SELECT, DO NOT VOTE: OPTIMAL LIST PAC LEARNING AND EXACT TRANSDUCTION

Pahan Dewasurendra
Johns Hopkins University

ABSTRACT

List learning permits a classifier to return ℓ labels and counts a prediction as correct when the target appears anywhere in the list. Its realizable sample complexity is characterized qualitatively by the ℓ-DS dimension d, but every known quantitative bound carries list-size dependence beyond d. We remove it. For every class of finite ℓ-DS dimension d, a randomized learner uses

$$O\left(\frac{d+\log(1/\delta)}{\varepsilon}\right)$$

samples. The algorithm does not aggregate lists. It samples a logarithmic number of prefix one-inclusion predictors and uses a fresh validation block to select one. This turns a high-probability average-risk inequality into a single list predictor without the $\ell+1$ loss of Top-ℓ voting.

The rate is minimax sharp, uniformly in ℓ. The lower bound uses functions with at most d nondefault coordinates over $2\ell+1$ labels. We also solve fixed-design list transduction exactly. If μ_ℓ is maximum excess hyperedge density, optimal randomized risk is μ_ℓ/m, while optimal deterministic risk is $\lceil \mu_\ell \rceil/m$. A one-coordinate class separates them by the sharp factor $\ell+1$. This is the list-valued extension of the Hall-complexity characterization for ordinary multiclass transduction. The same excess calculation strengthens the known i.i.d. transductive lower bound by $\ell+1$. List length can still reduce d, sometimes dramatically, but it creates no additional polynomial price once d is fixed.

1 INTRODUCTION

In multiclass list learning, a predictor returns a set of at most ℓ labels at every input. The loss is zero when the true label belongs to that set. This model captures top-ℓ classification, recommendation menus, and label ambiguity. It also serves as a structural primitive in modern multiclass learning theory (Brukhim et al., 2022; Charikar & Pabbaraju, 2023; Hanneke et al., 2024).

The qualitative theory is now clean. Charikar and Pabbaraju proved that a class is realizably ℓ-list PAC learnable exactly when its ℓ-DS dimension is finite (Charikar & Pabbaraju, 2023). Recent algebraic advances established a sharp Sauer lemma and showed that the maximum excess density of every one-inclusion hypergraph is at most the ℓ-DS dimension (Hanneke et al., 2026b; Pabbaraju, 2026). These results remove the longstanding $d^{3/2}$ barrier in the dependence on dimension.

The dependence on list length remained unresolved. Write $d = d_{\mathrm{DS}}^{(\ell)}(\mathcal{H})$. The current realizable upper bound is

$$O\left(\frac{\ell(d+\log(1/\delta))}{\varepsilon}\right),$$

while the prior dimension lower bound is $\Omega(d/(\ell\varepsilon))$ (Hanneke et al., 2024). Combining it with the standard confidence lower bound gives $\Omega(d/(\ell\varepsilon)+\log(1/\delta)/\varepsilon)$. Thus the dimension terms differ by ℓ^2, and even the known upper confidence term pays ℓ. This is not a cosmetic gap. It asks whether returning a longer list creates a statistical cost beyond the change it induces in the dimension itself.

We show that it does not. The minimax realizable sample complexity among classes of ℓ-DS dimension at most d is

$$\Theta\left(\frac{d+\log(1/\delta)}{\varepsilon}\right),$$

with universal constants independent of ℓ and the number of labels. The upper bound holds class by class. The lower bound is witnessed by a fixed sparse-support class over $2\ell+1$ labels.

The upper-bound idea is elementary once the right intermediate statement is exposed. The one-inclusion list predictor has leave-one-out error at most d. The high-probability conversion of Aden-Ali et al. (2023), as instantiated by Pabbaraju (2026), shows that the average population risk of a suffix of prefix predictors is $O((d+\log(1/\delta))/n)$. The previous algorithm takes a Top-ℓ vote. If the aggregate list misses the target, only a $1/(\ell+1)$ fraction of the constituent lists need miss it, which introduces the factor $\ell+1$.

Our learner selects rather than votes. It samples $O(\log(1/\delta))$ prefix predictors. Markov's inequality ensures that one sampled candidate has risk within a constant factor of the average. A fresh validation block selects a good candidate using relative concentration. Validation costs $O(1/\varepsilon)$, not $O(1/\varepsilon^2)$, because the best candidate already has risk $O(\varepsilon)$. The selected object is still one ℓ-list predictor, so no list-size amplification occurs.

The lower bound explains why the resulting rate is the right one. Consider functions that equal a default label except on at most d coordinates, where one of 2ℓ active labels may appear. This class has ℓ-DS dimension exactly d. Under a sparse product prior, an unseen coordinate is active with probability proportional to ε, and conditional on activity its label is uniform among 2ℓ possibilities. Even a list of length ℓ misses with constant conditional probability. There are $\Theta(d/\varepsilon)$ possible coordinates, so fewer than a constant multiple of d/ε samples leave enough independent uncertainty to force error above ε.

We also revisit the transductive quantity that underlies one-inclusion learning. Existing lower bounds count every hyperedge of size larger than ℓ and then use the worst possible miss probability $1/(\ell+1)$. The exact conditional miss probability on an edge e is $(|e|-\ell)/|e|$. After averaging over targets, the factor $|e|$ cancels and leaves precisely the excess $(|e|-\ell)_+$. Linear-programming duality then gives more than a lower bound. It identifies maximum excess density as the exact randomized minimax load, while integral orientations give its ceiling. For $\ell=1$, this recovers the Hall-complexity characterization of Asilis et al. (2024). For general ℓ, it reveals a sharp factor-$\ell+1$ value of randomization and improves the i.i.d. lower bound by the same factor.

Two recent results require a careful distinction. The sharp Sauer lemma and excess-density theorem of Hanneke et al. (2026b); Pabbaraju (2026) supply the structural input to our upper bound. The subsequent ListCascade analysis of Hanneke et al. (2026a) learns a $(2L-1)$-list at a rate controlled by $d_{\mathrm{DS}}^{(\lceil L/2 \rceil)} \log K$, as an ingredient for bandit feedback over K labels. It neither returns an L-list controlled by $d_{\mathrm{DS}}^{(L)}$ nor removes label-range dependence. Our theorem gives the exact output size, the matching dimension, and no K factor. The transductive online model of Hanneke & Shaeiri (2025) is sequential and governed by different Littlestone-type dimensions.

Contributions. Our results are:

- A factor-free realizable upper bound $O((d+\log(1/\delta))/\varepsilon)$ for every class with ℓ-DS dimension d.
- A matching minimax lower bound over a label alphabet of size $2\ell+1$, including the confidence term with no dependence on ℓ.
- Exact fixed-design values: μ_ℓ/m for randomized learners and $\lceil \mu_\ell \rceil/m$ for deterministic learners, with a sharp factor-$\ell+1$ separation.
- A factor-$\ell+1$ strengthening of the prior i.i.d. lower bound and, through a finite-seed cover argument, removal of ℓ from the fast d/ε term in agnostic list learning up to logarithmic factors.

Result	Upper bound	Lower bound
Prior work	$O(\ell(d+\log(1/\delta))/\varepsilon)$	$\Omega(d/(\ell\varepsilon)+\log(1/\delta)/\varepsilon)$
This work	$O((d+\log(1/\delta))/\varepsilon)$	$\Omega((d+\log(1/\delta))/\varepsilon)$

Table 1: Realizable ℓ-list PAC sample complexity as a function of ℓ-DS dimension d. The new upper bound is classwise. Lower bounds are minimax over classes of dimension d.

2 Setup and main result

Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. An ℓ-list predictor is a map $g : \mathcal{X} \to \{A \subseteq \mathcal{Y} : |A| \leq \ell\}$. For a distribution P on $\mathcal{X} \times \mathcal{Y}$, define $\mathrm{err}_P(g) = \mathbb{P}_{(X,Y)\sim P}[Y \notin g(X)]$. A distribution is realizable by $\mathcal{H}$ if $Y = h^\star(X)$ almost surely for some $h^\star \in \mathcal{H}$. Randomized learners include their internal randomness in the probability of failure.

For a finite restriction $F \subseteq \mathcal{Y}^m$, two vectors are i-neighbors when they differ only in coordinate i. The set F is an ℓ-pseudo-cube if every vector has at least ℓ neighbors in every direction. A sequence of inputs is ℓ-DS shattered when the restriction of $\mathcal{H}$ contains a finite ℓ-pseudo-cube. The maximum shattered length is $d^{(\ell)}_{\mathrm{DS}}(\mathcal{H})$, abbreviated d below. This is the dimension that characterizes ℓ-list learnability (Charikar & Pabbaraju, 2023).

Theorem 1 (Optimal realizable list sample complexity)**.** *There are universal constants $c, C > 0$ with the following properties.*

1. *For every $\ell \geq 1$, every class $\mathcal{H}$ with $d^{(\ell)}_{\mathrm{DS}}(\mathcal{H}) = d < \infty$, and every $\varepsilon, \delta \in (0, 1/2)$, there is a randomized ℓ-list learner with*
$$m^{(\ell)}_{\mathcal{H}}(\varepsilon, \delta) \leq C\frac{d + \log(1/\delta)}{\varepsilon}.$$
2. *For every $\ell, d \geq 1$, there is a class $\mathcal{H}_{d,\ell} \subseteq [2\ell+1]^{\mathbb{N}}$ with $d^{(\ell)}_{\mathrm{DS}}(\mathcal{H}_{d,\ell}) = d$ such that, for $\varepsilon \in (0, 1/32)$ and $\delta \in (0, 1/8)$,*
$$m^{(\ell)}_{\mathcal{H}_{d,\ell}}(\varepsilon, \delta) \geq c\frac{d + \log(1/\delta)}{\varepsilon}.$$

The constants do not depend on ℓ, d, or the label alphabet.

The theorem does not say that larger lists are useless. Increasing ℓ can substantially decrease $d^{(\ell)}_{\mathrm{DS}}(\mathcal{H})$, even changing it from infinite to finite. The theorem says that after this change is captured by d, no further polynomial factor in ℓ is present.

3 Selection removes the voting penalty

We isolate the statistical step first. It applies to any data-dependent collection of predictors with bounded loss, not only to list learning.

Lemma 2 (Select from a low-risk average)**.** *Let $g_1, \ldots, g_N$ be predictors constructed from a training sample, and let $R_P(g_j) \in [0, 1]$ denote their population risks. Suppose*
$$\mathbb{P}\left[\frac{1}{N}\sum_{j=1}^{N} R_P(g_j) \leq r\right] \geq 1 - \eta.$$
Independently sample R indices uniformly with replacement from $[N]$. On a fresh validation sample of size v, return the sampled candidate of minimum empirical risk. For every $\zeta \in (0, 1)$, the returned predictor $\widehat{g}$ satisfies
$$R_P(\widehat{g}) \leq 16r + \frac{24}{v}\log\frac{2R}{\zeta}$$
with probability at least $1 - \eta - 4^{-R} - \zeta$.

The proof is in Appendix A. On the event that the average is at most r, at least three quarters of the candidates have risk at most $4r$. Thus one of R random candidates is good except with probability 4^{-R}. A relative validation inequality compares risks at scale r, which is why v scales as $1/r$, rather than $1/r^2$.

We now apply the lemma. The one-inclusion list predictor of Charikar & Pabbaraju (2023) is symmetric in its training sample. Recent orientation and density results imply that its leave-one-out error on every realizable sample is at most d (Hanneke et al., 2026b; Pabbaraju, 2026). The high-probability leave-one-out conversion of Aden-Ali et al. (2023) therefore gives the following statement. For a sample $S = (Z_1, \dots, Z_n)$, let g_t be the one-inclusion list predictor trained on the prefix $S_{\le t}$. When n is divisible by four,

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} \mathrm{err}_P(g_t) \le 4.82 \left(\frac{d}{n} + \frac{1}{n} \log \frac{2}{\eta} \right) \tag{1}$$

with probability at least $1 - \eta$. Equation (1) is exactly the intermediate inequality used by Pabbaraju (2026).

The prior proof forms the Top-ℓ vote of all g_t. If this aggregate misses y, only a $1/(\ell+1)$ fraction of the constituent lists must miss y. Multiplying Equation (1) by $\ell + 1$ gives the previous upper bound.

Instead set $\eta = \delta/4$, sample

$$R = \left\lceil \frac{\log(4/\delta)}{\log 4} \right\rceil$$

prefix indices from the suffix, and validate the corresponding list predictors on a fresh block of size n. Lemma 2, with $v = n$ and $\zeta = \delta/2$, gives

$$\mathrm{err}_P(\widehat{g}) \le \frac{77.12\big(d + \log(8/\delta)\big) + 24 \log(4R/\delta)}{n} \tag{2}$$

with probability at least $1-\delta$. Since $\log(4R/\delta) = O(\log(1/\delta))$, two blocks of size n prove the upper half of Theorem 1. The numerical constants are displayed only to make the reduction auditable. No effort is made to optimize them.

The argument also covers infinite label spaces. The compactness extension of the one-inclusion list orientation supplies the same leave-one-out bound in that setting (Charikar & Pabbaraju, 2023; Pabbaraju, 2026). Candidate sampling and validation involve only finitely many predictors and require no further restriction on $\mathcal{Y}$.

Why validation remains fast. Ordinary additive uniform convergence over R candidates would suggest $v = O(1/\varepsilon^2)$. That is unnecessary here. A candidate of risk $O(\varepsilon)$ already exists. Relative concentration separates such a candidate from every candidate of risk larger than a constant multiple of ε using $O(\log(R/\delta)/\varepsilon)$ labels.

4 Exact randomized and deterministic list transduction

The selection proof uses an average-risk guarantee generated by one-inclusion orientations. We next solve the corresponding fixed-design game exactly. This extends the Hall-complexity theorem of Asilis et al. (2024) from one predicted label to a list of ℓ labels.

For finite $F \subseteq \mathcal{Y}^m$, its one-inclusion hypergraph has one edge for each coordinate i and each fixed labeling of the other $m - 1$ coordinates. Define

$$\mathrm{Dens}_\ell(F) = \frac{1}{|F|} \sum_e (|e| - \ell)_+, \qquad \mu_\ell(F) = \max_{\varnothing \ne G \subseteq F} \mathrm{Dens}_\ell(G).$$

In fixed-design transduction, the learner observes f_{-i}, where $f \in F$ is the target and i is the hidden coordinate, and returns at most ℓ labels for coordinate i. Let $\rho_\ell^{\mathrm{rand}}(F)$ and $\rho_\ell^{\mathrm{det}}(F)$ be the minimax risks when the learner may randomize and when it must be deterministic, respectively.

Theorem 3 (Exact list transduction)**.** *For every nonempty finite $F \subseteq \mathcal{Y}^m$,*

$$\rho_\ell^{\mathrm{rand}}(F) = \frac{\mu_\ell(F)}{m}, \qquad \rho_\ell^{\mathrm{det}}(F) = \frac{\lceil \mu_\ell(F) \rceil}{m}.$$

Both equalities are attained by one-inclusion list orientations.

Proof. Fix nonempty $G \subseteq F$ and draw the target uniformly from G. On an edge e, every target produces the same observation, and a list covers at most ℓ of its $|e|$ hidden labels. Summing misses over targets in e gives at least $(|e| - \ell)_+$. Summing over edges and dividing by $m|G|$ gives average error at least $\mathrm{Dens}_\ell(G)/m$. Maximizing over G proves the lower bound $\mu_\ell(F)/m$, even for randomized learners.

For the randomized upper bound, let $z_{e,v}$ be the probability that vertex $v \in e$ is omitted from the list on edge e. The feasible vectors satisfy $0 \le z_{e,v} \le 1$ and $\sum_{v \in e} z_{e,v} = (|e| - \ell)_+$. The minimum possible maximum expected miss load is the linear program

$$\tau_\ell(F) = \min_z \max_{v \in F} \sum_{e \ni v} z_{e,v}.$$

Its dual has a weight $a_v \ge 0$ on each vertex with $\sum_v a_v = 1$. If $b_e = (|e| - \ell)_+$ and $a_{e,(1)} \le \cdots \le a_{e,(|e|)}$ are the incident weights, the dual objective is

$$\sum_e \sum_{j=1}^{b_e} a_{e,(j)}.$$

For $U_t = \{v : a_v \ge t\}$, the layer-cake identity turns this objective into

$$\int_0^\infty \sum_e (|e \cap U_t| - \ell)_+ \, dt \le \mu_\ell(F) \int_0^\infty |U_t| \, dt = \mu_\ell(F).$$

Uniform weights on a subset attaining $\mu_\ell(F)$ give equality. Thus $\tau_\ell(F) = \mu_\ell(F)$. The feasible polytope on each edge is the convex hull of integral omission sets, so independent edgewise randomization realizes the fractional solution. This proves the first equality.

For deterministic learners, every maximum miss load is an integer. The randomized lower bound therefore implies a deterministic load of at least $\lceil \mu_\ell(F) \rceil$. The list-orientation theorem used by Hanneke et al. (2026b); Pabbaraju (2026) supplies an orientation with maximum load at most this ceiling. Dividing by m proves the second equality. Appendix B gives the LP dual in full. □

The two values can differ by $\ell + 1$, and this factor is sharp. Take $m = 1$ and let F contain the $\ell + 1$ possible labels. Then $\mu_\ell(F) = 1/(\ell + 1)$. A randomized learner can omit each label with equal probability, while every deterministic list misses one fixed target. Hence

$$\rho_\ell^{\mathrm{rand}}(F) = \frac{1}{\ell + 1}, \qquad \rho_\ell^{\mathrm{det}}(F) = 1.$$

For $\ell = 1$, regularity of the one-inclusion hypergraph gives $\mu_1(F) = m - \mathrm{Hall}(F)$, so Theorem 3 recovers the randomized Hall-complexity formula of Asilis et al. (2024) and its deterministic rounding counterpart.

The same edge calculation strengthens the i.i.d. transductive lower bound of Charikar & Pabbaraju (2023). Define

$$\mu_{\ell,\mathcal{H}}(m) = \sup_{x_1,\ldots,x_m} \sup_{\varnothing \neq F \subseteq \mathcal{H}|_{(x_1,\ldots,x_m)}} \mathrm{Dens}_\ell(F),$$

where the outer supremum is over distinct inputs.

Corollary 4 (i.i.d. transductive lower bound)**.** *For any class $\mathcal{H}$, every learner trained on $m - 1$ i.i.d. realizable examples has worst-case expected test error at least*

$$\frac{\mu_{\ell,\mathcal{H}}(m)}{em}.$$

The unseen-coordinate event from Charikar & Pabbaraju (2023, Theorem 6) contributes $m^{-1}(1 - 1/m)^{m-1} \ge 1/(em)$. The exact edge miss replaces their factor $1/(\ell+1)$ by excess density. Details appear in Appendix C. Since $\lceil \mu_{\ell,\mathcal{H}}(m) \rceil \le d_{\mathrm{DS}}^{(\ell)}(\mathcal{H})$ (Pabbaraju, 2026), fixed-design and i.i.d. list transduction both have scale d/m in the worst case.

5 A MATCHING PAC LOWER BOUND

Fix $q = 2\ell$ and define the sparse-support class

$$\mathcal{H}_{d,\ell} = \{h : \mathbb{N} \to \{0, 1, \ldots, q\} : |\operatorname{supp}(h)| \le d\}, \qquad \operatorname{supp}(h) = \{x : h(x) \neq 0\}. \tag{3}$$

Proposition 5 (Dimension and density of the hard class)**.** *The class in Equation (3) has ℓ-DS dimension exactly d. On a restriction to $n \ge d$ coordinates, its excess density is*

$$\frac{n(q+1-\ell)\sum_{s=0}^{d-1}\binom{n-1}{s}q^s}{\sum_{s=0}^{d}\binom{n}{s}q^s},$$

which converges to $d(q+1-\ell)/q$ as $n \to \infty$. For $q = 2\ell$, the limit is at least $d/2$.

For the dimension lower bound, any d coordinates contain the full product over labels $\{0, 1, \ldots, \ell\}$. The upper bound has a direct proof. Suppose an ℓ-pseudo-cube exists on k coordinates, and choose one of its patterns with maximum support. If that pattern is zero at coordinate i, every i-neighbor changes zero to a nonzero label and has larger support, a contradiction. The maximal pattern is therefore nonzero on all k coordinates, so $k \le d$. The density formula follows by counting hyperedges. An edge has size $q+1$ exactly when the fixed labels on the other coordinates have support at most $d-1$.

We sketch the PAC lower bound. Let $\varepsilon \le 1/32$, take

$$n = \left\lfloor \frac{d}{16\varepsilon} \right\rfloor, \qquad p = 8\varepsilon,$$

and put the uniform input distribution on $[n]$. Generate a random target by choosing labels independently at each coordinate: label zero with probability $1-p$, and each active label in $[q]$ with probability p/q. The expected support is at most $d/2$. With probability at least $1 - \exp(-d/6)$, the support is at most d, so conditioning on this event produces a prior supported on $\mathcal{H}_{d,\ell}$.

Suppose the learner receives fewer than $n/4$ examples. At least $3n/4$ coordinates are unseen, deterministically. Before conditioning on the support event, labels at unseen coordinates remain independent. For any predicted list L of size ℓ,

$$\mathbb{P}[Y \notin L] \ge \frac{p}{2}.$$

Indeed, if $0 \in L$, the list contains at most $\ell - 1$ of the $q = 2\ell$ active labels. If $0 \notin L$, it already misses the default label with probability $1-p$. Thus the number of errors on unseen coordinates stochastically dominates a binomial variable with mean at least $3n\varepsilon$. A Chernoff bound shows that the population error exceeds ε except with probability at most $\exp(-d/26)$. Removing targets outside the class costs at most $\exp(-d/6)$. For $d \ge 48$, the conditioned failure probability is larger than a universal constant. For smaller d, the confidence construction below supplies the same d/ε lower bound after adjusting the universal constant. Appendix D gives the complete argument.

For the confidence term, place mass 2ε on one rare coordinate and the remaining mass on an anchor. Choose the rare label uniformly among $q = 2\ell$ active labels. If the rare coordinate is absent from the sample, every ℓ-list misses its label with probability at least $1/2$. Hence failure probability is at least $\frac{1}{2}(1-2\varepsilon)^m$, which yields $m = \Omega(\log(1/\delta)/\varepsilon)$. Taking the larger of the dimension and confidence lower bounds proves the second half of Theorem 1.

6 CONSEQUENCES AND LIMITATIONS

Agnostic list learning. The agnostic reduction of Pabbaraju (2026) uses a constant-error realizable learner to construct a finite list cover, then applies multiplicative weights and inside-menu compression. Our learner has a finite seed space because its only random choices are prefix indices. Including those indices as side information extends the minimax cover argument without changing its polynomial terms. Appendix E proves this finite-seed conversion. Substitution gives

$$\widetilde{O}\left(\frac{d_{\mathrm{DS}}^{(\ell)}}{\varepsilon} + \frac{\ell^4 d_{\mathrm{Nat}}^{(\ell)} + \log(1/\delta)}{\varepsilon^2}\right).$$

The suppressed factors gain only logarithms in the realizable sample size and confidence.

What remains open. The agnostic square-root term still carries polynomial dependence on ℓ. It is also unknown whether $d^{(\ell)}_{\mathrm{Nat}}/\varepsilon^2$ is necessary when the comparator is the best single-label hypothesis, rather than the best list predictor. Our result settles the realizable term and the worst-case realizable sample complexity. It does not resolve these agnostic questions.

Randomization. The PAC learner samples prefix indices during training, then returns one fixed list predictor. It does not randomize at test time. Deterministic selection over all prefixes is possible, but the direct analysis introduces an avoidable $\log n$ factor. In contrast, randomization is intrinsic to the exact value of the fixed-design game. The one-coordinate example shows that forbidding prediction-time randomization can increase minimax transductive risk by $\ell + 1$.

7 Conclusion

The apparent list-size penalty in realizable PAC learning comes from aggregation, not from information. High-probability suffix averaging already produces many good one-inclusion list predictors at the factor-free scale. Sampling and validating one preserves list length and removes the $\ell + 1$ voting loss. A sparse-support construction shows that the resulting d/ε rate is unavoidable. In fixed-design transduction, excess density is the exact fractional miss load and its ceiling is the exact integral load, exposing a separate sharp $\ell + 1$ value of randomization. List length influences learning through the ℓ-DS dimension, but once that dimension is fixed, there is no additional polynomial price.

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. Optimal PAC bounds without uniform convergence. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science*, pp. 1203–1223, 2023. doi: 10.1109/FOCS57990.2023.00071.

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. doi: 10.1109/FOCS54457.2022.00093.

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.

Steve Hanneke and Amirreza Shaeiri. A trichotomy for list transductive online learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22009–22022, 2025.

Steve Hanneke, Shay Moran, and Qian Zhang. Improved sample complexity for multiclass PAC learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 42798–42839, 2024. doi: 10.52202/079017-1356.

Steve Hanneke, Qinglin Meng, Shay Moran, and Amirreza Shaeiri. PAC learning with bandit feedback: Sharp sample complexity in the realizable setting. *arXiv preprint arXiv:2605.25678*, 2026a.

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

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

A Proof of the selection lemma

We first record a relative validation inequality.

Lemma 6 (Finite relative validation)**.** *Let $g_1, \dots, g_R$ be fixed predictors with losses in $[0,1]$, independent of a validation sample of size v. Let p_j and $\widehat{p}_j$ be their population and empirical risks. With probability at least $1-\zeta$, simultaneously for all $j \in [R]$,*

$$p_j \le 2\widehat{p}_j + b, \qquad \widehat{p}_j \le 2p_j + b, \qquad b = \frac{8}{v}\log\frac{2R}{\zeta}.$$

Consequently, if $\widehat{j} \in \arg\min_j \widehat{p}_j$, then $p_{\widehat{j}} \le 4\min_j p_j + 3b$.

Proof. Fix j, abbreviate $p = p_j$, $\widehat{p} = \widehat{p}_j$, and let $L = \log(2R/\zeta)$, so $b = 8L/v$. If $p \le b$, the event $p > 2\widehat{p} + b$ is impossible. If $p > b$, that event implies $\widehat{p} < p/2$, whose probability is at most $\exp(-vp/8) \le \exp(-L)$ by the multiplicative Chernoff bound. The event $\widehat{p} > 2p + b$ implies $\widehat{p} - p > p + b$. Bernstein's inequality bounds its probability by

$$\exp\left(-\frac{v(p+b)^2}{2p + 2(p+b)/3}\right) \le \exp\left(-\frac{3v(p+b)}{8}\right) \le \exp(-3L).$$

A union bound over the two events and R predictors proves the simultaneous statement. On this event, for $j^\star \in \arg\min_j p_j$,

$$p_{\widehat{j}} \le 2\widehat{p}_{\widehat{j}} + b \le 2\widehat{p}_{j^\star} + b \le 4p_{j^\star} + 3b.$$

□

Proof of Lemma 2. Condition on the event $N^{-1}\sum_j R_P(g_j) \le r$. Markov's inequality shows that at most one quarter of the candidates have risk greater than $4r$. Therefore, the probability that all R sampled candidates have risk greater than $4r$ is at most 4^{-R}. Conditional on the training sample and sampled indices, Lemma 6 applies to the independent validation block. Whenever a candidate of risk at most $4r$ was sampled and the validation event holds,

$$R_P(\widehat{g}) \le 4(4r) + 3\frac{8}{v}\log\frac{2R}{\zeta} = 16r + \frac{24}{v}\log\frac{2R}{\zeta}.$$

The three failure probabilities are η, 4^{-R}, and ζ. □

B Fractional list orientations

We prove the linear-programming identity used in Theorem 3. Write $V = F$, let E be its one-inclusion hyperedges, and set $b_e = (|e| - \ell)_+$. A fractional omission orientation is a collection $z_{e,v} \in [0,1]$ such that

$$\sum_{v \in e} z_{e,v} = b_e \qquad \text{for every } e \in E.$$

Its load at v is $L_v(z) = \sum_{e \ni v} z_{e,v}$. Since $\max_v L_v(z) = \max_{a \in \Delta(V)} \sum_v a_v L_v(z)$, the finite minimax theorem gives

$$\begin{aligned}\min_z \max_v L_v(z) &= \max_{a \in \Delta(V)} \min_z \sum_{e \in E}\sum_{v \in e} a_v z_{e,v} \\ &= \max_{a \in \Delta(V)} \sum_{e \in E}\sum_{j=1}^{b_e} a_{e,(j)}, \end{aligned} \tag{4}$$

where $a_{e,(1)} \le \cdots \le a_{e,(|e|)}$ are the weights incident to e. The second equality follows because the minimizing edgewise allocation puts its b_e units on the smallest incident weights.

Fix $a \in \Delta(V)$ and let $U_t = \{v : a_v \ge t\}$. The layer-cake formula gives

$$\sum_{j=1}^{b_e} a_{e,(j)} = \int_0^\infty (|e \cap U_t| - \ell)_+ \, dt.$$

Indeed, if $r = |e \cap U_t|$, then exactly $(r - \ell)_+$ of the $b_e = |e| - \ell$ smallest weights exceed level t. Intersecting every edge of F with U_t produces the one-inclusion hypergraph of the restricted class U_t. Therefore,

$$\sum_e \sum_{j=1}^{b_e} a_{e,(j)} = \int_0^\infty |U_t| \operatorname{Dens}_\ell(U_t)\, dt$$
$$\leq \mu_\ell(F) \int_0^\infty |U_t|\, dt = \mu_\ell(F) \sum_v a_v = \mu_\ell(F).$$

Conversely, let $G \subseteq F$ attain maximum excess density and set $a_v = 1/|G|$ on G, with zero weight elsewhere. On an edge e, the sum of the b_e smallest weights is exactly $(|e \cap G| - \ell)_+/|G|$. The objective in Equation (4) is then $\operatorname{Dens}_\ell(G) = \mu_\ell(F)$. Hence the fractional optimum equals $\mu_\ell(F)$.

For each edge, the polytope $\{z \in [0,1]^e : \sum_v z_v = b_e\}$ is the convex hull of the indicator vectors of its size-b_e subsets. A distribution over integral omission sets therefore realizes every feasible fractional vector. Randomizing independently on each edge realizes the optimal fractional orientation as a randomized list predictor.

C THE I.I.D. EXCESS-DENSITY LOWER BOUND

Proof of Corollary 4. Fix distinct inputs $x_1, \dots, x_m$ and a finite nonempty $F \subseteq \mathcal{H}|_{(x_1,\dots,x_m)}$. Let D be uniform on these m inputs and draw the target labeling uniformly from F. For an edge e in direction i, consider the event that the test input is x_i, none of the $m-1$ training inputs equals x_i, and the target belongs to e. This event has probability

$$\frac{1}{m}\left(1-\frac{1}{m}\right)^{m-1}\frac{|e|}{|F|}.$$

Conditional on the event, every target in e produces the same observed training labels. Any output list covers at most ℓ of the $|e|$ possible hidden labels, so its conditional average error is at least $(|e|-\ell)_+/|e|$. Summing over edges gives expected error at least

$$\frac{1}{m}\left(1-\frac{1}{m}\right)^{m-1}\frac{1}{|F|}\sum_e (|e|-\ell)_+ \geq \frac{\mathrm{Dens}_\ell(F)}{em}.$$

The probabilistic method supplies a fixed target in F, and maximizing over the restriction and subfamily proves the claim. □

D COMPLETE PAC LOWER BOUND

D.1 PROOF OF PROPOSITION 5

Restrict $\mathcal{H}_{d,\ell}$ to $n \geq d$ coordinates. Any chosen d coordinates support the full product $\{0,1,\dots,\ell\}^d$, so the ℓ-DS dimension is at least d. Conversely, suppose a finite ℓ-pseudo-cube is supported on k coordinates. Choose a pattern in it with maximum number of nonzero entries. If its entry at coordinate i were zero, any i-neighbor would change that entry to a nonzero label while fixing all others, contradicting maximality. The pattern is nonzero on every one of the k coordinates. Since every pattern is realized by a function with support at most d, we have $k \leq d$.

There are $\sum_{s=0}^{d}\binom{n}{s}q^s$ restricted patterns. Fix a coordinate. An edge has size $q+1$ exactly when the labels on the other $n-1$ coordinates have support at most $d-1$. There are $\sum_{s=0}^{d-1}\binom{n-1}{s}q^s$ such edges. Every remaining edge is a singleton and has zero excess. Multiplying the number of nontrivial edges by $n(q+1-\ell)$ and dividing by the class size gives the displayed density formula. Finally,

$$\frac{n\binom{n-1}{d-1}q^{d-1}}{\binom{n}{d}q^d} = \frac{d}{q},$$

and the degree-d terms dominate the two finite sums as $n \to \infty$. This proves the stated limit.

D.2 THE DIMENSION TERM

Fix $d \geq 48$, $\varepsilon \leq 1/32$, $q = 2\ell$, $n = \lfloor d/(16\varepsilon)\rfloor$, and $p = 8\varepsilon$. Let P_X be uniform on $[n]$. Under an auxiliary product prior, independently set $H_i = 0$ with probability $1-p$, and set $H_i = j$ with probability p/q for each $j \in [q]$. Let $E = \{|\operatorname{supp}(H)| \leq d\}$.

The support size is binomial with mean $np \leq d/2$. The binomial upper tail is monotone in its mean, and the multiplicative Chernoff bound at mean $d/2$ gives

$$\mathbb{P}(E^c) \leq (e/4)^{d/2} \leq e^{-d/6}. \tag{5}$$

Consider any randomized learner receiving $m < n/4$ examples. Condition on the sampled input coordinates, their labels, and the learner's random seed. At least $n - m > 3n/4$ coordinates remain unseen. Their labels are independent under the auxiliary prior. For the predicted list L_i at an unseen coordinate,

$$\mathbb{P}[H_i \notin L_i] \geq p/2.$$

If $0 \in L_i$, then at most $\ell - 1$ active labels are covered and the miss probability is at least $p(1-(\ell-1)/q) \geq p/2$. If $0 \notin L_i$, the miss probability is at least $1-p \geq 3/4 \geq p/2$.

Let Z be the number of misclassified unseen coordinates. It stochastically dominates a binomial random variable with mean

$$\nu > \frac{3n}{4}\frac{p}{2} = 3n\varepsilon.$$

The population error exceeds ε whenever $Z > n\varepsilon$. The multiplicative Chernoff bound therefore gives

$$\mathbb{P}[Z \le n\varepsilon] \le \exp(-2\nu/9) \le \exp(-2n\varepsilon/3) \le \exp(-d/26),$$

where the final inequality uses $n\varepsilon \ge d/16 - \varepsilon$, $d \ge 48$, and $\varepsilon \le 1/32$.

Under the auxiliary prior, every learner thus has failure probability at least $1 - e^{-d/26}$. Conditioning the target prior on E, which makes it supported on $\mathcal{H}_{d,\ell}$, reduces this probability by at most $e^{-d/6}$. Hence

$$\mathbb{P}[\mathrm{err}_{P_X}(\widehat{g}) > \varepsilon \mid E] \ge 1 - e^{-d/26} - e^{-d/6} > \frac{1}{4}.$$

An averaging argument supplies a fixed target $h \in \mathcal{H}_{d,\ell}$ for which failure over the training sample and learner randomness is larger than $1/4$. Since $m < n/4$, this gives $m = \Omega(d/\varepsilon)$.

For $1 \le d < 48$, the confidence construction below with a fixed confidence, say $\delta = 1/8$, gives $\Omega(1/\varepsilon) = \Omega(d/\varepsilon)$ after decreasing the universal constant by a factor of 48.

D.3 The confidence term

Use two inputs x_0, x_1, with probabilities $1 - 2\varepsilon$ and 2ε. For each $j \in [q]$, let the target h_j satisfy $h_j(x_0) = 0$, $h_j(x_1) = j$, and equal zero elsewhere. These targets belong to $\mathcal{H}_{d,\ell}$ for every $d \ge 1$.

Draw J uniformly from $[q]$. On the event that the sample contains only x_0, the learner has no information about J. Its list at x_1 contains at most $\ell = q/2$ possible values, so it misses J with probability at least $1/2$. A miss gives population error $2\varepsilon > \varepsilon$. Averaging over J,

$$\mathbb{P}[\mathrm{err}(\widehat{g}) > \varepsilon] \ge \frac{1}{2}(1 - 2\varepsilon)^m.$$

For $\varepsilon \le 1/4$, $\log(1 - 2\varepsilon) \ge -4\varepsilon$. Thus the failure probability exceeds δ whenever

$$m < \frac{1}{4\varepsilon}\log\frac{1}{2\delta}.$$

Combining this with the dimension term and using $\max\{a, b\} \ge (a + b)/2$ proves Theorem 1.

E Agnostic substitution

The proof of the agnostic list-learning bound in Pabbaraju (2026, Appendix A.2) has three stages. Its first stage invokes a deterministic constant-error realizable learner inside a minimax argument, repeats it logarithmically many times, and obtains a finite list cover. We show explicitly that the finite randomness of our learner fits the same argument.

Fix constant accuracy and confidence, say $\varepsilon_0 = \delta_0 = 1/6$. Theorem 1 uses $k = O(d + 1)$ examples. Fix a deterministic choice of one-inclusion orientation and deterministic validation tie breaking. The remaining random seed consists of $R_0 = O(1)$ prefix indices, so its seed space Σ_k has size at most k^{R_0}. Write $A_s(T)$ for the deterministic list predictor produced from a length-k sample T and seed $s \in \Sigma_k$. For every realizable distribution P,

$$\mathbb{P}_{T\sim P^k,\, s}\left[\mathrm{err}_P(A_s(T)) \le \frac{1}{6}\right] \ge \frac{5}{6},$$

and therefore

$$\mathbb{E}_{T,s}\,\mathrm{err}_P(A_s(T)) \le \frac{11}{36} < \frac{1}{3}. \tag{6}$$

Now fix a finite realizable sample S of size m. Consider the zero-sum game whose columns are examples $z \in S$, whose rows are pairs $(T, s) \in S^k \times \Sigma_k$, and whose payoff is the list loss of $A_s(T)$ on z. For every distribution P over columns, Equation (6) implies that some row has expected payoff

below $1/3$. The minimax theorem therefore supplies a distribution Q over rows such that every fixed example is missed with probability below $1/3$ under Q.

Draw $j = \lceil \log_3(2m) \rceil$ rows independently from Q and take the union of their output lists. A fixed example is missed by all rows with probability below 3^{-j}. A union bound over S shows that with positive probability every example is covered. Hence there exist j rows whose union, of size at most $j\ell$, covers S.

This cover has a finite description. It stores $kj = O(d \log m)$ sample entries and j seeds. The number of possible descriptions is at most

$$m^{kj} |\Sigma_k|^j \leq m^{kj} k^{R_0 j}.$$

Thus the seed contributes only $O(\log k \log m)$ bits, dominated by the existing logarithmic cover terms. The multiplicative-weights and inside-menu compression stages of Pabbaraju (2026) are unchanged. Their previous $k = O(\ell d)$ is replaced by $k = O(d)$, which yields

$$\widetilde{O}\left(\frac{d_{\mathrm{DS}}^{(\ell)}}{\varepsilon} + \frac{\ell^4 d_{\mathrm{Nat}}^{(\ell)} + \log(1/\delta)}{\varepsilon^2} \right).$$