TWO DIMENSIONS GOVERN AGNOSTIC MULTICLASS TRANSDUCTION

Pahan Dewasurendra
Johns Hopkins University

Nuwan Dewasurendra
MiraCosta College

ABSTRACT

In transductive classification, an adversary fixes a labeled population, one label is hidden uniformly, and the learner sees all remaining labels. For binary classes, agnostic transductive and PAC learning have the same minimax rate. Whether this extends to multiclass learning was open, especially for unbounded label spaces where uniform convergence can fail.

We resolve the question up to logarithmic factors. For every multiclass class $\mathcal{H}$ with DS dimension d_{DS} and Natarajan dimension d_{N}, the optimal agnostic transductive excess error satisfies

$$\widetilde{\Theta}\left(\frac{d_{\mathrm{DS}}}{n}+\sqrt{\frac{d_{\mathrm{N}}}{n}}\right).$$

The result holds for arbitrary label spaces. The two terms are both necessary. A DS pseudo-cube gives the realizable d_{DS}/n obstruction, while a Natarajan cube with repeated points and fair labels gives the agnostic $\sqrt{d_{\mathrm{N}}/n}$ obstruction.

The upper bound uses a random-reservation principle. The learner deliberately ignores a constant fraction of the visible labels, which makes the true test point uniform in a large unseen block. We combine realizable compression, a label-space reduction, and inside-menu agnostic compression across this finite-population split. A new without-replacement multiplicative-weights lemma preserves the fast d_{DS}/n term. Consequently, agnostic multiclass PAC and transductive learning obey the same two-dimension law up to logarithmic factors.

1 INTRODUCTION

The transductive model isolates one of the simplest prediction problems with dependent data. An adversary chooses n labeled examples. One index is hidden uniformly at random. The learner sees the full unlabeled population and every label except the hidden one, then predicts that label. Its loss is compared with the empirical loss of the best fixed hypothesis in a benchmark class.

For realizable classification, transductive learning is tightly connected to one-inclusion graphs and to PAC learning (Haussler et al., 1994; Daniely & Shalev-Shwartz, 2014; Asilis et al., 2024). The agnostic setting is subtler. A learner must compete additively with the best hypothesis on each adversarial population. Generic transformations from PAC learners to transductive learners lose a factor of $1/\varepsilon$ in sample complexity (Asilis et al., 2024). In the other direction, a recent transformation from transductive to PAC learning has only a class-independent validation cost (Dughmi et al., 2025). For binary classification, a symmetrization of the agnostic one-inclusion graph completes the converse and gives the optimal rate $\Theta(\sqrt{d/n})$, where d is VC dimension. The corresponding multiclass question was left open.

Multiclass learning over an arbitrary label space has an additional difficulty: no single dimension controls its agnostic sample complexity. The DS dimension characterizes realizable learnability (Daniely & Shalev-Shwartz, 2014; Brukhim et al., 2022). Recent work showed that the Natarajan dimension reappears in the small-excess agnostic regime (Cohen et al., 2026). A subsequent sharp density theorem removed the remaining polynomial gap in the DS term (Pabbaraju, 2026). Combin-

ing these results, the agnostic PAC sample complexity is, up to logarithms,

$$\frac{d_{\mathrm{DS}}}{\varepsilon} + \frac{d_{\mathrm{N}}}{\varepsilon^2}.$$

This progress does not itself yield a transductive learner. PAC proofs average over independent samples, while transductive guarantees must hold for every fixed population. In particular, empirical minimization over a finite reconstructed family can have constant leave-one-out error even when the family has only n members. This is the instability behind the open problem.

We show that the two models nevertheless have the same multiclass rate, up to logarithmic factors. Let $\epsilon^{\mathrm{tr}}_{\mathcal{H}}(n)$ denote the optimal agnostic transductive excess error. Our main theorem gives

$$c \min\left\{1, \frac{d_{\mathrm{DS}}}{n} + \sqrt{\frac{d_{\mathrm{N}}}{n}}\right\} \le \epsilon^{\mathrm{tr}}_{\mathcal{H}}(n) \le C \min\left\{1, \frac{d_{\mathrm{DS}} \log^2(en)}{n} + \sqrt{\frac{d_{\mathrm{N}} \log^3(en)}{n}}\right\}.$$

The result permits an infinite or uncountable label space.

The upper bound rests on a simple change in how the visible sample is used. Given the hidden index, the learner chooses three disjoint random blocks from the $n-1$ visible examples and ignores all other visible labels. Equivalently, the three blocks can be chosen first from the full population, after which the hidden index is uniform in their complement. The ignored points turn the single hidden label into a uniformly random point of a large unseen block. This permits finite-population generalization without requiring leave-one-out stability.

We then adapt the three-stage label-space reduction of Cohen et al. (2026). The first block creates a finite cover using realizable compression. The second runs multiplicative weights to reduce the pointwise label space. The third performs inside-menu agnostic compression. Two standard compression arguments transfer to random splits. The key new ingredient is a finite-population multiplicative-weights lemma. Although the examples arrive without replacement, the expected number of rewards earned by sampled experts is constant because each expert can newly cover only a disjoint part of the population. This preserves the $1/n$ approximation term.

The lower bounds reveal why both dimensions remain necessary in the transductive model. For the Natarajan term, repeat each shattered point and independently assign one of its two witness labels. The hidden label is an independent fair bit, while the best hypothesis chooses each block majority after seeing the entire population. For the DS term, orient the coordinate fibers of a pseudo-cube. Every orientation has average outdegree proportional to the pseudo-cube dimension, even after the population is padded by repetitions.

Our result answers the multiclass instance of the open question of Dughmi et al. (2025). It is specific enough to exploit the structure of multiclass learning. We do not give a black-box PAC-to-transductive reduction for every bounded loss. The random-reservation argument may, however, be useful whenever a PAC construction factors through small reconstructed families.

Contributions.

1. We characterize agnostic multiclass transductive error up to logarithmic factors by separate DS and Natarajan terms, for arbitrary label spaces.
2. We introduce a random-reservation reduction that transfers a structured three-stage PAC learner to every fixed finite population.
3. We prove a without-replacement multiplicative-weights menu lemma with miss probability $O(\log |F|/n)$.
4. We give direct transductive lower bounds from Natarajan cubes and DS pseudo-cubes.

2 Setting and Main Result

Let X be an instance space, let Y be an arbitrary label space, and let $\mathcal{H} \subseteq Y^X$ be nonempty. A labeled population is $S = (z_1, \dots, z_n) \in (X \times Y)^n$, where $z_i = (x_i, y_i)$. Repeated instances and

repeated labeled examples are allowed. Write

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}.$$

A randomized transductive learner receives the full instance sequence $x_{1:n}$, an index i, and the labeled sequence S_{-i}. It outputs a label for x_i. Its agnostic excess error on S is

$$\mathcal{E}_S(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_A\{A(x_{1:n}, S_{-i}, i) \neq y_i\} - \min_{h\in\mathcal{H}} L_S(h). \tag{1}$$

The optimal worst-population error is

$$\epsilon^{\mathrm{tr}}_{\mathcal{H}}(n) = \inf_{A} \sup_{S\in(X\times Y)^n} \mathcal{E}_S(A).$$

This agrees with the agnostic transductive criterion of Asilis et al. (2024) and Dughmi et al. (2025).

We recall the two dimensions. A sequence $x_{1:d}$ is Natarajan shattered if there are pairs $a_j \neq b_j$ such that $\mathcal{H}|_{x_{1:d}}$ contains every vector in $\prod_j\{a_j, b_j\}$. The largest d is the Natarajan dimension d_{N} (Natarajan, 1989). A finite nonempty set $F \subseteq Y^d$ is a d-dimensional pseudo-cube if every $f \in F$ has a neighbor differing only in coordinate j, for every $j \in [d]$. The largest d for which a trace of $\mathcal{H}$ contains such a pseudo-cube is the DS dimension d_{DS} (Daniely & Shalev-Shwartz, 2014; Brukhim et al., 2022).

Theorem 1 (Two-dimension transductive law)**.** *There are universal constants $c, C > 0$ such that every nonempty $\mathcal{H} \subseteq Y^X$ and every $n \geq 2$ satisfy*

$$\epsilon^{\mathrm{tr}}_{\mathcal{H}}(n) \geq c\min\left\{1, \frac{d_{\mathrm{DS}}}{n} + \sqrt{\frac{d_{\mathrm{N}}}{n}}\right\}, \tag{2}$$

$$\epsilon^{\mathrm{tr}}_{\mathcal{H}}(n) \leq C\min\left\{1, \frac{d_{\mathrm{DS}}\log^2(en)}{n} + \sqrt{\frac{d_{\mathrm{N}}\log^3(en)}{n}}\right\}. \tag{3}$$

The conventions are that the upper bound is vacuous when $d_{\mathrm{DS}} = \infty$, and that a class of Natarajan dimension zero has zero transductive excess error.

Define $m^{\mathrm{tr}}_{\mathcal{H}}(\varepsilon) = \min\{n : \sup_{N\geq n} \epsilon^{\mathrm{tr}}_{\mathcal{H}}(N) \leq \varepsilon\}$. Inverting the rates gives the sample-complexity form.

Corollary 2 (PAC and transductive rates)**.** *Up to polylogarithmic factors,*

$$m^{\mathrm{tr}}_{\mathcal{H}}(\varepsilon) = \widetilde{\Theta}\left(\frac{d_{\mathrm{DS}}}{\varepsilon} + \frac{d_{\mathrm{N}}}{\varepsilon^2}\right).$$

This is the same two-parameter law as agnostic PAC learning, obtained by combining Cohen et al. (2026) and Pabbaraju (2026).

Why a generic ERM transfer fails. Suppose a learner trains an empirical minimizer over a finite family on the $n-1$ visible labels. Even with one perfect comparator and only n additional hypotheses, tie-breaking can choose, for each held-out index i, a hypothesis that is correct everywhere except at i. Its leave-one-out error is one. A finite cardinality bound is useful only when the test point lies in a large block that was not used to choose the output. Random reservation creates precisely this block.

3 Three finite-population lemmas

We state the tools used by the upper bound. All populations below are finite multisets of indexed examples. Uniform samples are over indices, so duplicate values cause no ambiguity.

3.1 REALIZABLE COMPRESSION COVER

A deterministic selection scheme $\mathcal{A} = (\kappa, \rho)$ of size $k(m)$ selects an ordered tuple of at most $k(m)$ input examples, with repetition allowed, and reconstructs a predictor from that tuple. It is a realizable compression scheme if its reconstruction realizes every realizable input sequence. Recent one-inclusion and boosting results imply the following fact.

For a loss ℓ, an ℓ-sample compression scheme additionally requires its reconstruction on every input sequence to have empirical ℓ-loss no larger than the empirical ℓ-loss of any comparator in $\mathcal{H}$. Thus empirical optimality is part of the guarantee, not an extra property of the reconstruction.

Proposition 3 (Compression ingredients)**.** *If $\mathcal{H}$ has finite DS dimension d_{DS}, it admits a deterministic realizable compression scheme of size*

$$k_1(m) \le C d_{\mathrm{DS}} \log(em).$$

For every menu $\mu : X \to 2^Y$ of size at most p, $\mathcal{H}$ admits a deterministic sample compression scheme for

$$\ell_\mu(g, (x, y)) = \mathbf{1}\{y \in \mu(x),\ g(x) \neq y\}$$

of size

$$k_2(m) \le C d_{\mathrm{N}} \log(ep) \log(em).$$

Both statements hold for arbitrary label spaces.

The first statement combines the density theorem of Pabbaraju (2026) with the weak-learning-to-compression construction of David et al. (2016). The exact construction is also given by Cohen et al. (2026, Corollary C.2). The second statement is Cohen et al. (2026, Proposition 3.6). Appendix A records the reduction and the infinite-label point.

For a fixed comparator h, let $A(h)$ denote the examples in A correctly labeled by h. Applying the first scheme to $A(h)$ produces a predictor that may disagree with h, but not on any point of A where h is correct.

Lemma 4 (Finite-population compression cover)**.** *Let R be a fixed population of size N, let A be a uniform size-m subset with $m \le N/2$, and let $\mathcal{A} = (\kappa, \rho)$ be a deterministic realizable compression scheme of size at most k. For fixed h, put $f_{h,A} = \mathcal{A}(A(h))$. Then*

$$\mathbb{E}_A \frac{1}{N-m} \sum_{(x,y) \in R \setminus A} \mathbf{1}\{h(x) = y \neq f_{h,A}(x)\} \le C \frac{(k+1)\log(eN)}{m}.$$

The proof is a compression union bound, but over the unseen finite population rather than a distribution. Every possible output is reconstructed from one of at most $(k+1)N^k$ tuples. If its bad fraction on $R \setminus A$ is u, the probability that A avoids all bad points is at most $e^{-um/2}$. We give details in Appendix B.

3.2 A WITHOUT-REPLACEMENT MENU LEMMA

Let $F \subseteq Y^X$ be finite and let $Z_1, \dots, Z_T$ be a random ordered sample without replacement from a fixed population R of size N. Multiplicative weights starts with uniform weights. Before seeing $Z_t = (X_t, Y_t)$, it draws f_t from the current distribution p_t. It assigns every $f \in F$ reward

$$r_t(f) = \mathbf{1}\{f(X_t) = Y_t,\ f_s(X_t) \neq Y_t \text{ for all } s < t\} \tag{4}$$

and updates by $w_{t+1}(f) = w_t(f) e^{r_t(f)/2}$. Its final menu is $\mu(x) = \{f_1(x), \dots, f_T(x)\}$.

Lemma 5 (Finite-population MW menu)**.** *If $T \le N/2$ and $U = R \setminus \{Z_1, \dots, Z_T\}$, then every $f^\star \in F$ satisfies*

$$\mathbb{E} \frac{1}{|U|} \sum_{(x,y) \in U} \mathbf{1}\{f^\star(x) = y,\ y \notin \mu(x)\} \le \frac{2 \log |F| + 3}{T}.$$

This is the main new technical lemma. Its proof has two parts. First, the expected reward of $f^\star$ at every round upper bounds its final unseen miss rate. This follows from exchangeability of the remaining permutation and monotonicity as the menu grows. Second, the expected cumulative

reward of the sampled experts is at most two. To see this, define $D_t \subseteq R$ as the points newly covered by f_t. The sets D_t are disjoint. Conditional on the past and f_t, the reward probability is at most $|D_t|/(N-T+1)$. Summing gives at most $N/(N-T+1) \le 2$. The multiplicative-weights regret inequality completes the proof. Appendix C is formal.

3.3 Agnostic compression across a split

The final tool converts empirical optimality on one random block into excess control on another.

Lemma 6 (Random-split agnostic compression)**.** *Let a fixed population R of size N be split uniformly into C and V, where $|C|, |V| \ge N/3$. Let $\ell \in [0,1]$, and let a deterministic ℓ-sample compression scheme of size at most k output $\widehat{h}$ on C. Then every fixed comparator h in the benchmark class satisfies*

$$\mathbb{E}\big[L_V^\ell(\widehat{h}) - L_V^\ell(h)\big] \le C\sqrt{\frac{(k+1)\log(eN)}{N}}.$$

Conditional on R, every possible output lies in a family of size at most $(k+1)N^k$. Hoeffding's comparison theorem for sampling without replacement (Hoeffding, 1963; Serfling, 1974) and a union bound control the difference between its losses on C and V. Empirical optimality on C then gives the result. Appendix D includes the calculation.

4 The transductive learner

For n below a universal constant, use an arbitrary predictor and enlarge the universal constant in Theorem 1. Henceforth fix a population S of sufficiently large size n, and let I be the uniform hidden index. Put $q = \lfloor n/4 \rfloor$. From the visible indices $[n] \setminus \{I\}$, choose disjoint blocks A, B, C, each of size q, uniformly in sequence. Give every block an independent uniform order. The learner ignores every visible label outside these blocks.

The joint distribution has an equivalent deferred-decision form:

$$A, B, C \text{ are chosen first from } S, \qquad V = S \setminus (A \cup B \cup C), \qquad I \sim \mathrm{Unif}(V). \tag{5}$$

Both orders assign the same probability to every disjoint tuple (A, B, C, I). Thus the expected test loss conditional on (A, B, C) is exactly $L_V(\widehat{h})$.

The learner uses Proposition 3 in three stages.

Stage 1: finite cover. Let (κ_1, ρ_1) be the realizable compression scheme. Construct

$$F_A = \left\{ \rho_1(t) : t \in \bigcup_{j=0}^{k_1(q)} A^j \right\}.$$

Then $|F_A| \le (q+1)^{k_1(q)+1}$. For every $h \in \mathcal{H}$, the reconstruction $f_{h,A} = \mathcal{A}_1(A(h))$ belongs to F_A.

Stage 2: label-space reduction. Run the procedure of Lemma 5 on the ordered block B, with expert set F_A and $T = q$. This produces a menu μ of size at most q.

Stage 3: inside-menu learning. Run the inside-menu compression scheme on C and output its reconstructed classifier $\widehat{h}$. Predict $\widehat{h}(x_I)$.

The algorithm is information valid. It selects A, B, C from visible examples and never uses labels in V, even when they are available.

Proof of the upper bound in Theorem 1. The result is trivial when $d_{\mathrm{N}} = 0$, so assume $d_{\mathrm{N}} \ge 1$. Fix

$$h^\star \in \arg\min_{h \in \mathcal{H}} L_S(h).$$

An empirical minimizer exists because the possible empirical errors lie in $\{0, 1/n, \dots, 1\}$.

Lemma 4, followed by random subdivision of $S \setminus A$, gives

$$\mathbb{E}L_V\big(\mathbf{1}\{h^\star(x) = y \neq f_{h^\star,A}(x)\}\big) \leq C\frac{d_{\mathrm{DS}}\log^2(en)}{n}. \tag{6}$$

Conditional on A, Lemma 5 applies to the remaining population because $q \leq (n-q)/2$. Since

$$\log|F_A| \leq Cd_{\mathrm{DS}}\log^2(en),$$

another subdivision step gives

$$\mathbb{E}L_V\big(\mathbf{1}\{f_{h^\star,A}(x) = y,\ y \notin \mu(x)\}\big) \leq C\frac{d_{\mathrm{DS}}\log^2(en)}{n}. \tag{7}$$

Conditional on (A, B, μ), the blocks C, V form a uniform split of the remaining population, with both sizes at least one third of it. The inside-menu scheme has size

$$k_2(q) \leq Cd_{\mathrm{N}}\log^2(en).$$

Lemma 6 yields

$$\mathbb{E}\left[L_V^\mu(\widehat{h}) - L_V^\mu(h^\star)\right] \leq C\sqrt{\frac{d_{\mathrm{N}}\log^3(en)}{n}}. \tag{8}$$

For every fixed menu and population,

$$\begin{aligned} L_V(\widehat{h}) - L_V(h^\star) \leq{}& L_V^\mu(\widehat{h}) - L_V^\mu(h^\star) \\ &+ L_V\big(\mathbf{1}\{h^\star(x) = y,\ y \notin \mu(x)\}\big). \end{aligned} \tag{9}$$

The final miss event is contained in the union of the events in (6) and (7). Combining the displays gives the right side of (3) without the cap at one.

Finally, (5) makes the expected test loss equal to $\mathbb{E}L_V(\widehat{h})$. The set V is a uniform subset of S, so $\mathbb{E}L_V(h^\star) = L_S(h^\star)$. The trivial upper bound one supplies the cap. □

5 Matching lower bounds

We prove the two terms separately. Since their maximum is at least half their sum, this proves (2).

5.1 The Natarajan obstruction

Let $d = \min\{d_{\mathrm{N}}, n\}$, and take Natarajan-shattered points $x_1, \ldots, x_d$ with witness labels $a_j \neq b_j$. Partition the n positions into nonempty blocks of sizes $k_j \in \{\lfloor n/d \rfloor, \lceil n/d \rceil\}$. Every position in block j has feature x_j. Label its copies independently and uniformly from $\{a_j, b_j\}$.

Conditional on every visible label, the hidden label remains a fair independent bit. Every learner therefore has expected error at least $1/2$. If $B_j \sim \mathrm{Bin}(k_j, 1/2)$, the best hypothesis makes $\min\{B_j, k_j - B_j\}$ errors in block j. Natarajan shattering realizes all blockwise majority choices. A hypothesis using a third label on x_j is dominated by one of these choices. Hence the expected excess is at least

$$\frac{1}{2n}\sum_{j=1}^{d}\mathbb{E}|2B_j - k_j|.$$

Khinchine's inequality gives $\mathbb{E}|2B_j - k_j| \geq \sqrt{k_j/2}$. Since $k_j \geq n/(2d)$, the excess is at least $c\sqrt{d/n}$. The probabilistic method fixes one deterministic labeling with at least this error.

5.2 The pseudo-cube obstruction

Let $d = \min\{d_{\mathrm{DS}}, n\}$. Project a DS witness onto any d witness coordinates. After duplicate vertices are removed, the projection is still a pseudo-cube. For $d \geq 2$, put one sample copy at each $x_2, \ldots, x_d$, and use every remaining position as a copy of x_1. Label the population by a vertex of the pseudo-cube, making it realizable.

Fix a possibly randomized learner. For every direction $i \geq 2$, vertices that agree outside coordinate i form a nontrivial fiber e. When x_i is hidden, all vertices in e give the learner exactly the same observation. Their labels at x_i are distinct. Consequently, the sum of their error probabilities is at least $|e| - 1$. Summing over all fibers,

$$\sum_e (|e| - 1) \geq \frac{1}{2} \sum_e |e| = \frac{|F|(d-1)}{2},$$

where F is the pseudo-cube. Some realizable vertex causes at least $(d-1)/2$ errors across the n possible hidden positions. Thus

$$\epsilon^{\mathrm{tr}}_{\mathcal{H}}(n) \geq \frac{d-1}{2n}.$$

For $d = 1$, the Natarajan obstruction already dominates the desired $1/n$ term. Appendix F gives the projection and randomized-orientation details.

6 Related work and discussion

Agnostic one-inclusion graphs. Asilis et al. (2024) formulate multiclass agnostic transduction as orienting a Hamming hypergraph with vertex credits equal to distance from $\mathcal{H}$. Their agnostic Hall complexity exactly characterizes optimal transductive error, but it is not bounded there by statistical dimensions. Dughmi et al. (2025) solve the binary case by showing that symmetrization enlarges discounted density until the full Boolean cube is reached, then invoke Rademacher complexity. That symmetrization is specific to two labels. Our proof instead constructs a transductive learner and exposes the separate DS and Natarajan roles.

Agnostic multiclass PAC learning. Cohen et al. (2026) introduce the three-stage cover, multiplicative-weights, and inside-menu architecture. Their bound is expressed through the optimal constant-error realizable complexity. Pabbaraju (2026) prove that one-inclusion density is at most DS dimension, closing the realizable gap and yielding the two-dimension PAC rate. Our random-reservation analysis transfers this architecture to a fixed population. The finite-population menu lemma is essential for retaining the fast DS term.

Other transductive models. Multiclass transductive online learning reveals the full instance sequence and then predicts many labels sequentially. Its rates are governed by level-constrained Littlestone-type dimensions (Hanneke et al., 2024; Hanneke & Wang, 2026). This differs from the batch leave-one-out model studied here. Recent work separates local regularizers from general multiclass transductive learners (Jafar et al., 2025), and a subsequent construction gives the corresponding separation for fixed local scores in PAC learning (Hou, 2026). Our learner is highly improper and does not contradict either separation.

Limitations and open questions. The logarithmic gap comes from boosting weak learners into compression, enumerating reconstructed families, and inside-menu compression. Removing it may require a direct weighted orientation of the multiclass agnostic Hamming hypergraph. Our result is information-theoretic. Like the underlying one-inclusion and compression constructions, it need not be computationally efficient. It also does not provide a black-box reduction for arbitrary bounded losses.

7 Conclusion

Agnostic multiclass transduction has the same two statistical scales as agnostic PAC learning. DS dimension controls the fast approximation term, while Natarajan dimension controls the square-root estimation term. Randomly reserving a large unseen block turns structured compression arguments into finite-population guarantees and bypasses the instability of direct leave-one-out transfer. This closes the PAC versus transductive question for multiclass classification up to logarithmic factors, including for unbounded label spaces.

Reproducibility statement

The setting, learner, and assumptions are specified in Sections 2–4. Complete proofs of every stated lemma and theorem are included after the references, including exact partition bookkeeping and randomized lower-bound quantifiers.

References

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Regularization and optimal multiclass learning. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 260–310, 2024.

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.

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.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Shaddin Dughmi, Yusuf Hakan Kalayci, and Grayson York. Is transductive learning equivalent to PAC learning? In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 418–443, 2025.

Steve Hanneke and Hongao Wang. Universal multiclass transductive online learning. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2026.

Steve Hanneke, Vinod Raman, Amirreza Shaeiri, and Unique Subedi. Multiclass transductive online learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting {0,1}-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Eric Hou. Local regularization does not characterize multiclass PAC learnability. *arXiv preprint arXiv:2607.23449*, 2026.

Sky Jafar, Julian Asilis, and Shaddin Dughmi. Local regularizers are not transductive learners. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2942–2957, 2025.

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.

Robert J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

A Compression ingredients

We give a self-contained derivation of the consequences collected in Proposition 3.

A.1 Realizable compression from DS dimension

If $d_{\mathrm{DS}} = 0$, every hypothesis has the same pointwise behavior and compression size zero suffices. Assume $d_{\mathrm{DS}} \geq 1$. The density theorem of Pabbaraju (2026, Theorem 1) and the standard one-inclusion orientation construction give the deterministic realizable learner in their Corollary 1.1. For $m_0 \leq C d_{\mathrm{DS}}$ training examples, run that learner with target error and failure probability both $1/6$. For every realizable distribution P, its bounded loss then satisfies

$$\mathbb{E}_{T \sim P^{m_0}} \operatorname{err}_P(A_0(T)) \leq \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

The density statement extends to infinite label spaces by their Remark 2, and the learner extends by the compactness argument in their Remark 4. Thus A_0 is a deterministic constant-error weak learner using $O(d_{\mathrm{DS}})$ examples for every label space.

The minimax and boosting construction of David et al. (2016), stated explicitly as Cohen et al. (2026, Corollary C.2), converts a deterministic weak learner using $O(d_{\mathrm{DS}})$ examples into a realizable compression scheme of size $O(d_{\mathrm{DS}} \log(en))$ on sequences of length at most n. Its compression tuple may repeat input examples, which is why our counting arguments use N^k rather than binomial coefficients.

A.2 Inside-menu compression

Fix a menu $\mu : X \to 2^Y$ of size p. For $h \in \mathcal{H}$, define a partial hypothesis that equals $h(x)$ when $h(x) \in \mu(x)$ and is undefined otherwise. On every finite support, this partial class has Natarajan dimension at most d_{N} and at most p active labels per point. Its one-inclusion learner has leave-one-out error $O(d_{\mathrm{N}} \log(ep)/m)$. Applying the same weak-learning-to-compression construction gives an ℓ_μ-sample compression scheme of size

$$O(d_{\mathrm{N}} \log(ep) \log(em)).$$

This is the construction and proof of Cohen et al. (2026, Proposition 3.6). The reconstruction depends only on the selected tuple, $\mathcal{H}$, and μ. It uses no additional side information. When $p = 1$, the unique menu label has zero inside-menu loss and no examples are needed.

B Proof of the finite-population compression cover

Proof of Lemma 4. Every possible reconstruction has the form $\rho(t)$, where t is an ordered tuple of elements of R of length at most k. Repetition is allowed. There are at most

$$M = \sum_{j=0}^{k} N^j \leq (k+1)N^k$$

such tuples.

Fix one tuple t, put $f = \rho(t)$, and define the indexed bad set

$$D_t = \{(x, y) \in R : h(x) = y \neq f(x)\}.$$

If $f = f_{h,A}$, then $A \cap D_t = \varnothing$, because $f_{h,A}$ realizes $A(h)$. If $|D_t|/(N-m) > u$, sampling without replacement gives

$$\begin{aligned}
\mathbb{P}(A \cap D_t = \varnothing) &= \frac{\binom{N-|D_t|}{m}}{\binom{N}{m}} \\
&\leq \left(1 - \frac{|D_t|}{N}\right)^m \\
&\leq \exp\left(-u\frac{(N-m)m}{N}\right) \leq e^{-um/2}.
\end{aligned}$$

A union bound yields

$$\mathbb{P}\left(\frac{|D_{\kappa(A(h))}|}{N-m} > u\right) \leq \min\{1, Me^{-um/2}\}.$$

Integrating over $u \in [0,1]$,

$$\mathbb{E}\frac{|D_{\kappa(A(h))}|}{N-m} \le \frac{2(\log M+1)}{m} \le C\frac{(k+1)\log(eN)}{m}.$$

□

C PROOF OF THE FINITE-POPULATION MW LEMMA

Proof of Lemma 5. Let $R_t = R \setminus \{Z_1, \dots, Z_{t-1}\}$. For a fixed benchmark $f^\star$, define

$$Q_t(x,y) = \mathbf{1}\{f^\star(x) = y,\ f_s(x) \neq y \text{ for all } s < t\}.$$

Conditional on the history before round t, Z_t is uniform in R_t, so

$$\mathbb{E}[r_t(f^\star) \mid \mathcal{F}_{t-1}] = \frac{1}{|R_t|}\sum_{z\in R_t} Q_t(z).$$

The final unseen set U is a uniform size-$(N-T)$ subset of R_t, conditional on the same history. Also $Q_{T+1}(z) \le Q_t(z)$ for every $z \in U$. Therefore

$$\mathbb{E}r_t(f^\star) \ge \mathbb{E}\frac{1}{|U|}\sum_{z\in U} Q_{T+1}(z).$$

Summing over t lower bounds $\mathbb{E}\sum_t r_t(f^\star)$ by T times the desired final miss rate.

The multiplicative-weights regret inequality at learning rate $1/2$ is

$$\sum_{t=1}^{T}\langle p_t, r_t\rangle \ge \frac{2}{3}\sum_{t=1}^{T} r_t(f^\star) - \frac{4}{3}\log|F|. \tag{10}$$

For completeness, let $W_t = \sum_{f\in F} w_t(f)$. For $0 \le u \le 1$ and $0 < \eta \le 1$,

$$e^{\eta u} \le 1 + \eta u + \eta^2 u^2 \le 1 + \eta(1+\eta)u.$$

Consequently,

$$\log\frac{W_{T+1}}{W_1} \le \eta(1+\eta)\sum_{t=1}^{T}\langle p_t, r_t\rangle, \qquad \log\frac{W_{T+1}}{W_1} \ge \eta\sum_{t=1}^{T} r_t(f^\star) - \log|F|.$$

Taking $\eta = 1/2$ gives (10).

It remains to bound the left side. Before Z_t is exposed, define

$$D_t = \{(x,y) \in R : f_t(x) = y,\ f_s(x) \neq y \text{ for all } s < t\}.$$

The sets $D_1, \dots, D_T$ are pairwise disjoint on every path. Conditional on the past and on f_t,

$$\mathbb{E}[r_t(f_t) \mid \mathcal{F}_{t-1}, f_t] = \frac{|D_t \cap R_t|}{|R_t|} \le \frac{|D_t|}{N-T+1}.$$

Since $f_t \sim p_t$ before Z_t is revealed,

$$\mathbb{E}\sum_{t=1}^{T}\langle p_t, r_t\rangle = \mathbb{E}\sum_{t=1}^{T} r_t(f_t) \le \frac{\mathbb{E}\sum_t |D_t|}{N-T+1} \le \frac{N}{N-T+1} \le 2.$$

Rearranging (10) gives

$$\mathbb{E}\sum_{t=1}^{T} r_t(f^\star) \le 3 + 2\log|F|.$$

Divide by T and use the first part of the proof. □

D PROOF OF RANDOM-SPLIT AGNOSTIC COMPRESSION

Proof of Lemma 6. Conditional on R, every output of the scheme belongs to

$$\mathcal{G}_R = \left\{ \rho(t) : t \in \bigcup_{j=0}^{k} R^j \right\}, \qquad |\mathcal{G}_R| \le (k+1)N^k.$$

For fixed g, Hoeffding's comparison theorem for sampling without replacement gives

$$\mathbb{P}\{|L_C^\ell(g) - L_R^\ell(g)| > s\} \le 2e^{-2|C|s^2}.$$

Because

$$L_C^\ell(g) - L_V^\ell(g) = \frac{N}{|V|}\left(L_C^\ell(g) - L_R^\ell(g)\right),$$

a union bound and tail integration imply

$$\mathbb{E}\sup_{g\in\mathcal{G}_R} |L_C^\ell(g) - L_V^\ell(g)| \le C\sqrt{\frac{\log(2|\mathcal{G}_R|)}{N}} \le C\sqrt{\frac{(k+1)\log(eN)}{N}}.$$

The same argument for the single fixed comparator h gives a smaller bound. Empirical optimality of the compression scheme yields $L_C^\ell(\widehat{h}) \le L_C^\ell(h)$. Hence

$$L_V^\ell(\widehat{h}) - L_V^\ell(h) \le [L_V^\ell - L_C^\ell](\widehat{h}) + [L_C^\ell - L_V^\ell](h).$$

Taking expectations proves the lemma. □

E DEFERRED DECISIONS AND THE COMPLETE UPPER-BOUND BOOKKEEPING

We verify the exact partition identity and the passage between remaining populations.

E.1 PARTITION IDENTITY

Ignore internal block orders, which are uniform under both procedures. Under the implemented procedure, a fixed disjoint tuple (A, B, C, I) has probability

$$\frac{1}{n}\binom{n-1}{q}^{-1}\binom{n-1-q}{q}^{-1}\binom{n-1-2q}{q}^{-1}.$$

Under deferred decisions, its probability is

$$\binom{n}{q}^{-1}\binom{n-q}{q}^{-1}\binom{n-2q}{q}^{-1}\frac{1}{n-3q}.$$

Both expressions equal

$$\frac{(q!)^3(n-3q-1)!}{n!}.$$

E.2 SUBDIVISION IDENTITIES

If $D \subseteq S \setminus A$ is fixed conditional on A, every later block and remainder is exchangeable within $S \setminus A$. In particular,

$$\mathbb{E}[L_V(\mathbf{1}_D) \mid A] = L_{S\setminus A}(\mathbf{1}_D).$$

After (A, B) are fixed, the same statement holds with $S \setminus (A \cup B)$. This justifies passing the bounds from Lemmas 4 and 5 to V.

E.3 INSIDE-MENU DECOMPOSITION

For any g, h and any example (x, y),

$$\begin{aligned}\mathbf{1}\{g(x) \ne y\} - \mathbf{1}\{h(x) \ne y\} \le \mathbf{1}\{y \in \mu(x), g(x) \ne y\} - \mathbf{1}\{y \in \mu(x), h(x) \ne y\} \\ + \mathbf{1}\{h(x) = y, y \notin \mu(x)\}.\end{aligned}$$

Averaging gives (9). If $h^\star(x) = y$ but $y \notin \mu(x)$, either $f_{h^\star,A}(x) \ne y$, or $f_{h^\star,A}(x) = y$ and this correct label is missed by the menu. These are exactly the events in (6) and (7).

E.4 SMALL POPULATIONS

For n below a universal constant, the upper bound follows after enlarging C because transductive excess error is at most one. For larger n, $q = \lfloor n/4 \rfloor$ satisfies all constant-fraction conditions used above. If $d_{\mathrm{N}} = 0$, no two hypotheses disagree at any point. The class has one pointwise behavior, which the learner can output with zero excess.

F COMPLETE LOWER-BOUND PROOFS

F.1 NATARAJAN LOWER BOUND

Let $d = \min\{d_{\mathrm{N}}, n\}$. Conditional on the random labels outside the hidden index, its label is uniform over the two witness labels. Predicting any third label has error one, so every randomized learner has conditional error at least $1/2$.

For block j, encode its labels as independent Rademacher variables $\sigma_{j,1}, \ldots, \sigma_{j,k_j}$. Then

$$\frac{k_j}{2} - \min\{B_j, k_j - B_j\} = \frac{1}{2}\left|\sum_{r=1}^{k_j} \sigma_{j,r}\right|.$$

The $p = 1$ Khinchine inequality gives

$$\mathbb{E}\left|\sum_{r=1}^{k_j} \sigma_{j,r}\right| \geq \sqrt{\frac{k_j}{2}}.$$

Since $k_j \geq \lfloor n/d \rfloor \geq n/(2d)$,

$$\mathbb{E}\mathcal{E}_S(A) \geq \frac{1}{2n}\sum_{j=1}^{d}\sqrt{\frac{k_j}{2}} \geq \frac{1}{4}\sqrt{\frac{d}{n}}.$$

This expectation is over a finite random choice of labelings. At least one labeling attains the expected lower bound.

F.2 PROJECTION OF A PSEUDO-CUBE

Let $F \subseteq Y^D$ be a pseudo-cube and retain a coordinate set $J \subseteq [D]$. Let $F|_J$ be the set of distinct projections. Fix $v \in F|_J$, choose a preimage $\widetilde{v} \in F$, and fix $j \in J$. The pseudo-cube property supplies $\widetilde{u} \in F$ differing from $\widetilde{v}$ only at j. Their projections remain distinct and differ only at j. Thus $F|_J$ is a $|J|$-dimensional pseudo-cube.

F.3 RANDOMIZED FIBER COUNTING

Fix the padded population described in Section 5. For a direction $i \geq 2$, let e be a coordinate fiber of F. All $v \in e$ induce exactly the same observed labeled population when the unique copy of x_i is hidden. The learner therefore uses one common distribution P_e over predicted labels. Distinct vertices in e have distinct labels at coordinate i, so

$$\sum_{v \in e} \mathbb{P}_{P_e}\{\widehat{y} \neq v_i\} = |e| - \sum_{v \in e} P_e(v_i) \geq |e| - 1.$$

Each vertex belongs to one fiber in each of the $d-1$ retained directions. Every fiber is nontrivial, so

$$\begin{aligned}\sum_{v \in F}\sum_{i=2}^{d} \mathbb{P}\{A \text{ errs at } x_i \mid v\} &\geq \sum_{e}(|e| - 1) \\ &\geq \frac{1}{2}\sum_{e}|e| \\ &= \frac{|F|(d-1)}{2}.\end{aligned}$$

Some v has at least $(d-1)/2$ expected errors. Its population is realizable, so the comparator loss is zero. Dividing by n proves the DS lower bound for randomized learners.