrightarrow \quad \mathbf{1}_A(x) = \mathbf{1}_A(y).$$

Fix $\alpha \in [0,1]$. The event

$$E_\alpha = \left\{ x \in \mathcal{X} : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \alpha \right\}$$

depends only on the asymptotic frequency of the coordinates. Changing finitely many coordinates of x does not affect this limit (when it exists), and therefore E_α is a tail event.

Since $P \in \mathcal{P}_{\mathrm{prod}}$ is a product probability measure on $\mathcal{X}$, Kolmogorov's 0–1 law applies. By this law, every tail event under a product measure has probability either 0 or 1 (see, e.g., Shiryaev (1996) or Durrett (2019)). Hence $P(E_\alpha) \in \{0,1\}$, as claimed. ∎

Proposition 45 *For every $P \in \mathcal{P}_{\mathrm{prod}}$ and every $F \in \mathcal{F}$, we have*

$$P(F) \in \{0,1\}.$$

Moreover, uniform estimation over $\mathcal{F}$ under $\mathcal{P}_{\mathrm{prod}}$ is trivial: with probability one over a single draw $X \sim P$,

$$\sup_{F\in\mathcal{F}} \left| \widehat{P}_1(F) - P(F) \right| = 0,$$

where $\widehat{P}_1(F) := \mathbf{1}\{X \in F\}$.

Proof Let $F = F_A$ for some finite $A \subseteq [0,1] \cap \mathbb{Q}$. Since the events $\{E_\alpha\}_{\alpha\in A}$ are disjoint,

$$P(F) = \sum_{\alpha\in A} P(E_\alpha).$$

By Lemma 44, each summand belongs to $\{0,1\}$, and hence so does $P(F)$.

Now fix $X \sim P$. Almost surely, X belongs to at most one of the sets E_α. Consequently, for every $F \in \mathcal{F}$, the value $P(F)$ is determined exactly by the membership of X in F. Thus $\widehat{P}_1(F) = P(F)$ for all $F \in \mathcal{F}$ almost surely. ∎

Proposition 45 shows that, in infinite product spaces, strong independence can render uniform estimation *trivial* even for classes with infinite VC dimension. The novelty of this example lies not merely in the coexistence of infinite VC dimension and estimability, but in the fact that *exact* estimation is possible from a single sample with probability one.

This phenomenon is inherently infinite-dimensional and has no analogue in finite product spaces. In particular, it demonstrates that neither VC dimension nor linear VC dimension can characterize uniform estimability or its rates in the infinite-product regime.

F.5. Example: Necessity of d dependence for lower bound

Lemma 46 (Fano via KL from the mean distribution) *Let $M \geq 2$ and let $V \sim \mathrm{Unif}([M])$. For each $v \in [M]$ let P_v be a distribution on $\mathcal{Y}$, and draw $Y \mid (V = v) \sim P_v$. Let $\widehat{V} = f(Y)$ be any estimator and set $P_e := \Pr(\widehat{V} \neq V)$. Define the mean (mixture) distribution*

$$\bar{P} \;:=\; \frac{1}{M}\sum_{v=1}^{M} P_v.$$

Then

$$\log M - \frac{1}{M}\sum_{v=1}^{M} \mathrm{KL}(P_v \| \bar{P}) \;\leq\; h(P_e) + P_e \log(M-1), \tag{16}$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy.

In particular,

$$P_e \;\geq\; 1 - \frac{\frac{1}{M}\sum_{v=1}^{M} \mathrm{KL}(P_v \| \bar{P}) + \log 2}{\log M}. \tag{17}$$

Proof Apply the Fano inequality to $X := V$ (with $|\mathcal{X}| = M$), Y, and $\tilde{X} := \widehat{V} = f(Y)$:

$$H(V \mid Y) \;\leq\; h(P_e) + P_e \log(M-1). \tag{18}$$

We now compute $H(V \mid Y)$. Let $p(v, y)$ be the joint law of (V, Y). Since V is uniform and $Y \mid (V = v) \sim P_v$,

$$p(v,y) = \frac{1}{M} P_v(y), \qquad p_Y(y) = \sum_{u=1}^{M} p(u,y) = \bar{P}(y).$$

Hence, by Bayes' rule,

$$p(v \mid y) = \frac{p(v,y)}{p_Y(y)} = \frac{P_v(y)}{M\,\bar{P}(y)}.$$

Using the definition of conditional entropy,

$$\begin{aligned}
H(V \mid Y) &= -\sum_{v=1}^{M} \int p(v,y)\, \log p(v \mid y)\, dy \\
&= -\sum_{v=1}^{M} \int \frac{1}{M} P_v(y)\, \log\Big(\frac{P_v(y)}{M\,\bar{P}(y)}\Big)\, dy \\
&= \log M - \frac{1}{M}\sum_{v=1}^{M} \int P_v(y) \log\Big(\frac{P_v(y)}{\bar{P}(y)}\Big)\, dy \\
&= \log M - \frac{1}{M}\sum_{v=1}^{M} \mathrm{KL}(P_v \| \bar{P}).
\end{aligned}$$

Plug this identity into (18) to obtain (16). Finally, using $h(P_e) \leq \log 2$ and $\log(M-1) \leq \log M$ yields (17). ∎

Lemma 47 (KL for biased product Bernoullis) *Fix $d \in \mathbb{N}$ and $\nu \in (0, \frac{1}{4})$. For each $\theta \in \{\pm 1\}^d$ define the product distribution*

$$P_\theta \;:=\; \bigotimes_{i=1}^{d} Ber\big(\tfrac{1}{2} + \theta_i \nu\big)\,.$$

Then for any $\theta, \theta' \in \{\pm 1\}^d$,

$$\mathrm{KL}(P_\theta \| P_{\theta'}) \;=\; \Delta(\theta, \theta') \cdot D_\nu,$$

where $\Delta(\theta, \theta') := |\{i : \theta_i \neq \theta'_i\}|$ is the Hamming distance and

$$D_\nu \;:=\; \mathrm{KL}\Big(Ber\big(\tfrac{1}{2} + \nu\big) \,\Big\|\, Ber\big(\tfrac{1}{2} - \nu\big)\Big) = 2\nu \log\left(\frac{\frac{1}{2} + \nu}{\frac{1}{2} - \nu}\right) = 2\nu \log\left(\frac{1 + 2\nu}{1 - 2\nu}\right).$$

Moreover, for all $\nu \in (0, \frac{1}{4})$,

$$8\nu^2 \;\leq\; D_\nu \;\leq\; \frac{32}{3}\,\nu^2.$$

Proof By additivity of KL divergence for product measures,

$$\mathrm{KL}(P_\theta \| P_{\theta'}) = \sum_{i=1}^{d} \mathrm{KL}\Big(Ber\big(\tfrac{1}{2} + \theta_i \nu\big) \,\Big\|\, Ber\big(\tfrac{1}{2} + \theta'_i \nu\big)\Big)\,.$$

If $\theta_i = \theta'_i$, the ith summand is 0. If $\theta_i \neq \theta'_i$, then $\{\theta_i, \theta'_i\} = \{+1, -1\}$ and the ith summand equals $\mathrm{KL}(Ber(\frac{1}{2} + \nu) \| Ber(\frac{1}{2} - \nu)) =: D_\nu$. Hence the sum has exactly $\Delta(\theta, \theta')$ nonzero terms, proving

$$\mathrm{KL}(P_\theta \| P_{\theta'}) = \Delta(\theta, \theta')\, D_\nu.$$

Let $p = \frac{1}{2} + \nu$ and $q = \frac{1}{2} - \nu$. Then

$$\begin{aligned} D_\nu &= p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} = p \log \frac{p}{q} + q \log \frac{q}{p} \\ &= (p - q) \log \frac{p}{q} = 2\nu \log\left(\frac{\frac{1}{2} + \nu}{\frac{1}{2} - \nu}\right) = 2\nu \log\left(\frac{1 + 2\nu}{1 - 2\nu}\right). \end{aligned}$$

Set $x := 2\nu \in (0, \frac{1}{2})$. For $x \in (0, 1)$ one has

$$2x \;\leq\; \log\left(\frac{1 + x}{1 - x}\right) \;\leq\; \frac{2x}{1 - x^2}.$$

Applying these inequalities,

$$D_\nu = 2\nu \log\left(\frac{1 + x}{1 - x}\right) \geq 2\nu \cdot 2x = 8\nu^2,$$

and

$$D_\nu \leq 2\nu \cdot \frac{2x}{1 - x^2} = \frac{8\nu^2}{1 - 4\nu^2} \leq \frac{32}{3}\nu^2,$$

since $\nu < 1/4$ implies $1 - 4\nu^2 \geq 3/4$. ∎

Lemma 48 (Hellinger separation for biased product Bernoullis) *Let $k \geq 1$ and define product measures on $\{0,1\}^k$ by*

$$P_+ := Ber\big(\tfrac{1}{2}+\nu\big)^{\otimes k}, \qquad P_- := Ber\big(\tfrac{1}{2}-\nu\big)^{\otimes k},$$

where $0 < \nu < 1/2$. Let

$$H^2(P,Q) \;:=\; 2\big(1-\rho(P,Q)\big), \qquad \rho(P,Q) \;:=\; \sum_x \sqrt{P(x)Q(x)}$$

denote the squared Hellinger distance and the Bhattacharyya coefficient. Then for every $0 < \varepsilon \leq 1$, if

$$\nu \;\geq\; \sqrt{\frac{\varepsilon}{2k}},$$

we have

$$H^2(P_+, P_-) \;\geq\; \varepsilon.$$

Proof For a single coordinate, let

$$P_+^{(1)} = Ber\big(\tfrac{1}{2}+\nu\big), \qquad P_-^{(1)} = Ber\big(\tfrac{1}{2}-\nu\big).$$

A direct computation gives

$$\rho\Big(P_+^{(1)}, P_-^{(1)}\Big) = \sqrt{1-4\nu^2}.$$

For product measures $P = \bigotimes_{i=1}^k P_i$ and $Q = \bigotimes_{i=1}^k Q_i$ on a finite product space, one has the multiplicativity

$$\rho(P,Q) = \prod_{i=1}^k \rho(P_i, Q_i).$$

Applying this with $P_i = P_+^{(1)}$ and $Q_i = P_-^{(1)}$ yields

$$\rho(P_+, P_-) = \big(\rho(P_+^{(1)}, P_-^{(1)})\big)^k = (1-4\nu^2)^{k/2},$$

and therefore

$$H^2(P_+, P_-) = 2\Big(1-(1-4\nu^2)^{k/2}\Big). \tag{19}$$

Since $4\nu^2 \leq 1$, we may use $(1-x)^m \leq e^{-mx}$ for $x \in [0,1]$ to obtain

$$(1-4\nu^2)^{k/2} \leq e^{-2k\nu^2}.$$

Substituting into (19) gives

$$H^2(P_+, P_-) \geq 2\big(1-e^{-2k\nu^2}\big).$$

If $\nu^2 \geq \varepsilon/(2k)$, then $2k\nu^2 \geq \varepsilon$, hence

$$H^2(P_+, P_-) \geq 2\big(1-e^{-\varepsilon}\big).$$

For $\varepsilon \in (0,1]$ we have $1-e^{-\varepsilon} \geq \varepsilon/2$, and therefore

$$H^2(P_+, P_-) \geq 2\cdot\frac{\varepsilon}{2} = \varepsilon,$$

as claimed. ∎

Proposition 49 ($\Omega(d)$ samples for TV/uniform estimation over the d–cube) *Let $\mathcal{X} = \{0,1\}^d$ and let $\mathcal{F} = 2^{\mathcal{X}}$. Fix $\varepsilon \in (0, 1/10)$ and $\delta \in (0, 1/3)$. Consider the family of* product *distributions*

$$\mathcal{P} := \{P_\theta : \theta \in \{\pm 1\}^d\}, \qquad P_\theta := \bigotimes_{i=1}^{d} Ber\left(\tfrac{1}{2} + \theta_i \nu\right),$$

where

$$\nu \;:=\; 4\sqrt{\frac{\varepsilon}{d}}.$$

Assume d is large enough so that $\nu < 1/4$.

Suppose an estimator $\widehat{P}$ based on n i.i.d. samples satisfies

$$\Pr_{S \sim P^{\otimes n}}\left[\sup_{F \in \mathcal{F}} |\widehat{P}(F) - P(F)| \le \varepsilon\right] \;\ge\; 1 - \delta \qquad \text{for all } P \in \mathcal{P}.$$

Then necessarily

$$n \;\ge\; c\,\frac{d}{\varepsilon},$$

for a universal constant $c > 0$. In particular, for constant ε, one must have $n = \Omega(d)$.

Proof Step 1: a large code with Hamming distance $\ge d/4$. Let $\Theta \subseteq \{\pm 1\}^d$ be such that for all distinct $\theta, \theta' \in \Theta$,

$$\Delta(\theta, \theta') := |\{i : \theta_i \ne \theta'_i\}| \;\ge\; d/4,$$

and $|\Theta| \ge 2^{c_1 d}$ for a universal constant $c_1 > 0$. Such a set exists by the Gilbert–Varshamov bound. We restrict attention to the subfamily $\{P_\theta : \theta \in \Theta\}$.

Step 2: separation in total variation. Fix distinct $\theta, \theta' \in \Theta$ and let $k := \Delta(\theta, \theta') \ge d/4$. Let $I := \{i \in [d] : \theta_i \ne \theta'_i\}$, so $|I| = k$. Since the Bhattacharyya coefficient factors over product measures, and the marginals coincide on $[d] \setminus I$, we have

$$H^2(P_\theta, P_{\theta'}) \;=\; H^2\Big(Ber(\tfrac{1}{2} + \nu)^{\otimes k}, Ber(\tfrac{1}{2} - \nu)^{\otimes k}\Big).$$

By Lemma 48 applied with $\varepsilon' := 8\varepsilon$, it suffices that

$$\nu \;\ge\; \sqrt{\frac{\varepsilon'}{2k}} \;=\; \sqrt{\frac{8\varepsilon}{2k}} \;=\; 2\sqrt{\frac{\varepsilon}{k}} \;\le\; 4\sqrt{\frac{\varepsilon}{d}},$$

where the last inequality uses $k \ge d/4$. With our choice $\nu = 4\sqrt{\varepsilon/d}$, we obtain

$$H^2(P_\theta, P_{\theta'}) \;\ge\; 8\varepsilon.$$

Since $\mathrm{TV}(P, Q) \ge \frac{1}{2} H^2(P, Q)$, it follows that

$$\mathrm{TV}(P_\theta, P_{\theta'}) \;\ge\; 4\varepsilon \qquad \text{for all distinct } \theta, \theta' \in \Theta.$$

Step 3: uniform estimation implies decoding. For distributions on a finite domain,

$$\mathrm{TV}(P, Q) = \sup_{F \subseteq \mathcal{X}} |P(F) - Q(F)|.$$

Thus, by assumption, with probability at least $1-\delta$,

$$\mathrm{TV}(\widehat{P}, P_\theta) \;\le\; \varepsilon.$$

Define the decoder

$$\widehat{\theta} \in \arg\min_{\theta' \in \Theta} \mathrm{TV}(\widehat{P}, P_{\theta'}).$$

On the event $\mathrm{TV}(\widehat{P}, P_\theta) \le \varepsilon$, separation implies $\widehat{\theta} = \theta$: indeed, for $\theta' \neq \theta$,

$$\mathrm{TV}(\widehat{P}, P_{\theta'}) \;\ge\; \mathrm{TV}(P_\theta, P_{\theta'}) - \mathrm{TV}(\widehat{P}, P_\theta) \;\ge\; 4\varepsilon - \varepsilon \;>\; \varepsilon.$$

Hence

$$\Pr[\widehat{\theta} = \theta] \;\ge\; 1-\delta. \tag{20}$$

Step 4: Fano via KL from the mean distribution. Let V be uniform on Θ, and let $S \mid (V=\theta) \sim (P_\theta)^{\otimes n}$. Let $\widehat{V} = \widehat{\theta}(S)$. By (20),

$$P_e := \Pr[\widehat{V} \neq V] \;\le\; \delta.$$

Let

$$\bar{Q} \;:=\; \frac{1}{|\Theta|} \sum_{\theta \in \Theta} (P_\theta)^{\otimes n}$$

be the mean distribution of S. Applying Lemma 46 gives

$$\log|\Theta| - \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathrm{KL}\big((P_\theta)^{\otimes n} \big\| \bar{Q}\big) \;\le\; h(P_e) + P_e \log(|\Theta|-1). \tag{21}$$

Using $P_e \le \delta$, $h(P_e) \le \log 2$, and $\log(|\Theta|-1) \le \log|\Theta|$, we obtain

$$\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathrm{KL}\big((P_\theta)^{\otimes n} \big\| \bar{Q}\big) \;\ge\; (1-\delta)\log|\Theta| - \log 2.$$

Step 5: upper bound the KL from the mean. By convexity of KL in its second argument,

$$\mathrm{KL}\big((P_\theta)^{\otimes n} \big\| \bar{Q}\big) \le \frac{1}{|\Theta|} \sum_{\theta' \in \Theta} \mathrm{KL}\big((P_\theta)^{\otimes n} \big\| (P_{\theta'})^{\otimes n}\big)\,.$$

Averaging over θ yields

$$\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathrm{KL}\big((P_\theta)^{\otimes n} \big\| \bar{Q}\big) \le \frac{n}{|\Theta|^2} \sum_{\theta, \theta' \in \Theta} \mathrm{KL}(P_\theta \| P_{\theta'}).$$

By Lemma 47,

$$\mathrm{KL}(P_\theta \| P_{\theta'}) = \Delta(\theta, \theta')\, D_\nu \le d \cdot D_\nu.$$

Using the upper bound $D_\nu \le \frac{32}{3}\nu^2$, we get

$$\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathrm{KL}\big((P_\theta)^{\otimes n} \big\| \bar{Q}\big) \;\le\; n \cdot C\, d\, \nu^2$$

for a universal constant $C > 0$.

Step 6: conclude the sample complexity lower bound. Combining the lower bound from Step 4 with the upper bound from Step 5 and using $\log|\Theta| \geq c_1 d$, we obtain

$$n \cdot C\, d\, \nu^2 \;\geq\; (1-\delta)\log|\Theta| - \log 2 \;\geq\; (1-\delta)c_1 d - \log 2.$$

Since $\delta \leq 1/3$, the right-hand side is at least $c_2 d$ for a universal constant $c_2 > 0$ (absorbing the additive $-\log 2$ term into the constant for all sufficiently large d). Hence

$$n \cdot C\, d\, \nu^2 \;\geq\; c_2 d,$$

and therefore

$$n \;\geq\; \frac{c_2}{C\,\nu^2}.$$

Recalling that $\nu = 4\sqrt{\varepsilon/d}$ (so that $\nu^2 = 16\,\varepsilon/d$), this yields

$$n \;\geq\; \frac{c_2}{C} \cdot \frac{d}{16\,\varepsilon} \;=\; \frac{c_2}{16C}\,\frac{d}{\varepsilon}.$$

Setting $c := c_2/(16C)$ completes the proof. ■