THE LIST SPECTRUM OF PRODUCT CLASSES

Pahan Dewasurendra
Johns Hopkins University

ABSTRACT

How many predictions per example are necessary to learn several classification tasks simultaneously? For a class $\mathcal{H}$, let $K(\mathcal{H})$ be the least list size that makes realizable PAC learning possible. Hanneke, Moran, and Waknine asked for $K(\mathcal{H}_1 \otimes \mathcal{H}_2)$, leaving a gap between $(K(\mathcal{H}_1) - 1)(K(\mathcal{H}_2) - 1)$ and $K(\mathcal{H}_1)K(\mathcal{H}_2)$.

We close the gap and determine a stronger risk spectrum. For nonempty classes with finite $k_j = K(\mathcal{H}_j)$, set $M = \prod_j k_j$. If a learner may output L tuple labels, then its optimal worst-case expected realizable error converges to

$$\left(1 - \frac{L}{M}\right)_+ .$$

Consequently, $K(\bigotimes_j \mathcal{H}_j) = \prod_j K(\mathcal{H}_j)$. The result is quantitative at every sample size. Its lower bound is a strong direct-product inequality in terms of the factor learning curves.

The proof turns infinite list-DS pseudo-cubes into explicit finite Bayes experiments. On an unseen coordinate, the posterior contains k_j competing labels. A difference of adjacent Bayes list risks lower bounds the k_j-th posterior mass. Under independent products, these masses multiply, while a list of size L omits at least $M - L$ cells of the top-posterior grid. This posterior order-statistic argument also shows that optimal weak multiclass accuracies multiply. We complement the list theorem with an exact fixed-marginal tensorization under standard minimax regularity and dimension-dependent rates for uniform and agnostic product learning.

1 INTRODUCTION

Direct-sum questions ask how the resources needed for one task scale when several independent instances must be solved together. In classification, the Cartesian product $\mathcal{H}_1 \otimes \mathcal{H}_2$ predicts a pair of labels and incurs zero-one loss when either coordinate is wrong. Separate learning gives immediate upper bounds, but it need not reveal the optimal dependence on the number or structure of the tasks.

Hanneke et al. (2024b) formulated a collection of direct-sum problems for PAC learning, uniform convergence, agnostic learning, compression, and combinatorial dimensions. Their sharpest qualitative question concerns list learning. A k-list learner gets k guesses at the label. If $K(\mathcal{H})$ is the least k for which $\mathcal{H}$ is k-list PAC learnable, they proved

$$(K(\mathcal{H}_1) - 1)(K(\mathcal{H}_2) - 1) \le K(\mathcal{H}_1 \otimes \mathcal{H}_2) \le K(\mathcal{H}_1)K(\mathcal{H}_2).$$

The upper bound outputs all pairs from two marginal lists. The lower bound combines list-DS pseudo-cubes, but its gap is largest when one factor is ordinarily learnable and remains nontrivial for every fixed list size.

We prove that the Cartesian upper bound is exact. More strongly, we determine the asymptotic minimax error for every list budget. Let $\mathfrak{R}_k(n; \mathcal{H})$ be the smallest worst-case expected realizable error of a randomized k-list learner trained on n examples. For $k_j = K(\mathcal{H}_j) < \infty$ and $M = \prod_j k_j$, our main theorem gives

$$\lim_{n\to\infty} \mathfrak{R}_L\left(n; \bigotimes_j \mathcal{H}_j\right) = \left(1 - \frac{L}{M}\right)_+ . \tag{1}$$

Thus every missing slot below the threshold has an exact asymptotic price. In particular, an $(M-1)$-list learner retains error $1/M$, while an M-list learner achieves vanishing error.

The lower bound does not follow by tensorizing the previous nonlearnability statement. That route loses one label in every factor. Our proof instead extracts a posterior order statistic from the sharp boundary between learnable and nonlearnable list sizes. If $k = K(\mathcal{H}) \geq 2$, the $(k-1)$-DS dimension is infinite by the characterization of Charikar & Pabbaraju (2023). An arbitrarily high-dimensional pseudo-cube supports a finite prior for which a $(k-1)$-list has Bayes error arbitrarily close to $1/k$. Under the same prior, a k-list has Bayes error at most $\mathfrak{R}_k(n;\mathcal{H})$. The difference of these two risks is the expected k-th largest posterior label probability.

For a product experiment, the coordinate posteriors are independent. Their top k_j labels form a grid of M tuple labels. Every cell has posterior mass at least the product of the boundary order statistics. Any L-list omits at least $M - L$ cells. This gives the finite-sample inequality

$$\mathfrak{R}_L\left(n; \bigotimes_j \mathcal{H}_j\right) \geq (M-L) \prod_j \left(\frac{1}{k_j} - \mathfrak{R}_{k_j}(n;\mathcal{H}_j)\right)_+ . \tag{2}$$

Taking $n \to \infty$ proves the lower half of (1). For the upper half, form the Cartesian list and retain a uniformly random L-sublist.

The theorem has an immediate boosting interpretation. Brukhim et al. (2023) prove that the best weak PAC accuracy $\gamma(\mathcal{H})$ satisfies $K(\mathcal{H})\gamma(\mathcal{H}) = 1$. Exact list multiplicativity therefore implies exact weak-accuracy multiplicativity. Solving more tasks simultaneously multiplies the smallest nontrivial accuracy rather than adding weak edges.

We also revisit the learning-curve questions of Hanneke et al. (2024b). For a known marginal, product Bayes risks tensorize exactly. This yields the minimax identity $1 - \prod_j (1 - \mathfrak{R}_j)$ whenever the standard minimax equality with finite priors holds, including every finite game. For unrestricted agnostic learning, no formula in the one-factor curve alone is possible. The recent separation of More et al. (2026) gives two binary classes with identical one-factor rates and different product rates. We give a complementary dimension-dependent result. In the nondegenerate regime, the latest sharp multiclass theorems imply an agnostic product rate with a linear realizable term and a square-root noisy term.

Contributions.

1. We determine the complete asymptotic list-risk spectrum of a finite Cartesian product and prove the finite-sample strong direct-product inequality (2).
2. We resolve the minimal-list-size question exactly, including heterogeneous factors and infinite list numbers, and deduce multiplicativity of optimal weak PAC accuracy.
3. We prove a fixed-marginal Bayes and minimax tensorization theorem for all-coordinate zero-one loss.
4. We derive product-class rates from graph, DS, and Natarajan dimensions, while identifying the low-dimensional exceptions exposed by the latest agnostic separation.

2 Setting

Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a nonempty concept class. A randomized k-list learner maps a labeled sample $S \in (\mathcal{X} \times \mathcal{Y})^n$ to a measurable function $A(S) : \mathcal{X} \to 2^{\mathcal{Y}}$ satisfying $|A(S)(x)| \leq k$. For a marginal D on $\mathcal{X}$, a target $h \in \mathcal{H}$, and an independent $X \sim D$, its expected realizable error is

$$\mathbb{E}_{S\sim D_h^n, A, X} \mathbf{1}\{h(X) \notin A(S)(X)\}, \qquad D_h = \mathrm{Law}(X, h(X)).$$

Define the minimax list risk

$$\mathfrak{R}_k(n;\mathcal{H}) = \inf_A \sup_D \sup_{h\in\mathcal{H}} \mathbb{E}_{S\sim D_h^n, A, X} \mathbf{1}\{h(X) \notin A(S)(X)\}. \tag{3}$$

Set $\mathfrak{R}_0(n;\mathcal{H}) = 1$. Allowing lists of size at most rather than exactly k is immaterial because lists can be padded whenever needed.

Vanishing minimax expected risk is equivalent to realizable list PAC learnability. A high-probability guarantee gives vanishing expectation after clipping the loss at one. Conversely, Markov's inequality turns vanishing expectation into any fixed error and confidence guarantee. We therefore define

$$K(\mathcal{H}) = \min\{k \geq 1 : \mathfrak{R}_k(n; \mathcal{H}) \to 0\},$$

with $K(\mathcal{H}) = \infty$ if no finite list works.

The list-DS characterization gives the same threshold if learner randomization is forbidden. Randomization is used only to state the exact subcritical spectrum in its simplest form.

For classes $\mathcal{H}_j \subseteq \mathcal{Y}_j^{\mathcal{X}_j}$, their Cartesian product is

$$\bigotimes_{j=1}^{r} \mathcal{H}_j = \{(x_1, \ldots, x_r) \mapsto (h_1(x_1), \ldots, h_r(x_r)) : h_j \in \mathcal{H}_j\}.$$

The label is a tuple and the loss is one unless the entire true tuple appears in the predicted list. This is all-coordinate zero-one loss, not averaged Hamming loss.

We use the list-DS characterization. The ordinary DS dimension was introduced by Daniely & Shalev-Shwartz (2014) and shown to characterize multiclass PAC learnability by Brukhim et al. (2022). A finite set $F \subseteq \mathcal{Y}^d$ is a q-pseudo-cube if, for every $f \in F$ and coordinate $i \in [d]$, at least q other members agree with f outside i and have distinct values at i. A class q-DS shatters $x_{1:d}$ if its trace contains such an F. The q-DS dimension is the largest shattered d. The theorem of Charikar & Pabbaraju (2023) states that $\mathcal{H}$ is q-list PAC learnable if and only if its q-DS dimension is finite. Notice the indexing: a q-pseudo-cube has coordinate lines of size at least $q + 1$.

3 The Exact List Spectrum

Theorem 1 (Finite-sample direct product and exact spectrum)**.** *Let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be nonempty classes with finite $k_j = K(\mathcal{H}_j)$, and set $M = \prod_{j=1}^{r} k_j$. For every $n \geq 0$ and integer $0 \leq L < M$,*

$$\mathfrak{R}_L\left(n; \bigotimes_{j=1}^{r} \mathcal{H}_j\right) \geq (M - L) \prod_{j=1}^{r} \left(\frac{1}{k_j} - \mathfrak{R}_{k_j}(n; \mathcal{H}_j)\right)_+, \tag{4}$$

$$\mathfrak{R}_L\left(n; \bigotimes_{j=1}^{r} \mathcal{H}_j\right) \leq 1 - \frac{L}{M} + \frac{L}{M} \sum_{j=1}^{r} \mathfrak{R}_{k_j}(n; \mathcal{H}_j). \tag{5}$$

For $L \geq M$, the product risk is at most $\sum_j \mathfrak{R}_{k_j}(n; \mathcal{H}_j)$. Consequently, for every fixed integer $L \geq 0$,

$$\lim_{n \to \infty} \mathfrak{R}_L\left(n; \bigotimes_{j=1}^{r} \mathcal{H}_j\right) = \left(1 - \frac{L}{M}\right)_+. \tag{6}$$

The limit is a strong converse. Below M, more samples cannot overcome an insufficient prediction budget. The asymptotic success probability is exactly the fraction L/M of the critical Cartesian list that the learner can retain. For example, the r-th power of any class with $K(\mathcal{H}) = 2$ has critical list size 2^r, and every fixed $L < 2^r$ has limiting error $1 - L/2^r$. Classes with $K = 2$ and infinite ordinary DS dimension are constructed by Charikar & Pabbaraju (2023).

Corollary 2 (List number and weak accuracy)**.** *For nonempty classes,*

$$K\left(\bigotimes_{j=1}^{r} \mathcal{H}_j\right) = \prod_{j=1}^{r} K(\mathcal{H}_j),$$

with the convention that a product containing infinity is infinite. If all list numbers are finite and $\gamma(\mathcal{H})$ denotes the optimal weak PAC accuracy of Brukhim et al. (2023), then

$$\gamma\left(\bigotimes_{j=1}^{r} \mathcal{H}_j\right) = \prod_{j=1}^{r} \gamma(\mathcal{H}_j).$$

For two factors, Corollary 2 replaces both sides of the previous interval by $K(\mathcal{H}_1)K(\mathcal{H}_2)$. Factors with $K = 1$ are not exceptions. They leave the critical list size unchanged and contribute their ordinary learning error to the finite-sample bounds.

4 Proof of the spectrum theorem

The proof has three ingredients. First, a pseudo-cube makes the boundary list size Bayes-hard. Second, adjacent Bayes list risks expose one posterior order statistic. Third, posterior order statistics multiply across independent experiments.

4.1 Boundary posteriors from pseudo-cubes

Fix a finite prior Π on pairs (D, h) with $h \in \mathcal{H}$. After observing $S \sim D_h^n$ and an unlabeled test point $X \sim D$, let

$$p_{(1)}(S, X) \geq p_{(2)}(S, X) \geq \cdots$$

be the ordered posterior probabilities of $h(X)$, padded by zeros. The Bayes q-list risk is

$$\mathfrak{B}_q(\Pi) = 1 - \mathbb{E} \sum_{a=1}^{q} p_{(a)}(S, X).$$

Hence adjacent list risks satisfy the exact identity

$$\mathbb{E} p_{(q)}(S, X) = \mathfrak{B}_{q-1}(\Pi) - \mathfrak{B}_q(\Pi). \tag{7}$$

Lemma 3 (Boundary posterior)**.** *Let $k = K(\mathcal{H}) < \infty$. For every n and $\eta > 0$, there is a finite prior with*

$$\mathbb{E} p_{(k)}(S, X) \geq \frac{1 - \eta}{k} - \mathfrak{R}_k(n; \mathcal{H}).$$

For $k = 1$, one may replace $(1 - \eta)/k$ by one.

Proof. Assume $k \geq 2$. Minimality of k and the list-DS characterization imply that the $(k-1)$-DS dimension of $\mathcal{H}$ is infinite. Choose a $(k-1)$-pseudo-cube F on d coordinates, where d is large enough that $(1 - 1/d)^n \geq 1 - \eta$. Use the uniform prior on its distinct trace vectors, selecting one representative hypothesis per vector, and let D be uniform on the d coordinates.

Condition on the test coordinate not appearing among the n training coordinates. Reveal to the learner every target label outside the test coordinate. This additional information can only reduce Bayes risk. The compatible vertices form one coordinate line of F. The prior conditional on that line is uniform, and the pseudo-cube property gives at least k distinct labels on it. Even with the extra information, a $(k-1)$-list therefore misses with probability at least $1/k$. Thus

$$\mathfrak{B}_{k-1}(\Pi) \geq \frac{(1 - 1/d)^n}{k} \geq \frac{1 - \eta}{k}.$$

For every prior, its Bayes k-list risk is at most the worst-case minimax risk $\mathfrak{R}_k(n; \mathcal{H})$. Equation (7) proves the claim. If $k = 1$, use $\mathfrak{B}_0 = 1$, $\mathfrak{B}_1 \leq \mathfrak{R}_1(n; \mathcal{H})$, and $p_{(1)} = \mathfrak{B}_0 - \mathfrak{B}_1$. □

Repeated training coordinates cause no difficulty in Lemma 3. The hard event is precisely that the test coordinate is absent. Revealing the entire complement subsumes all repeated observations and leaves a coordinate line with the required number of distinct labels.

4.2 A top-grid inequality

Lemma 4 (Omitted posterior grid)**.** *For each $j \in [r]$, let p_j be a probability vector with ordered entries $p_{j,(1)} \geq p_{j,(2)} \geq \cdots$. Set $M = \prod_j k_j$. Every list of at most $L < M$ atoms from the product distribution $\bigotimes_j p_j$ has omitted mass at least*

$$(M - L) \prod_{j=1}^{r} p_{j,(k_j)}.$$

Proof. The Cartesian product of the top k_j coordinates contains M atoms. A list of size L omits at least $M - L$ of them. Every omitted atom in this grid has mass at least $\prod_j p_{j,(k_j)}$. Summing their masses proves the claim. □

4.3 Combining the factors

Proof of Theorem 1. For each factor, take the finite prior from Lemma 3, with a common parameter η. Draw the factor targets, training coordinates, and test coordinates independently. This defines a finite prior over realizable problems for the product class. Conditional on the full product sample and test tuple, the posterior law of the tuple label factorizes as $\bigotimes_j p_j$.

Lemma 4 lower bounds the conditional Bayes L-list error. The factor experiments are independent, so taking expectations gives

$$\mathfrak{B}_L \geq (M-L)\prod_j \mathbb{E}p_{j,(k_j)} \geq (M-L)\prod_j \left(\frac{1-\eta}{k_j} - \mathfrak{R}_{k_j}(n;\mathcal{H}_j)\right)_+ .$$

The minimax risk is at least the Bayes risk under every finite prior. Letting $\eta \downarrow 0$ proves (4).

For the upper bound, run independent near-minimax k_j-list learners on the coordinate projections of the product sample. Their Cartesian list has size $m \leq M$, and it omits the true tuple with probability at most $\sum_j \mathfrak{R}_{k_j}(n;\mathcal{H}_j)$, up to an arbitrarily small optimization slack. If $m > L$, choose a uniformly random L-subset. If $m \leq L$, keep the whole list. Conditional on the true tuple being present, it is retained with probability at least L/M. Therefore the success probability is at least

$$\frac{L}{M}\left(1 - \sum_j \mathfrak{R}_{k_j}(n;\mathcal{H}_j)\right),$$

which proves (5). For $L \geq M$, retain the full Cartesian list. Finally, $\mathfrak{R}_{k_j}(n;\mathcal{H}_j) \to 0$ for every factor, so the finite-sample bounds converge to (6). □

If one factor has infinite list number and the product were L-list learnable, fix arbitrary targets and point-mass marginals in the other factors. Project every output tuple list onto the hard coordinate. The projected list has size at most L, and its error is no larger than the tuple-list error. This would make the hard factor finitely list learnable, a contradiction. Corollary 2 now follows from Theorem 1 and the identity $K(\mathcal{H})\gamma(\mathcal{H}) = 1$ of Brukhim et al. (2023).

5 Fixed-marginal tensorization

The same posterior factorization gives an exact finite-sample statement for ordinary classification, subject only to the usual distinction between minimax and least-favorable-prior values. Fix a known marginal D and let $R(n; D, \mathcal{H})$ be the randomized minimax one-list risk, with the supremum only over $h \in \mathcal{H}$. Let

$$\underline{R}(n; D, \mathcal{H}) = \sup_{\Pi} \inf_{A} \mathbb{E}_{h\sim\Pi, S, X, A} \mathbf{1}\{A(S)(X) \neq h(X)\},$$

where the supremum is over finitely supported priors on $\mathcal{H}$.

Theorem 5 (Fixed-marginal sandwich)**.** *For r classes and marginals,*

$$1 - \prod_{j=1}^{r}(1 - \underline{R}(n; D_j, \mathcal{H}_j)) \leq R\left(n; \bigotimes_j D_j, \bigotimes_j \mathcal{H}_j\right) \leq 1 - \prod_{j=1}^{r}(1 - R(n; D_j, \mathcal{H}_j)). \quad (8)$$

Whenever finite-prior minimax equality holds in every factor, both inequalities are equalities. In particular, for every finite game,

$$R(n; D^r, \mathcal{H}^r) = 1 - (1 - R(n; D, \mathcal{H}))^r .$$

For the upper bound, run independent factor learners. Their correctness events are independent for every fixed target tuple. For the lower bound, take independent nearly least-favorable priors. Conditional posterior correctness maxima multiply, so Bayes all-coordinate success is the product of the factor Bayes successes. Appendix B gives the full argument and states sufficient regularity conditions.

In the small-risk regime, $1-(1-R)^r = rR(1+O(rR))$. Thus a fixed marginal offers no asymptotic improvement over separate learning before the all-coordinate loss saturates. This gives a sharp answer to the fixed-marginal question of Hanneke et al. (2024b) for finite games and, more generally, whenever the standard minimax identity is valid.

6 What tensorizes in agnostic and uniform learning

The list spectrum is universal because it is governed by a qualitative boundary. Agnostic excess-risk curves are not. More et al. (2026) exhibit binary classes $\mathcal{F}$ and $\mathcal{G}$ with $\mathfrak{R}^{\mathrm{agn}}(n;\mathcal{F}) \asymp \mathfrak{R}^{\mathrm{agn}}(n;\mathcal{G}) \asymp n^{-1/2}$, but $\mathcal{F}^r$ retains rate $n^{-1/2}$ while $\mathcal{G}^r$ has rate $\min\{1, \sqrt{r/n}\}$. The one-factor curve and r therefore cannot determine the product curve.

Combinatorial dimensions recover a positive law away from the exceptional one-dimensional regime. Let $d_{\mathrm{DS}}(\mathcal{H})$ and $d_{\mathrm{N}}(\mathcal{H})$ be the DS and Natarajan dimensions (Natarajan, 1989). Combining the sharp realizable theorem of Pabbaraju (2026), the two-parameter agnostic theorem of Cohen et al. (2026), and the product-dimension bounds of Hanneke et al. (2024b) yields the following consequence. A proof is included in Appendix C.

Corollary 6 (Agnostic products in the nondegenerate regime)**.** *Let $\mathcal{H}$ have a finite label space, $d_{\mathrm{DS}}(\mathcal{H}) = d \geq 2$, and $d_{\mathrm{N}}(\mathcal{H}) = q \geq 3$. Up to logarithmic factors, the agnostic PAC sample complexity of $\mathcal{H}^r$ is*

$$\widetilde{\Theta}\left(\frac{rd}{\varepsilon} + \frac{rq + \log(1/\delta)}{\varepsilon^2}\right).$$

Equivalently, its expected excess-risk scale is

$$\widetilde{\Theta}\left(\frac{rd}{n} + \sqrt{\frac{rq}{n}}\right),$$

clipped at one.

The assumptions $d \geq 2$ and $q \geq 3$ are substantive. Product dimensions satisfy $d_{\mathrm{DS}}(\mathcal{H}^r) \geq r(d-1)+1$ and $d_{\mathrm{N}}(\mathcal{H}^r) \geq r(q-2)+2$. Neither lower bound grows with r at the smallest dimension, exactly where the separation of More et al. (2026) operates.

There is a parallel distribution-free statement for empirical uniform convergence. If $g = d_{\mathrm{G}}(\mathcal{H}) \geq 2$, then

$$r(g-1) \leq d_{\mathrm{G}}(\mathcal{H}^r) \leq Crg\log(2r).$$

The lower bound uses one anchor in each graph-shattered factor. The upper bound observes that the product loss is the OR of r lifted factor losses and applies the growth-function product bound. Thus the worst-case expected uniform-deviation curve lies between constant multiples of

$$\min\left\{1, \sqrt{\frac{r(g-1)}{n}}\right\} \quad \text{and} \quad \min\left\{1, \sqrt{\frac{rg\log(2r)}{n}}\right\}.$$

This concerns the empirical process of a product concept class under arbitrary distributions. Holzman et al. (2026) study a different problem, namely arbitrary event families on product domains under distributions compatible with the product structure, and characterize when a nonempirical uniform estimator exists via linear VC dimension.

7 Related work and discussion

List learnability. Charikar & Pabbaraju (2023) introduce the list-DS dimensions and prove that finite k-DS dimension characterizes k-list PAC learnability. Brukhim et al. (2023) give a boosting-based proof and identify optimal weak accuracy with reciprocal minimal list size. Hanneke et al.

(2026) prove the sharp Sauer inequality for list-DS dimension. Pabbaraju (2026) establishes the optimal DS dependence of realizable multiclass and list sample complexity. Our result uses the qualitative characterization at the single boundary $k-1$ versus k, then derives a product theorem that is not a dimension inequality.

Direct sums. The product operation and open questions are due to Hanneke et al. (2024b), whose longer list-compression work develops the pseudo-cube product bounds behind the previous interval for K (Hanneke et al., 2024a). Nachum et al. (2018) prove a direct sum for the information complexity of proper consistent binary learning. Suruga (2024) proves direct-sum theorems for oracle and query complexity. These resources, losses, and protocols differ from the list-risk spectrum studied here.

Limitations. The exact spectrum is information-theoretic and may use improper, randomized, computationally inefficient learners. Randomization is needed for the concise upper spectrum below the critical list size, though the threshold identity for K does not depend on it. The fixed-marginal equality is stated with its minimax regularity condition rather than assuming that every infinite measurable game has a least-favorable finite prior. The agnostic and uniform corollaries retain logarithmic gaps and nondegeneracy conditions.

8 Conclusion

The least PAC list size tensorizes exactly under Cartesian products, and the threshold is only one point on a complete linear risk spectrum. The proof isolates a reusable mechanism: adjacent Bayes list risks reveal a posterior order statistic, and a product of those order statistics controls every omitted cell in a top-posterior grid. This also turns weak multiclass accuracy into an exactly multiplicative resource. Fixed-marginal ordinary risk obeys a different all-coordinate tensorization, while unrestricted agnostic curves require more structural information than their one-factor rates. Together, these results resolve the sharpest qualitative direct-sum question and delineate which learning resources multiply, add, or admit no curve-only law.

Reproducibility statement

All assumptions and minimax quantities are defined in Section 2. Section 4 proves the main theorem. The appendices prove the regularity, fixed-marginal, and dimension claims. Appendix D describes finite checks of the two algebraic inequalities used in the paper.

References

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science*, pp. 943–955, 2022.

Nataly Brukhim, Amit Daniely, Yishay Mansour, and Shay Moran. Multiclass boosting: Simple and intuitive weak learning criteria. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1403–1425, 2023.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1713–1726, 2023. doi: 10.1145/3564246.3585190.

Alon Cohen, Liad Erez, Steve Hanneke, Tomer Koren, Yishay Mansour, Shay Moran, and Qian Zhang. Sample complexity of agnostic multiclass classification: Natarajan dimension strikes back. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*, pp. 722–733, 2026. doi: 10.1145/3798129.3800787.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 287–316, 2014.

Steve Hanneke, Shay Moran, and Tom Waknine. List sample compression and uniform convergence. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2360–2388, 2024a.

Steve Hanneke, Shay Moran, and Tom Waknine. Open problem: Direct sums in learning theory. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 5325–5329, 2024b.

Steve Hanneke, Qinglin Meng, Shay Moran, and Amirreza Shaeiri. An optimal sauer lemma over k-ary alphabets. *arXiv preprint arXiv:2604.12952*, 2026.

Ron Holzman, Shay Moran, and Alexander Shlimovich. Uniform laws of large numbers in product spaces. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3224–3279, 2026.

Mihir More, Aritra Das, and Debayan Gupta. A rate separation for agnostic direct sums. *arXiv preprint arXiv:2608.06951*, 2026.

Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. A direct sum result for the information complexity of learning. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1547–1568, 2018.

Balas K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

Chirag Pabbaraju. The optimal sample complexity of multiclass and list learning. *arXiv preprint arXiv:2604.24749*, 2026.

Daiki Suruga. Direct sum theorems beyond query complexity. *arXiv preprint arXiv:2408.15570*, 2024. Revised January 2025.

Aad W. van der Vaart and Jon A. Wellner. A note on bounds for VC dimensions. In *High Dimensional Probability V: The Luminy Volume*, volume 5 of *Institute of Mathematical Statistics Collections*, pp. 103–107. Institute of Mathematical Statistics, 2009. doi: 10.1214/09-IMSCOLL508.

A Expected risk and PAC learnability

We record the equivalence used to define K. Suppose a fixed k-list learner has the PAC property. For any $a > 0$, request population error at most a with failure probability at most a. Since loss is bounded by one, its expected population error is at most $2a$ for all sufficiently large samples. Taking the infimum over learners gives $\mathfrak{R}_k(n; \mathcal{H}) \to 0$.

Conversely, choose a learner with worst-case expected error at most $2\mathfrak{R}_k(n; \mathcal{H})$. Markov's inequality gives

$$\sup_{D,h} \mathbb{P}\{\mathrm{err}_{D_h}(A(S)) > \varepsilon\} \leq \frac{2\mathfrak{R}_k(n; \mathcal{H})}{\varepsilon}.$$

For any $\varepsilon, \delta > 0$, the right side is at most δ for all sufficiently large n. Optimization slack can be sent to zero. Thus vanishing minimax expectation and list PAC learnability are equivalent.

If one factor has $K = \infty$, suppose its product with nonempty classes were L-list learnable. Fix one target and a point-mass marginal in every other coordinate. Simulate the product learner using the hard factor sample, append the fixed coordinates, and project every predicted tuple onto the hard coordinate. Projection produces at most L labels. Whenever the projected list misses the hard target, the original tuple list also misses. This contradicts $K = \infty$.

B Proof of fixed-marginal tensorization

For factor j, let an algorithm have worst-target error at most $R_j + \eta$. Run the algorithms with independent random seeds on the coordinate projections of the product sample. For a fixed target tuple, the factor samples, test coordinates, and seeds are independent. The probability that every

coordinate is correct is therefore the product of the factor correctness probabilities and is at least $\prod_j (1 - R_j - \eta)$. Letting $\eta \downarrow 0$ proves the upper bound in (8).

For the lower bound, fix finitely supported priors Π_j on the targets. After observing (S_j, X_j), let $m_j(S_j, X_j)$ be the largest posterior probability of a factor label. Under the product prior and product marginal, the posterior of the tuple label factorizes. Its largest atom has mass $\prod_j m_j$. Hence the product Bayes error is

$$1 - \mathbb{E} \prod_j m_j = 1 - \prod_j \mathbb{E} m_j = 1 - \prod_j (1 - \mathfrak{B}_1(\Pi_j)).$$

The minimax product risk is at least every Bayes risk. Taking each prior arbitrarily close to its finite-prior Bayes envelope gives the lower bound in (8).

If $\underline{R}(n; D_j, \mathcal{H}_j) = R(n; D_j, \mathcal{H}_j)$ for every factor, the bounds coincide. This equality follows directly from finite zero-sum minimax when the induced statistical game is finite. It also holds under the standard compactness and lower-semicontinuity hypotheses of statistical decision theory. The sandwich remains valid without any such regularity.

C Product dimensions and learning rates

C.1 Graph dimension

The graph dimension of $\mathcal{H}$ is the VC dimension of its correctness class $\{(x, y) \mapsto \mathbf{1}\{h(x) = y\} : h \in \mathcal{H}\}$. For the product class, correctness is the AND of the lifted factor correctness indicators, or equivalently product loss is the OR of factor losses. On any m product examples, the number of product loss patterns is at most the product of the factor growth functions. If the positive graph dimensions are g_j and $G = \sum_j g_j$, Sauer's lemma gives

$$\Pi_{\otimes_j \mathcal{H}_j}(m) \le \prod_j \left(\frac{em}{g_j}\right)^{g_j}.$$

The standard VC union calculation (van der Vaart & Wellner, 2009) implies $d_{\mathrm{G}}(\bigotimes_j \mathcal{H}_j) \le CG \log(2r)$.

For the lower bound, take a graph-shattered set of size g_j in each factor with $g_j \ge 1$, and reserve one point as its anchor. For every nonanchor point of factor j, create a product point using that point in coordinate j and anchors in all other coordinates. To realize any binary pattern on the resulting $\sum_j (g_j - 1)_+$ points, ask each factor hypothesis to match its witness on the anchor and on exactly the selected nonanchors. All other coordinates match their anchors, so product correctness on a point is controlled by its unique nonanchor coordinate. Therefore

$$\sum_j (g_j - 1)_+ \le d_{\mathrm{G}}\left(\bigotimes_j \mathcal{H}_j\right) \le C \left(\sum_j g_j\right) \log(2r).$$

Classical VC lower and upper bounds for expected uniform deviation give the rates stated in Section 6.

C.2 DS, Natarajan, and realizable dimensions

Let $d_j = d_{\mathrm{DS}}(\mathcal{H}_j)$ and $q_j = d_{\mathrm{N}}(\mathcal{H}_j)$. The constructions of Hanneke et al. (2024b) iterate to

$$d_{\mathrm{DS}}\left(\bigotimes_j \mathcal{H}_j\right) \ge \sum_j d_j - (r - 1), \tag{9}$$

$$\sum_j q_j - 2(r - 1) \le d_{\mathrm{N}}\left(\bigotimes_j \mathcal{H}_j\right) \le \sum_j q_j. \tag{10}$$

The realizable dimension $d_{\mathrm{RE}}(\mathcal{H})$ of Cohen et al. (2026) is the minimax sample size needed for a fixed constant expected error. Pabbaraju (2026) prove the realizable tail bound corresponding to sample complexity $O((d_{\mathrm{DS}} + \log(1/\delta))/\varepsilon)$. Integrating this tail gives a factor learner with expected error $O(d_j/n)$. Running these learners on the coordinate projections gives product expected error $O((\sum_j d_j)/n)$. Hence

$$d_{\mathrm{RE}}\left(\bigotimes_j \mathcal{H}_j\right) \le C \sum_j d_j.$$

Conversely, realizable dimension is at least a constant multiple of DS dimension. Equation (9) supplies the matching lower bound whenever $\sum_j d_j - (r-1)$ is a constant fraction of $\sum_j d_j$.

The main theorem of Cohen et al. (2026), sharpened using Pabbaraju (2026), states

$$M_{\mathrm{agn}}(\varepsilon, \delta; \mathcal{G}) = \widetilde{\Theta}\left(\frac{d_{\mathrm{RE}}(\mathcal{G})}{\varepsilon} + \frac{d_{\mathrm{N}}(\mathcal{G}) + \log(1/\delta)}{\varepsilon^2}\right).$$

Apply this to $\mathcal{G} = \mathcal{H}^r$. If $d = d_{\mathrm{DS}}(\mathcal{H}) \ge 2$, then (9) is $\Theta(rd)$. If $q = d_{\mathrm{N}}(\mathcal{H}) \ge 3$, then (10) is $\Theta(rq)$. This proves Corollary 6. Standard inversion of the two sample-complexity terms gives the expected excess-risk display, up to logarithms and clipping at one.

D FINITE VERIFICATION CHECKS

The proof of Theorem 1 is analytic. We nevertheless checked its posterior inequality on 20,000 independent triples of random finite probability vectors. For each trial, the script selected three list ranks and checked zero budget, the threshold-minus-one budget, and one random intermediate budget. After duplicate budgets were removed, 54,276 inequalities of the form

$$1 - \sum_{a=1}^{L} q_{(a)} \ge (M - L) \prod_j p_{j,(k_j)}$$

were tested, where q is the flattened product posterior. The minimum slack was -5.3×10^{-16}, within floating-point roundoff.

We also solved eight exact randomized minimax linear programs for all binary concepts on a two-point domain. The cases use $n \in \{0, 1\}$, product powers $r \in \{2, 3\}$, and marginals $(1/2, 1/2)$ and $(0.7, 0.3)$. Every instance matched the fixed-marginal prediction $1 - (1 - R)^r$, with maximum absolute discrepancy 3.4×10^{-16}.