Local Regularization Does Not Characterize Multiclass PAC Learnability

Eric Hou

July 24, 2026

Abstract

Local regularization assigns each hypothesis a test-point-dependent score and predicts with a minimum-score hypothesis consistent with the sample. Asilis et al. asked whether this principle characterizes multiclass PAC learnability. We give a negative answer. There is a countable class of Daniely–Shalev-Shwartz dimension at most two with realizable PAC sample complexity

$$O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\right),$$

that no local regularizer learns. Hypotheses are edges of complete graphs and instances are tournaments. At a test tournament, the scores fix an edge ranking while the training sample independently removes competitors. Cyclic triangles force enough inversions that surviving competitors produce constant population error at arbitrarily large sample sizes.

1 Introduction

A central question in statistical learning theory is which simple algorithmic principles characterize learnability. In binary classification, every consistent empirical risk minimizer learns any PAC-learnable class with nearly optimal sample complexity. The multiclass setting is less rigid. Proper learning can fail even for learnable classes, and general optimal learners use sample-dependent one-inclusion structures rather than a fixed ordering of hypotheses [3, 4].

Asilis et al. proposed local regularization as a possible replacement for empirical risk minimization [1, 2]. A local regularizer is a fixed score $\psi(h, x)$ depending on a hypothesis h and a test point x. Given a labeled sample, an induced learner may choose, separately at each test point, the prediction of any minimum-score hypothesis in the realizable version space. The resulting predictor may be improper, but the scores are fixed before the sample is observed. Asilis et al. asked whether every PAC-learnable multiclass class is learned by some local regularizer.

Jafar, Asilis, and Dughmi gave a negative answer in the transductive model [5]. Their construction forces a mistake at a withheld point through a preference cycle. A PAC lower bound needs more. The withheld point may carry negligible population mass, and repeated independent samples may reveal a distinguished hidden point.

Our counterexample uses distributions over tournaments. At scale q, the hypotheses are the edges of K_q, and an instance is a tournament on its vertices. An edge hypothesis returns the head of that edge. For a target incidence (e, v), the hard distribution is uniform over tournaments that orient e toward v, so every observed label is v. The sample filters the competing edges incident to v; an independent tournament then tests the pointwise ranking fixed by the regularizer. Error is spread across many instances rather than concentrated at one distinguished point.

Theorem 1.1 (Main theorem)**.** *There exists a countable multiclass hypothesis class $\mathcal{H}$ satisfying*

$$m_{\mathcal{H}}(\varepsilon,\delta) \le \left\lceil \frac{\log(2/\delta)}{\varepsilon} \right\rceil \qquad \text{for all } \varepsilon,\delta \in (0,1),$$

and $\mathrm{DSdim}(\mathcal{H}) \le 2$*, such that no local regularizer learns $\mathcal{H}$.*

More quantitatively, for every local regularizer ψ there is a deterministic learner A_ψ induced by ψ such that, for

$$\varepsilon_0 = \frac{1}{96e}, \qquad \delta_0 = \frac{1}{192e},$$

there are arbitrarily large sample sizes n and $\mathcal{H}$-realizable distributions D for which

$$\mathbb{P}_{S\sim D^n}\left[L_D(A_\psi(S)) > \varepsilon_0\right] > \delta_0.$$

An explicit learner identifies the target after seeing both endpoint labels, while the two-label image structure gives the DS-dimension bound. For the lower bound, fixed tie-breaking turns the local scores into a strict edge order at each tournament. The hard sample independently removes competitors, and cyclic triangles leave enough inversions to give constant expected risk for some target incidence.

2 Preliminaries

Let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ a label space, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. For a finite labeled sample

$$S = ((x_1,y_1),\dots,(x_n,y_n)),$$

define its realizable version space by

$$V_{\mathcal{H}}(S) = \{h \in \mathcal{H} : h(x_i) = y_i \text{ for every } i \in [n]\}.$$

The sample is realizable when $V_{\mathcal{H}}(S) \neq \varnothing$. For a distribution D on $\mathcal{X}\times\mathcal{Y}$ and a predictor $g : \mathcal{X} \to \mathcal{Y}$, let

$$L_D(g) = \mathbb{P}_{(x,y)\sim D}[g(x) \neq y].$$

The distribution D is $\mathcal{H}$-realizable if some $h^\star \in \mathcal{H}$ has zero D-risk.

A learner maps every finite sample to a predictor. It PAC learns $\mathcal{H}$ if there is a function $m_A : (0,1)^2 \to \mathbb{N}$ such that, for every $\mathcal{H}$-realizable distribution D, every $\varepsilon,\delta \in (0,1)$, and every $n \ge m_A(\varepsilon,\delta)$,

$$\mathbb{P}_{S\sim D^n}[L_D(A(S)) > \varepsilon] \le \delta.$$

The unrestricted sample complexity $m_{\mathcal{H}}$ is the pointwise infimum of m_A over all learners A.

Definition 2.1 (Local regularization)**.** A local regularizer is a function

$$\psi : \mathcal{H}\times\mathcal{X} \to \mathbb{R}_{\ge 0}.$$

A learner A is induced by ψ if, for every nonempty realizable sample S and every $x \in \mathcal{X}$,

$$A(S)(x) \in \left\{ h(x) : h \in \operatorname*{arg\,min}_{g\in V_{\mathcal{H}}(S)} \psi(g,x) \right\}.$$

The behavior of A on empty or unrealizable samples is arbitrary. We say that ψ learns $\mathcal{H}$ if every learner induced by ψ PAC learns $\mathcal{H}$.

Remark 2.2 (Nonvacuity and attainment). The induction condition is restricted to nonempty realizable samples because $V_{\mathcal{H}}(S)$ is empty on an unrealizable sample. There is a second issue on infinite version spaces: the score infimum need not be attained, in which case the argmin is empty and the universal statement over induced learners becomes vacuous.

For the class constructed below, every nonempty realizable version space is finite. Each local regularizer therefore attains its minimum wherever the induction condition is used, so no separate well-definedness assumption is required.

3 The tournament class

For an integer $q \geq 3$, let

$$V_q = [q], \qquad E_q = \binom{V_q}{2}, \qquad M_q = |E_q| = \binom{q}{2}.$$

Let $\mathcal{T}_q$ be the set of all tournaments on V_q. For $\tau \in \mathcal{T}_q$ and $e = \{u, v\} \in E_q$, write $\mathrm{head}_\tau(e) \in e$ for the head of the directed edge e in τ.

Define the countable instance and label spaces

$$\mathcal{X} = \bigsqcup_{q \geq 3} (\{q\} \times \mathcal{T}_q), \qquad \mathcal{Y} = \bigsqcup_{q \geq 3} (\{q\} \times V_q).$$

For every $q \geq 3$ and edge $e = \{u, v\} \in E_q$, define $h_{q,e} : \mathcal{X} \to \mathcal{Y}$ by

$$h_{q,e}(s, \tau) = \begin{cases} (q, \mathrm{head}_\tau(e)), & s = q, \\ (q, \min e), & s \neq q. \end{cases} \tag{1}$$

Here, in the second line, the tournament component belongs to $\mathcal{T}_s$ and is ignored. The constant off-block value makes $h_{q,e}$ a total function without mixing blocks; since every output retains the first coordinate q, one observed label already identifies the block of any consistent hypothesis. Let

$$\mathcal{H} = \{h_{q,e} : q \geq 3,\ e \in E_q\}.$$

This class is countable. Each hypothesis uses exactly two labels:

$$\mathrm{im}(h_{q,\{u,v\}}) = \{(q, u), (q, v)\}. \tag{2}$$

Moreover, distinct hypotheses have distinct image sets.

3.1 An explicit PAC learner

Proposition 3.1. *For every $\varepsilon, \delta \in (0, 1)$,*

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(2/\delta)}{\varepsilon} \right\rceil.$$

Proof. On a nonempty realizable sample, let $A^\star$ output $h_{q,\{u,v\}}$ if two distinct labels (q, u) and (q, v) appear, and the constant predictor $x \mapsto y$ if only one label y appears. Its behavior on empty or unrealizable samples is arbitrary.

Fix a realizable distribution with target $h_{q,\{u,v\}}$, and let α be the probability of label (q, u). Once both labels appear, Equation (2) identifies the target and the risk is zero. If only (q, u) appears, the

event has probability α^n and the risk is $1-\alpha$; if only (q, v) appears, the corresponding quantities are $(1-\alpha)^n$ and α. Hence

$$\begin{aligned}\mathbb{P}\big[L_D(A^\star(S)) > \varepsilon\big] &\le \mathbf{1}\{1-\alpha > \varepsilon\}\alpha^n + \mathbf{1}\{\alpha > \varepsilon\}(1-\alpha)^n \\ &\le 2e^{-\varepsilon n}.\end{aligned}$$

This is at most δ at the threshold stated in the proposition. □

3.2 A dimension bound

We use the standard Daniely–Shalev-Shwartz dimension. A finite nonempty set $P \subseteq \mathcal{Y}^d$ is a d-dimensional pseudo-cube if every $z \in P$ has, for every coordinate $i \in [d]$, a point $z' \in P$ that differs from z at coordinate i and agrees at all other coordinates. A d-tuple of instances is DS-shattered when the corresponding restriction of the class contains a d-dimensional pseudo-cube.

The image condition rules out a three-dimensional pseudo-cube: the required coordinate neighbors would force two distinct hypotheses to have the same two-label image.

Lemma 3.2. *Suppose every hypothesis in a class $\mathcal{G}$ uses at most two labels and the map $g \mapsto \mathrm{im}(g)$ is injective. Then* $\mathrm{DSdim}(\mathcal{G}) \le 2$.

Proof. It suffices to exclude a three-dimensional pseudo-cube, since projection onto any subset of coordinates preserves the pseudo-cube property. Suppose that $P \subseteq \mathcal{Y}^3$ is such a pseudo-cube in a restriction of $\mathcal{G}$. Every vector in P uses at most two labels. There is a vector using exactly two: if a chosen vector is constant, any coordinate-neighbor differs in one coordinate and therefore uses two labels. After permuting coordinates, write this vector as (a, a, b) with $a \neq b$.

Its neighbor in the first coordinate has the form (c, a, b) with $c \neq a$. It already contains a and b and uses at most two labels, so $c = b$. Let $g, g' \in \mathcal{G}$ realize (a, a, b) and (b, a, b). These hypotheses are distinct, but their global image sets both contain a, b and have size at most two. Thus

$$\mathrm{im}(g) = \mathrm{im}(g') = \{a, b\},$$

contrary to injectivity. □

Corollary 3.3. *The tournament class satisfies* $\mathrm{DSdim}(\mathcal{H}) \le 2$.

Proof. Apply Lemma 3.2 using Equation (2). □

4 No local regularizer learns the class

Because $\mathcal{H}$ is countable, fix an enumeration

$$\mathcal{H} = \{g_1, g_2, \ldots\}.$$

Fix an arbitrary local regularizer ψ. At each test point x, refine its scores to a strict total order by declaring

$$g_i \prec_x g_j \quad \Longleftrightarrow \quad (\psi(g_i, x), i) <_{\mathrm{lex}} (\psi(g_j, x), j). \tag{3}$$

For a nonempty realizable sample S, the first coordinate of any observed label fixes one block, so $V_{\mathcal{H}}(S)$ is finite. It therefore has a unique $\prec_x$-least element. Let $A_\psi(S)(x)$ be the label of this element, and define A_ψ arbitrarily on empty or unrealizable samples. The chosen hypothesis minimizes $\psi(\cdot, x)$ on the version space, hence A_ψ is induced by ψ.

A regularizer learns $\mathcal{H}$ only if every induced learner is PAC. It is therefore enough to show that A_ψ is not PAC.

4.1 Hard constant-label distributions

Fix $q \geq 3$, an edge $e \in E_q$, and an endpoint $v \in e$. Define $D_{e,v}$ to be the uniform distribution on

$$\left\{ \big((q,\tau),(q,v)\big) : \tau \in \mathcal{T}_q,\ \mathrm{head}_\tau(e) = v \right\}. \tag{4}$$

The distribution is realizable by $h_{q,e}$, and all labels are equal to (q, v).

Lemma 4.1 (Version-space structure)**.** *Let $S \sim D_{e,v}^n$ with $n \geq 1$. A hypothesis is consistent with S if and only if it is $h_{q,f}$ for an edge f incident to v and every sampled tournament orients f toward v. In particular, $h_{q,e}$ is always consistent. For distinct competitors $f \neq e$ incident to v, the consistency events are mutually independent, and each has probability*

$$p_n = 2^{-n}.$$

Proof. The label (q, v) forces a consistent hypothesis to lie in block q and to correspond to an edge incident to v. For such an edge f, Equation (1) gives the label (q, v) on a sampled tournament exactly when f is oriented toward v. This condition holds automatically for $f = e$ by Equation (4).

The unordered edges of a uniform tournament have independent unbiased orientations. Conditioning on the orientation of e leaves all other orientations independent and unbiased, and the sampled tournaments are independent. Each non-target competitor therefore survives with probability 2^{-n}, with mutual independence across competitors. □

4.2 Risk as a weighted inversion count

Write

$$d = q - 1.$$

For every incidence (e, v) and every test tournament $\tau \in \mathcal{T}_q$, define

$$\rho_{e,v}^{(n)}(\tau) = \mathbb{P}_{S \sim D_{e,v}^n}\big[A_\psi(S)(q,\tau) \neq (q,v)\big]. \tag{5}$$

We use this definition for every τ, although the formula below concerns the case in which e is incoming at v.

Fix a test tournament $\tau \in \mathcal{T}_q$ and a vertex $v \in V_q$. Restrict the total order in Equation (3) at (q, τ) to the d edges incident to v, and list them as

$$e_1^v \prec_{(q,\tau)} e_2^v \prec_{(q,\tau)} \cdots \prec_{(q,\tau)} e_d^v,$$

where we identify an edge with its corresponding block-q hypothesis. Define

$$b_j^v(\tau) = \begin{cases} 0, & \mathrm{head}_\tau(e_j^v) = v, \\ 1, & \mathrm{head}_\tau(e_j^v) \neq v. \end{cases}$$

Thus $b_j^v = 0$ denotes an incoming edge at v, and $b_j^v = 1$ an outgoing edge.

When the target is e_r^v, write

$$\rho_{v,r}^{(n)}(\tau) = \rho_{e_r^v,v}^{(n)}(\tau).$$

For a small example, suppose four incident edges are ordered as outgoing, incoming, outgoing, incoming, with the last edge as the target. The target is always present. If the first edge survives, it is selected and the prediction is wrong; if it is removed but the second edge survives, the prediction is correct; if the first two are removed and the third survives, the prediction is wrong. In general, an error occurs exactly when the first surviving edge before an incoming target is outgoing.

Lemma 4.2 (Weighted inversions)**.** *If $b_r^v(\tau) = 0$, then*

$$\rho_{v,r}^{(n)}(\tau) = \sum_{\substack{j<r \\ b_j^v(\tau)=1}} p_n(1-p_n)^{j-1}. \tag{6}$$

Consequently, if

$$I_v(\tau) = \left|\left\{(j,r) : j < r,\ b_j^v(\tau) = 1,\ b_r^v(\tau) = 0\right\}\right|,$$

then

$$\sum_{r:b_r^v(\tau)=0} \rho_{v,r}^{(n)}(\tau) \geq p_n(1-p_n)^{d-1} I_v(\tau). \tag{7}$$

Proof. The target e_r^v always survives, while every other incident edge survives independently with probability p_n. If $j < r$, all edges in positions $1, \ldots, j$ are non-target competitors. The learner selects the outgoing edge in position j exactly when that edge survives and the preceding $j-1$ edges do not, an event of probability

$$p_n(1-p_n)^{j-1}.$$

These selection events are disjoint, and an error occurs precisely when the first surviving edge before the target is outgoing. This proves Equation (6).

Summing over incoming target positions gives

$$\sum_{r:b_r^v=0} \rho_{v,r}^{(n)}(\tau) = \sum_{j:b_j^v=1} p_n(1-p_n)^{j-1}\left|\{r > j : b_r^v = 0\}\right|.$$

The summands on the left use the hard distributions associated with their respective target incidences; the display is a numerical sum, not a probability under a common training sample. Each weight on the right is at least $p_n(1-p_n)^{d-1}$, and the unweighted sum is $I_v(\tau)$. □

Set

$$n_q = \lceil \log_2(q-1) \rceil, \qquad p_q = 2^{-n_q}. \tag{8}$$

Then

$$\frac{1}{2d} \leq p_q \leq \frac{1}{d}. \tag{9}$$

Since

$$\left(1 - \frac{1}{d}\right)^{d-1} \geq e^{-1} \qquad (d \geq 2),$$

Equation (9) yields

$$p_q(1-p_q)^{d-1} \geq \frac{1}{2ed}. \tag{10}$$

Let

$$I(\tau) = \sum_{v \in V_q} I_v(\tau)$$

and

$$W(\tau) = \sum_{v \in V_q} \sum_{r:b_r^v(\tau)=0} \rho_{v,r}^{(n_q)}(\tau).$$

Applying Equation (10) to Equation (7) gives

$$W(\tau) \geq \frac{I(\tau)}{2e(q-1)}. \tag{11}$$

4.3 Cyclic triangles force inversions

Let $T(\tau)$ denote the number of directed cyclic triangles in the tournament τ.

Lemma 4.3 (Triangle charging)**.** *For every tournament $\tau \in \mathcal{T}_q$ and every strict total ordering of its edge hypotheses,*

$$I(\tau) \geq T(\tau).$$

Proof. Consider a directed cyclic triangle

$$u \longrightarrow v \longrightarrow w \longrightarrow u$$

with edges

$$a = \{u, v\}, \qquad b = \{v, w\}, \qquad c = \{w, u\}.$$

The three comparisons use the same strict total order at (q, τ). If there were no inversion at any vertex, then the incoming edge would precede the outgoing edge at each vertex, giving

$$a \prec b \prec c \prec a.$$

Every cyclic triangle therefore contains an inversion.

An inversion is an ordered pair of distinct incident edges. It determines the common vertex and the two remaining endpoints, hence a unique triangle. Inversions belonging to distinct triangles are therefore disjoint, so $I(\tau) \geq T(\tau)$. □

For a uniformly random tournament, each triple of vertices is cyclic with probability $1/4$. Therefore

$$\mathbb{E}_{\tau \sim \mathrm{Unif}(\mathcal{T}_q)} T(\tau) = \frac{1}{4}\binom{q}{3}. \tag{12}$$

Lemma 4.3, Equation (11), and the expectation in Equation (12) give

$$\mathbb{E}_\tau W(\tau) \geq \frac{1}{8e(q-1)}\binom{q}{3}. \tag{13}$$

4.4 Averaging over target incidences

For an incidence (e, v) with $e \in E_q$ and $v \in e$, let

$$R^{(q)}_{e,v} = \mathbb{E}_{S \sim D^{n_q}_{e,v}} L_{D_{e,v}}(A_\psi(S)).$$

Here the risk averages over the training sample and an independent test tournament from the same hard distribution. An unconditioned uniform tournament points a fixed edge toward either endpoint with probability $1/2$, which accounts for the factor of 2 in the proof below.

Lemma 4.4 (Incidence averaging)**.** *For every $q \geq 3$,*

$$\frac{1}{2M_q} \sum_{e \in E_q} \sum_{v \in e} R^{(q)}_{e,v} = \frac{1}{M_q} \mathbb{E}_{\tau \sim \mathrm{Unif}(\mathcal{T}_q)} W(\tau). \tag{14}$$

Proof. For a fixed incidence (e, v), population risk is the error probability on an independent test tournament conditioned on $\mathrm{head}_\tau(e) = v$:

$$R_{e,v}^{(q)} = \mathbb{E}_{\tau|\mathrm{head}_\tau(e)=v}\rho_{e,v}^{(n_q)}(\tau).$$

A uniform tournament orients e toward either endpoint with probability $1/2$. Thus

$$R_{e,v}^{(q)} = 2\mathbb{E}_\tau\left[\mathbf{1}\{\mathrm{head}_\tau(e) = v\}\rho_{e,v}^{(n_q)}(\tau)\right].$$

For fixed τ and e, exactly one endpoint is the head. Summing over all $2M_q$ incidences therefore gives

$$\sum_{e\in E_q}\sum_{v\in e} R_{e,v}^{(q)} = 2\mathbb{E}_\tau W(\tau).$$

Division by $2M_q$ proves Equation (14). □

Substituting Equation (13) into Equation (14) yields

$$\begin{aligned}\frac{1}{2M_q}\sum_{e\in E_q}\sum_{v\in e} R_{e,v}^{(q)} &\geq \frac{1}{8e(q-1)M_q}\binom{q}{3} \\ &= \frac{q-2}{24e(q-1)} \geq \frac{1}{48e}. \end{aligned} \tag{15}$$

Hence, for every $q \geq 3$, some incidence (e_q, v_q) satisfies

$$R_{e_q,v_q}^{(q)} \geq c_0, \qquad c_0 = \frac{1}{48e}. \tag{16}$$

4.5 Failure of PAC learning

Theorem 4.5. *No local regularizer learns $\mathcal{H}$.*

Proof. Fix ψ and the induced learner A_ψ constructed from Equation (3). For each q, choose (e_q, v_q) as in Equation (16), and let

$$Z_q = L_{D_{e_q,v_q}}(A_\psi(S)), \qquad S \sim D_{e_q,v_q}^{n_q}.$$

Then $0 \leq Z_q \leq 1$ and $\mathbb{E}Z_q \geq c_0$. For every $\eta \in (0, 1)$,

$$\mathbb{E}Z_q \leq \eta + \mathbb{P}[Z_q > \eta].$$

Taking $\eta = c_0/2$ gives

$$\mathbb{P}\left[Z_q > \frac{c_0}{2}\right] \geq \frac{c_0}{2}.$$

Set

$$\varepsilon_0 = \frac{c_0}{2} = \frac{1}{96e}, \qquad \delta_0 = \frac{c_0}{4} = \frac{1}{192e}.$$

Then, for every q,

$$\mathbb{P}_{S\sim D_{e_q,v_q}^{n_q}}\left[L_{D_{e_q,v_q}}(A_\psi(S)) > \varepsilon_0\right] \geq 2\delta_0 > \delta_0.$$

The sample sizes

$$n_q = \lceil \log_2(q-1) \rceil$$

are unbounded. Given a proposed PAC threshold m, choose $q \geq 2^m + 1$. Then $n_q \geq m$, and the realizable distribution D_{e_q,v_q} violates the $(\varepsilon_0, \delta_0)$ guarantee at sample size n_q. Thus A_ψ is not PAC. Since it is induced by the arbitrary regularizer ψ, that regularizer does not learn $\mathcal{H}$. □

Proof of Theorem 1.1. The sample-complexity upper bound is Proposition 3.1, the DS-dimension bound is Corollary 3.3, and the impossibility for local regularization is Theorem 4.5. □

5 Discussion

The learner in Proposition 3.1 succeeds after both endpoint labels of the target edge appear. A local regularizer has less freedom: at each tournament it fixes a ranking of the edge hypotheses before seeing the sample. Training data may remove edges from that ranking, but it cannot reorder those that remain.

This differs from the transductive lower bound of Jafar, Asilis, and Dughmi [5], which forces an error at a single withheld point. One such point may have negligible mass in a PAC distribution, and repeated sampling may reveal it. The present construction instead uses a constant-label distribution supported on exponentially many tournaments. Sampling filters many competitors, while population risk is evaluated on a fresh tournament from the same distribution.

The theorem applies to local scores fixed independently of the sample. It does not address the unsupervised local-regularization framework of Asilis et al. [2], where the rule may depend on the unlabeled sample.

6 Conclusion

The tournament class is realizably PAC learnable with sample complexity

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

and has Daniely–Shalev-Shwartz dimension at most two. Nevertheless, every local regularizer admits an induced learner that fails at arbitrarily large sample sizes. Under the nonvacuous formulation used here, local regularization therefore does not characterize realizable multiclass PAC learnability.

References

[1] Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Open problem: Can local regularization learn all multiclass problems? In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 5301–5305, 2024.

[2] 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*, pages 260–310, 2024.

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

[4] 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*, pages 287–316, 2014.

[5] 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*, pages 2942–2957, 2025.