Local Regularizers Are Not Transductive Learners

Sky Jafar JAFAR.SKY@GMAIL.COM
University High School

Julian Asilis ASILIS@USC.EDU
Shaddin Dughmi SHADDIN@USC.EDU
University of Southern California

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We partly resolve an open question raised by Asilis et al. (2024b,a): whether the algorithmic template of *local regularization* — an intriguing generalization of explicit regularization, a.k.a. structural risk minimization — suffices to learn all learnable multiclass problems. Specifically, we provide a negative answer to this question in the transductive model of learning. We exhibit a multiclass classification problem which is learnable in both the transductive and PAC models, yet cannot be learned transductively by any local regularizer. The corresponding hypothesis class, and our proof, are based on principles from cryptographic secret sharing. We outline challenges in extending our negative result to the PAC model, leaving open the tantalizing possibility of a PAC/transductive separation with respect to local regularization.

Keywords: Multiclass classification, transductive learning, regularization, PAC learning

1. Introduction

Understanding the power of various algorithmic templates for supervised learning is a core concern of both computational and statistical learning theory. This understanding arguably serves two purposes: a descriptive one by explaining the success of approaches employed in practice, and a prescriptive one which conveniently circumscribes the search space of promising algorithms for the practitioner. Most appealing are algorithmic approaches which are simple, natural, mirror what is seen in practice, and are powerful enough to learn in a rich variety of settings.

The most compelling success story in this vein is that of *empirical risk minimization (ERM)*. In practice, ERM and its approximations (such as gradient descent) are the work-horses of machine learning. In theory, ERM characterizes learnability for both binary classification (Vapnik, 1982; Blumer et al., 1989) and agnostic real-valued regression (Alon et al., 1997).[1] In second place is perhaps *Structural Risk Minimization (SRM)*, a template for generalizations of ERM which trade off empirical risk with a user-specified measure $\psi(\cdot)$ of model complexity, often referred to as a *regularizer*. In trading off the fit of a model h with its complexity $\psi(h)$, SRM protects against overfitting by encoding a preference for "simple" models, à la Occam's razor. Approaches such as ridge regression, Lasso, and other instantiations of SRM have seen much success in applied machine learning. On the theoretical front, ridge regression learns successfully for a large class of convex and smooth problems, and SRM in the abstract is known to characterize non-uniform learnability (Shalev-Shwartz and Ben-David, 2014).

1. That said, the sample complexity of ERM is slightly suboptimal for binary classification (Hanneke, 2016), and more significantly suboptimal for regression (Vaškevičius and Zhivotovskiy, 2023).

These natural and instructive algorithmic characterizations have been largely limited to small or low-dimensional label spaces as in regression and binary classification. The next frontier in this regard is the family of multiclass classification problems, where neither ERM, SRM, nor any "simple" algorithmic approach is known to learn whenever learning is possible. As evidence that any such characterization must employ more sophisticated algorithmic templates than ERM or SRM, Daniely and Shalev-Shwartz (2014) show that no *proper* learner can succeed in general; i.e., an optimal learner must sometimes "stitch together" different hypotheses to form its predictor. Improper algorithms which are optimal (or near-optimal) for multiclass classification have been described through orienting *one-inclusion graphs* (Rubinstein et al., 2006; Daniely and Shalev-Shwartz, 2014; Aden-Ali et al., 2023), and as a combination of *sample compression* and *list learning* (Brukhim et al., 2022). While lending remarkable structural insight into multiclass learning, neither approach is "simple" in any meaningful sense, nor reminiscent of approaches employed in practice.

Our starting point for this paper, and perhaps the closest thing to the sort of algorithmic characterization we seek for multiclass problems, is the recent work of Asilis et al. (2024b). They demonstrate that every multiclass problem can be learned by an *unsupervised local SRM (UL-SRM)* learner, a generalization of SRM on two fronts. First, the complexity of a hypothesis h is described *locally* per test point x via a *local regularizer* $\psi(h, x)$. Second, this local regularizer $\psi(\,\cdot\,,\,\cdot\,)$ is derived from the unlabeled data in what resembles an unsupervised pre-training stage.[2] The former generalization, locality, is what makes such a learner improper, and therefore appears indispensible by the impossibility result of Daniely and Shalev-Shwartz (2014). Whether the second generalization, which allows learning the regularizer from unlabeled data, can be dispensed with is less clear. This was posed as an open problem by Asilis et al. (2024a), where they conjecture that dependence on the unlabeled data is indeed necessary; i.e., that *local structural risk minimization* — a.k.a. *local regularization* — is insufficient for learning all learnable multiclass problems.

The present paper resolves this conjecture in the affirmative for the transductive model of learning: we construct a hypothesis class which is learnable in both the transductive and PAC models, and yet cannot be transductively learned by a local regularizer. Our hypothesis class can be viewed as a cryptographic generalization of the *first Cantor class* of Daniely and Shalev-Shwartz (2014), incorporating ideas from *secret-sharing*. Each of our hypotheses divides a large domain arbitrarily in half, with each half receiving its own label. We ensure that each hypothesis is uniquely identified by its two labels, which by itself suffices to render the class learnable. However, the two halves "share a secret," such that witnessing any one label reveals next to nothing about the hypothesis or the identity of the other label. While this does not obstruct learnability in general, we show that it does obstruct local regularizers: a transductive learner with vanishing error must exhibit a "cycle" in the typical test point's preferences over hypotheses, and therefore cannot be a local regularizer.

Although we were unable to extend our result to the PAC model, we suspect that this is a failure of our proof techniques rather than of our construction. We conjecture therefore that our same hypothesis class is in fact not learnable by any local regularizer in the PAC model. This would be in keeping with the tight relationship between the transductive and PAC models: learnability is equivalent in the two models, with sample-efficient reductions in both directions — see the discussion in (Dughmi et al., 2025; Dughmi, 2025) and references therein. Whereas the reduction from PAC to transductive learning preserves the form of the learner, the converse reduction does not,

2. We note that both local regularization and unsupervised pre-training have been successfully employed in the theory and practice of machine learning — see Wolf and Donner (2008); Prost et al. (2021); Vaškevičius and Zhivotovskiy (2023); Ge et al.; Azoury and Warmuth (2001); Vovk (2001).

obstructing black-box extensions of our result to the PAC model. Non-black-box approaches also appear challenging, and we outline the difficulties in Section 5. Irrespective of the outcome for the PAC model, we argue that our result for the transductive model is interesting in its own right due to the tight relationship to the PAC model in most other respects, the promising hypothesis class we design and employ for our proof, and the intricacies involved in proving our main theorem (which uses a coupling argument at its center). Furthermore, the prospect of separating the transductive and PAC models with respect to learnability by local regularizers — in the event that our conjecture is misguided — is even more tantalizing.

2. Preliminaries

2.1. Notation

For a natural number $d \in \mathbb{N}$, we denote $[d] = \{1, \ldots, d\}$. For Z a set, we use Z^* to denote the set of all finite sequences in Z, i.e., $Z^* = \bigcup_{n \in \mathbb{N}} Z^n$. In particular, $\{0,1\}^*$ denotes the set of all finite binary strings. For $A, B \in \{0,1\}^d$ binary strings of equal length, $A \oplus B$ refers to their entrywise XOR. We use $\sigma_0(A)$ and $\sigma_1(A)$ to refer to the entries at which A takes the value of 0 or 1, respectively. That is, $\sigma_0(A) = \big\{i \in [d] : A(i) = 0\big\}, \sigma_1(A) = \big\{i \in [d] : A(i) = 1\big\}$. We use $e(i)$ to denote the ith standard basis vector, whose length will always be clear from context (i.e., the vector with a 1 in its ith entry and zeroes elsewhere). For a statement P, $[P]$ denotes the Iverson bracket of P, i.e., $[P] = 1$ when P is true and 0 otherwise. For a function $f : A \to B$, we let $\mathrm{im}(f) = \{f(a) : a \in A\}$ denote the image of f in B.

2.2. Learning Theory

Unlabeled datapoints x are drawn from a **domain** $\mathcal{X}$ and labeled by an element of the **label set** $y \in \mathcal{Y}$. Pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are referred as labeled datapoints or *examples*. A **training set** or training sequence is a tuple of labeled datapoints $S \in (\mathcal{X} \times \mathcal{Y})^n$. When clear from context, we will refer to labeled and unlabeled datapoints simply as *datapoints*. A function $f \colon \mathcal{X} \to \mathcal{Y}$ is a **classifier** or **predictor**. A collection of classifiers $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is referred to as a **hypothesis class**; one of its elements $h \in \mathcal{H}$ is a **hypothesis**. Throughout the paper we employ the 0-1 loss function $\ell_{0-1} \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ defined by $\ell_{0-1}(y, y') = \big[y \neq y'\big]$. When $S = (x_i, y_i)_{i \in [n]}$ is a training set and $f \in \mathcal{Y}^{\mathcal{X}}$ a classifier, the average loss incurred by f on S is referred to as its **empirical risk** and denoted $L_S(f)$. That is, $L_S(f) = \frac{1}{n} \sum_{i \in [n]} \ell_{0-1}\big(f(x_i), y_i\big)$.

A **learner** is a function from training sets to classifiers, e.g., $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$. We focus on learning in the realizable case, for which the purpose of a learner is to deduce the behavior of a *ground truth* function $h^* \in \mathcal{H}$ given its behavior on a finite training set. More precisely, we adopt the transductive model of learning, as employed by Haussler et al. (1994) and originally introduced by Vapnik and Chervonenkis (1974) and Vapnik (1982).

Definition 1 *The **transductive model** of learning is that in which the following sequence of steps take place:*

1. *An adversary selects a collection of n unlabeled datapoints $S = (x_i)_{i \in [n]}$ and a hypothesis $h^* \in \mathcal{H}$.*
2. *The unlabeled datapoints S are displayed to the learner.*

3. *An index $j \in [n]$ is selected uniformly at random. The learner then receives the training set $(x_i, h^*(x_i))_{i \neq j}$.*
4. *The learner is prompted to predict the label of x_j, i.e., $h^*(x_j)$.*

We refer to the choice of S and h^* parameterizing the above steps as a *transductive instance*, and often collect this information into a pair, i.e., (S, h^*). The **transductive error** incurred by a learner on an instance (S, h^*) is its average loss at the test point x_j, over the uniformly random choice of $j \in [n]$, i.e.,

$$L^{\text{Trans}}_{S,h^*}(\mathcal{A}) = \frac{1}{n} \sum_{i \in [n]} \big[\mathcal{A}(S_{-i}, h^*)(x_i) \neq h^*(x_i)\big],$$

where $\mathcal{A}(S_{-i}, h^*)(x_i)$ denotes $\mathcal{A}$'s output on the sample consisting of datapoints in $S_{-i} = S \setminus x_i$ labeled by h^*. Naturally, a class is transductively learnable if it can be learned to vanishingly small error on increasingly large datasets.

Definition 2 *A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is **transductively learnable** if there exists a learner $\mathcal{A}$ and function $m\colon (0,1) \to \mathbb{N}$ such that for any $\varepsilon \in (0,1)$ and $h^* \in \mathcal{H}$, if $S \in \mathcal{X}^*$ has length $|S| \geq m(\varepsilon)$, then $L^{\text{Trans}}_{S,h^*}(\mathcal{A}) \leq \varepsilon$.*

The pointwise minimal function m such that there exists a learner $\mathcal{A}$ satisfying Definition 2 is referred to as the *transductive sample complexity* of learning $\mathcal{H}$, denoted $m_{\text{Trans},\mathcal{H}}$, and learners which attain this sample complexity are said to be *optimal*. (For our purposes, however, it will suffice to consider the property of learnability, without emphasizing sample complexities.) Notably, the transductive model of learning bears intimate connections with the Probably Approximately Correct (PAC) model of learning (Valiant, 1984), which considers underlying probability distributions and requires learners to perform well on randomly drawn test points when trained on i.i.d. training points. In particular, there is an equivalence between sample complexities in both models (up to a logarithmic factor), and techniques from transductive learning have recently been employed to establish the first characterizations of learnability for both multiclass classification and realizable regression (Brukhim et al., 2022; Attias et al., 2023). We defer a dedicated discussion of the PAC model to Section 5.

In the landscape of learning, perhaps the most fundamental learners are given by ERM and SRM, which operate by selecting a hypothesis h in the underlying class $\mathcal{H}$ with lowest empirical risk (possibly balanced with an inductive bias over hypotheses in $\mathcal{H}$, in the case of SRM).

Definition 3 *A learner $\mathcal{A}$ is an **empirical risk minimization** (ERM) learner for a class $\mathcal{H}$ if for all samples S,*

$$\mathcal{A}(S) \in \arg\min_{h \in \mathcal{H}} L_S(h).$$

Definition 4 *Let $\mathcal{H}$ be a hypothesis class. A **regularizer** for $\mathcal{H}$ is a function $\psi : \mathcal{H} \to \mathbb{R}_{\geq 0}$. A learner $\mathcal{A}$ is a **structural risk minimization** (SRM) learner for a class $\mathcal{H}$ if there exists a regularizer ψ such that for all samples S,*

$$\mathcal{A}(S) \in \arg\min_{h \in \mathcal{H}} \Big(L_S(h) + \psi(h)\Big).$$

2.3. Local Regularization

In realizable binary classification, ERM learners are known to succeed on all learnable classes, and furthermore to attain nearly-optimal sample complexity (Vapnik and Chervonenkis, 1974; Blumer et al., 1989; Ehrenfeucht et al., 1989). For multiclass learning over arbitrary label sets, however, these learning strategies are known to fail as a consequence of Daniely and Shalev-Shwartz (2014, Theorem 1). In particular, Daniely and Shalev-Shwartz (2014) establish that there exist multiclass problems which can only be learned by **improper learners**: learners which may emit classifiers outside of the underlying hypothesis class $\mathcal{H}$. (Note that ERM and SRM learners, in being phrased as arg min's over $\mathcal{H}$, are necessarily **proper**.) More recently, Asilis et al. (2025) expanded upon this result by demonstrating that there exist learnable multiclass problems which cannot be learned by any *aggregation* of a finite number of proper learners, ruling out such strategies as majority voting of ERM learners, which has recently found success in binary classification (Aden-Ali et al., 2024; Høgsgaard et al., 2024).

As such, multiclass learning in full generality requires different algorithmic blueprints than ERM and SRM. Perhaps the simplest such blueprint which has been considered is that of *local regularization*, as introduced by Asilis et al. (2024b). Put simply, local regularization augments the regularizer — usually a function $\psi\colon \mathcal{H} \to \mathbb{R}_{\geq 0}$ — to additionally receive an unlabeled datapoint as input, i.e., $\psi\colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. The unlabeled datapoint is taken to be the test datapoint at which the learner is tasked with making a prediction. At each test datapoint $x \in \mathcal{X}$, the learner then takes the behavior of the hypothesis $\widehat{h}$ with minimal value of $\psi(\widehat{h}, x)$, subject to $L_S(\widehat{h}) = 0$.

This bears two crucial advantages over classical regularization.

1. Local regularization models the *location-dependent* complexity of a hypotheses. It may be that h acts as a simple function on a region $U \subseteq \mathcal{X}$ yet as a complex function on $V \subseteq \mathcal{X}$ (and vice versa for $h' \in \mathcal{H}$). A local regularizer ψ can model this behavior as $\psi(h, u) < \psi(h', u)$ for $u \in U$ and $\psi(h, v) > \psi(h', v)$ for $v \in V$. A classical regularizer, in contrast, is obligated to assign each of h and h' with a single value encoding their global complexity.

2. Local regularizers can induce improper learners, by emitting classifiers that "stitch together" various hypotheses of $\mathcal{H}$. In the previous case, for instance, one can imagine a learner $\mathcal{A}$ induced by a local regularizer which takes the behavior of h on $U \subseteq \mathcal{X}$ and of h' on $V \subseteq \mathcal{X}$.

Definition 5 (Asilis et al. (2024a)) *A **local regularizer** for a hypothesis class $\mathcal{H}$ is a function $\psi\colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. A learner $\mathcal{A}$ is said to be **induced** by ψ if for all samples S and datapoints $x \in \mathcal{X}$,*

$$\mathcal{A}(S)(x) \in \left\{ h(x) : h \in \operatorname*{arg\,min}_{h \in L_S^{-1}(0)} \psi(h, x) \right\}.$$

*We say that ψ **transductively learns** $\mathcal{H}$ if all learners it induces are transductive learners for $\mathcal{H}$.*

Remark 6 *Restricting to the set of hypotheses which incur zero empirical error in Definition 5, rather than minimizing the sum of hypotheses' empirical error and regularization value, has the effect of simplifying the analysis of regularization. Furthermore, simply normalizing a regularizer to take outputs in the range $[0, 1/n]$ has the effect of equating the two perspectives for samples of cardinality at most n. (Though this procedure is not uniform with respect to all possible sample sizes.)*

Asilis et al. (2024a) posed the open problem of whether local regularization is sufficiently expressive to learn all multiclass problems possible.

Open Problem 1 (Asilis et al. (2024a)) *In multiclass classification, can all learnable hypothesis classes be learned by a local regularizer? If so, with optimal (or nearly optimal) sample complexity?*

In Theorem 18 we resolve Open Problem 1 for the transductive model of learning, by exhibiting a learnable class which cannot be transductively learned by any local regularizer. In Section 5 we discuss the possibility of extending this result to the PAC model, and describe some of the challenges involved.

2.4. Secret Sharing

We now provide a brief review of topics in *secret sharing*, as the proof of our primary result, Theorem 18, employs a hypothesis class whose structure is intimately related to concepts from the field.

Secret sharing refers to the fundamental task of distributing private information, a *secret*, among a group of *players*. A successful solution consists of a technique for distributing information from a single *dealer* to each player in such a manner that any individual player is incapable of recovering the secret, yet it can be revealed when the group works in concert. The information revealed to each player individually is referred to as a *share*. In the general setting, there are n players and the secret should be recoverable by any group of t cooperating players, but not by any smaller group. (More precisely, any smaller group should not even be able to deduce partial information concerning the secret.) A solution to this problem is referred to as a (t, n)-threshold scheme.[3]

The study of secret sharing dates to the work of Shamir (1979) and Blakley (1979); for a contemporary introduction, see Beimel (2011) or Krenn and Lorünser (2023). Shamir (1979) designed an elegant (t, n)-threshold scheme which proceeds as follows: Let $q > n$ be a sufficiently large prime to contain all possible secrets, $k \in [q]$ the secret, and select $a_1, \ldots, a_{t-1}$ uniformly at random from $[q]$. Lastly, define the polynomial $P(x) = k + \sum_{i=1}^{t-1} a_i x^i \mod q$, and distribute to the jth player the share $(j, P(j))$. Then the cooperation of any t players suffices to reveal P (and thus k) owing to Lagrange's interpolation theorem, yet any smaller collection of players can reveal no information concerning $k = P(0)$ owing to the uniformly random choices of $a_1, \ldots, a_{t-1}$.

For our purposes, however, it will suffice to consider the basic case of $t = n = 2$, for which there is a strikingly simple secret-sharing method referred to as the **one-time pad** (OTP) in cryptography. Namely, given a secret $C \in \{0,1\}^n$, the one-time pad selects a string A uniformly at random from $\{0,1\}^n$, and distributes to player one the share A and to player two the share $A \oplus C$. Individually, then, each player witnesses a uniformly random string of length n, yet the XOR of their shares is precisely the secret, $A \oplus (A \oplus C) = C$. In Section 4, we design a hypothesis class inspired by the one-time pad, for which — roughly speaking — each hypothesis h is represented by a secret and each of its two possible outputs reveals one share of the secret. Crucially, then, the information of one of h's secrets (i.e., its behavior on one half of the domain) maintains the identity of its remaining secret completely opaque (i.e., its behavior on the remaining half of the domain).[4]

3. In an even-more-general setting, the problem is defined by a collection of *accessor subsets*, i.e., sets of players that should be able to deduce the secret when cooperating (Ito et al., 1989).
4. Strictly speaking, for the hypothesis class we construct, one of h's "secrets" and its behavior on a region of the domain can reveal partial information concerning its other "secret."

3. GBDLS Classes

We devote this section to the development of two properties (of a hypothesis class) which will play a central role in the proof of Theorem 18: that of being *generalized binary* and of having *distinct label sets*. Upon defining each such property (of an underlying hypothesis class), we will demonstrate that neither is individually sufficient to ensure learnability. However, we show that all classes which have both properties — termed *generalized binary with distinct label sets* (GBDLS) — are necessarily learnable. Neither observation is particularly difficult to prove, but serves towards the eventual proof of Theorem 18, and we believe that the language of such hypothesis classes may be of independent interest.

Definition 7 *A hypothesis $h\colon \mathcal{X} \to \mathcal{Y}$ is **k-ary** if $|\mathrm{im}(h)| \leq k$. When $k = 2$, we say that h is a binary hypothesis.*

Definition 8 *A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is **k-ary** if $|\bigcup_{h\in\mathcal{H}} \mathrm{im}(h)| \leq k$. When $k = 2$, we say that $\mathcal{H}$ is binary.*

We have slightly overloaded the term *k-ary*, but there should be little risk of confusion: a hypothesis h is k-ary if and only if the class $\{h\}$ is k-ary.

Definition 9 *A hypothesis class is **generalized k-ary** if each $h \in \mathcal{H}$ is a k-ary hypothesis. When $k = 2$, we say that $\mathcal{H}$ is generalized binary.*

Definition 10 *A hypothesis class $\mathcal{H}$ **has distinct label sets** if $\mathrm{im}\colon \mathcal{H} \to 2^{\mathcal{Y}}$ is an injection. That is, $\mathrm{im}(h) \neq \mathrm{im}(h')$ when $h \neq h'$ and $h, h' \in \mathcal{H}$.*

It is not difficult to exhibit counter-examples demonstrating that neither the property of being generalized binary nor the property of having distinct label sets is sufficient to ensure that a hypothesis class $\mathcal{H}$ is learnable. For the former, take any binary class of infinite VC dimension. For the latter, take any nonlearnable class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and expand each of its hypotheses $h \in \mathcal{H}$ to be defined on an additional unlabeled datapoint $*$ by $h(*) = h|_{\mathcal{X}}$. However, we refer to classes with both properties as being *generalized binary with distinct label sets* (GBDLS), and we now demonstrate that GBDLS classes are always learnable.

We will appeal to the fact that learnability of hypothesis classes, in both the PAC and transductive models, is characterized by finiteness of the *DS dimension*, as demonstrated by the seminal work of Brukhim et al. (2022). Recall that the DS dimension measures the expressivity of classes $\mathcal{H}$ on finite collections of unlabeled datapoints, as we now describe.

Definition 11 (Daniely and Shalev-Shwartz (2014)) *A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$* DS-shatters *a finite set of points $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ if there exists a finite, non-empty set of functions $\mathcal{F} \subseteq \mathcal{H}$ such that for any $f \in \mathcal{F}$ and any $i \in [n]$, there exists $g \in \mathcal{F}$ such that $g(x_i) \neq f(x_i)$ yet $g(x_j) = f(x_j)$ for all $j \neq i$. The **DS dimension** of $\mathcal{H}$, denoted $\mathrm{DS}(\mathcal{H})$, is the cardinality of the largest DS-shattered set, or ∞ if arbitrarily large sets are shattered.*

Learnability of GBDLS classes follows from the fact that they necessarily enjoy a finite DS dimension — in fact, a DS dimension of at most 2.

Proposition 12 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a GBDLS hypothesis class. Then $\mathcal{H}$ is learnable, in both the PAC and transductive models.*

Proof We will demonstrate that $\mathcal{H}$ cannot DS-shatter any sequence of 3 distinct points, from which it follows that $\mathrm{DS}(\mathcal{H}) \leq 2$ and thus that $\mathcal{H}$ is learnable in both models (Brukhim et al., 2022). Fix any sequence of distinct points $S = (x_1, x_2, x_3) \subseteq \mathcal{X}$, and let $\mathcal{F} \subseteq \mathcal{H}|_S$ be a finite non-empty set of behaviors of $\mathcal{H}$ on S. It remains to show that there exists an $f \in \mathcal{F}$ and $i \in [3]$ for which f does not have an i-neighbor in $\mathcal{F}$ (i.e., a $g \in \mathcal{F}$ with $g|_{S\setminus\{x_i\}} = f|_{S\setminus\{x_i\}}$ and $g(x_i) \neq f(x_i)$).

To this end, select an arbitrary behavior $h \in \mathcal{F}$, and identify it with the sequence $(h(x_j))_{j\in[3]} \subseteq \mathcal{Y}$. Suppose that h is a constant behavior, i.e., that $h = (a, a, a)$ for a label $a \in \mathcal{Y}$. If h does not have a 3-neighbor $f \in \mathcal{F}$, then we are done. Otherwise, consider the 3-neighbor $f = (a, a, b)$ of $h \in \mathcal{F}$. Note that if h had been a non-constant behavior, then it would have taken the form $h = (a, a, b)$ to begin with (up to re-ordering of the points (x_1, x_2, x_3), owing to the fact that h is a binary behavior). We may thus consider a behavior $f = (a, a, b) \in \mathcal{F}$ where $a \neq b$, without loss of generality. Then it follows immediately that f does not have a 1-neighbor $g \in \mathcal{F}$. Any such g must take the form $g = (b, a, b)$, as $\mathcal{H}$ is generalized binary, and this yields that $\mathrm{im}(f) = \{a, b\} = \mathrm{im}(g)$, producing contradiction with the fact that $\mathcal{H}$ has distinct label sets. ■

4. Insufficiency of Local Regularization in the Transductive Model

We are now equipped to prove the primary result of the paper: There exists a learnable hypothesis class $\mathcal{H}_{\mathrm{otp}}$ which cannot be transductively learned by any local regularizer (Theorem 18). We begin by defining the class $\mathcal{H}_{\mathrm{otp}}$ which witnesses this separation. Notably, $\mathcal{H}_{\mathrm{otp}}$ is intimately related to secret sharing techniques, including the *one-time pad*. More precisely, each hypothesis $h \in \mathcal{H}_{\mathrm{otp}}$ is parameterized by two strings $A, B \in \{0,1\}^*$, and $h(i) = (0, A)$ or $h(i) = (1, B)$ depending upon the ith bit of $A \oplus B$. When a training set S contains two distinct labels, learning is trivial: only the ground truth function attains zero training error. When the data distribution places full measure on a single label, however, correctly predicting an unseen test point j requires a local regularizer which favors functions mapping j to that same label. (And, owing to the structure of the one-time pad, knowledge of $A \oplus B$ at the training points and complete knowledge of A reveals little information regarding $A \oplus B$ at an unseen test point.) Using a coupling argument, we demonstrate that no local regularizer can do so simultaneously for all labels.

Recall that for binary strings $A, B \in \{0,1\}^n$, $A \oplus B$ denotes their entrywise XOR. We set $\sigma_0(A) = \{i \in [n] : A(i) = 0\}$ and likewise $\sigma_1(A) = \{i \in [n] : A(i) = 1\}$. In the event that $|\sigma_0(A)| = |\sigma_1(A)|$, A is said to be *balanced*.

Definition 13 *Let $\mathcal{X} = \mathbb{N}$ and $\mathcal{Y} = \{0,1\} \times \{0,1\}^*$. For each $d \in \mathbb{N}$ and $A, B \in \{0,1\}^d$, define*

$$h_{A,B} \colon \mathcal{X} \longrightarrow \mathcal{Y}$$
$$x \longmapsto \begin{cases} (0, A) & (A \oplus B)(x \bmod d) = 0, \\ (1, B) & (A \oplus B)(x \bmod d) = 1. \end{cases}$$

Then the hypothesis class $\mathcal{H}_{\mathrm{otp}} \subseteq \mathcal{Y}^{\mathcal{X}}$ is defined as

$$\mathcal{H}_{\mathrm{otp}} = \Big\{ h_{A,B} : A, B \in \{0,1\}^*, |A| = |B|, A \oplus B \textit{ is balanced} \Big\}.$$

Remark 14 *For those familiar with the* first Cantor class *of Daniely and Shalev-Shwartz (2014), denoted $\mathcal{H}_\infty$, note that $\mathcal{H}_{\text{otp}}$ can be seen as generalizing this class. In particular, $\mathcal{H}_\infty$ is defined by "glueing together" the classes $\{\mathcal{H}_d\}_{d\in\mathbb{N}}$, where $\mathcal{H}_d \subseteq \mathcal{X}_d^{\mathcal{Y}_d}$, $\mathcal{X}_d$ is a set of size d, $\mathcal{Y}_d = 2^{\mathcal{X}_d} \cup \{*\}$, and $\mathcal{H}_d = \left\{h_A : A \subseteq \mathcal{X}_d, |A| = d/2\right\}$ where $h_A(x) = A$ if $x \in A$ and $*$ otherwise. Each such $\mathcal{H}_d$ can equivalently be defined in the language of Definition 13 by identifying each $A \subseteq \mathcal{X}_d$ with its characteristic vector, setting $h_A := h_{A,1^d}$, and considering the relabeling $(1, 1^d) \mapsto *$, $(0, A) \mapsto A$.*

Lemma 15 *$\mathcal{H}_{\text{otp}}$ is a GBDLS class.*

Proof It is immediate from Definition 13 that each $h \in \mathcal{H}$ has $|\text{im}(h)| = 2$, meaning $\mathcal{H}$ is generalized binary. Furthermore, if $f \in \mathcal{H}$ is such that $\text{im}(f) = \{(0, A), (1, B)\}$ then it must be that $f = h_{A,B}$. Thus $\mathcal{H}_{\text{otp}}$ has distinct label sets. ■

We first equate the task of learning with a local regularizer to learning with a local regularizer which is *locally injective*, i.e., injective on $\mathcal{H}$ for each fixed choice of test point $x \in \mathcal{X}$.

Definition 16 *A local regularizer $\psi\colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is **locally injective** if $\psi(\,\cdot\,, x)$ is injective for each $x \in \mathcal{X}$. That is, $\psi(h, x) \neq \psi(h', x)$ for any $x \in \mathcal{X}$ and $h \neq h' \in \mathcal{H}$.*

The following lemma establishes that for countable hypothesis classes $\mathcal{H}$, their local regularizers ψ can be assumed to be locally injective (i.e., to totally order $\mathcal{H}$ at each location $x \in \mathcal{X}$, rather than merely partially order). Note that in this case, ψ induces a unique learner, as there is no "tie-breaking" left to its induced learners.

Lemma 17 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a countable hypothesis class. Then $\mathcal{H}$ can be learned by a local regularizer if and only if it can be learned by one which is locally injective.*

Proof The backward direction is immediate. For the forward, let ψ be a local regularizer which learns $\mathcal{H}$, meaning that each of its induced learners succeeds on $\mathcal{H}$. Recall that for each $x \in \mathcal{X}$, $\psi(\cdot, x)$ defines a (strict) partial order over $\mathcal{H}$, i.e., with $h < h'$ if $\psi(h, x) < \psi(h', x)$. For each such x, let $\overline{\psi(\cdot, x)}$ be any completion of this partial order into a total ordering. As $\mathcal{H}$ is countable, $\overline{\psi(\cdot, x)}$ can be embedded into the real numbers. Define $\overline{\psi}(\cdot, x)$ using such an embedding, and consider the unique learner $\mathcal{A}$ induced by $\overline{\psi}$. As $\overline{\psi}$ acts as a completion of ψ at each $x \in \mathcal{X}$, then $\mathcal{A}$ is also a learner induced by ψ. Thus $\mathcal{A}$ learns $\mathcal{H}$, by our assumption on ψ, meaning $\overline{\psi}$ learns $\mathcal{H}$. ■

Theorem 18 *$\mathcal{H}_{\text{otp}}$ is a learnable hypothesis class, but cannot be transductively learned by any local regularizer.*

Proof Fix a local regularizer ψ and $d \in \mathbb{N}$. By Lemma 17, we may assume that ψ is locally injective, inducing a unique learner $\mathcal{A}$. Using a probabilistic argument, we will demonstrate that there exists a transductive learning instance (S, h^*) with $|S| = d$ for which ψ incurs error at least $\frac{1}{4}$. To this end, define the following independent random variables:

- C is drawn uniformly at random from all balanced strings of length $2d$.
- A is drawn uniformly at random from $\{0, 1\}^{2d}$.
- m_0 and m_1 are drawn uniformly at random from $[d]$.

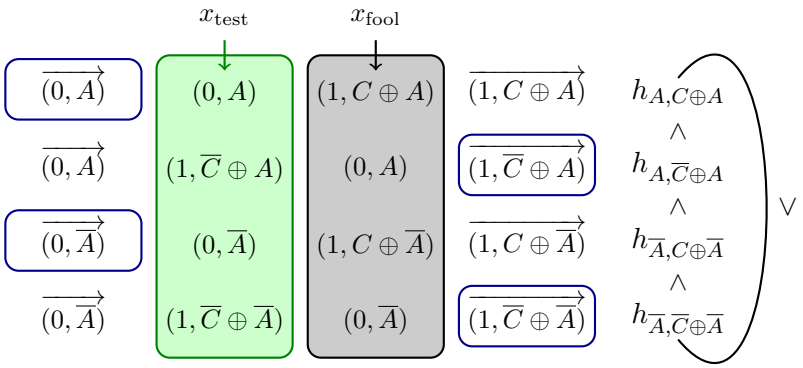

Figure 1: Depiction of the learning problems $\{(S_i, h_i^*)\}_{i \in [4]}$ in Theorem 18, with training sets S_i circumscribed in blue. Each row corresponds to one of the learning problems (S_i, h_i^*). Crucially, success of a local regularizer ψ on each learning problem imposes the ordering relations on the right-hand side, which collectively produce a cycle.

Further, define x_{test} to be the index of the m_0-th 0 entry in C, and x_{fool} to be the index of the m_1-th 1 entry in C. (Note that such entries exist for any value of $m_0, m_1 \in [d]$, owing to the fact that C is balanced.)

We now define four random variables $\{T_i\}_{i \in [4]}$ based upon the previous variables. Each one is parameterized by a tuple $(S_i, h_i^*)_{i \in [4]}$ and measures the performance of $\mathcal{A}$ at the test point x_{test} when trained the datapoints in $S_i \setminus \{x_{\text{test}}\}$ labeled by h_i^*, the ground truth function. The S_i and h_i^* are defined as follows:

(1.) Let $h_1^* = h_{A, C \oplus A}$, $S_1 = \sigma_0(C)$. Thus

$$T_1 = \Big[\mathcal{A}\big(S_1 \setminus \{x_{\text{test}}\}, h_1^*\big)(x_{\text{test}}) \neq h_1^*(x_{\text{test}})\Big],$$

where $\mathcal{A}\big(S_1 \setminus \{x_{\text{test}}\}, h_1^*\big)$ denotes the output of $\mathcal{A}$ when trained on the dataset consisting of the points in $S_1 \setminus \{x_{\text{test}}\}$ labeled by h_1^*.

(2.) Let $h_2^* = h_{A, \overline{C} \oplus A}$, where $\overline{C}$ equals C with its x_{test}-th and x_{fool}-th entries each flipped.[5] Let $S_2 = \sigma_1(\overline{C}) = \sigma_1(C) \cup \{x_{\text{test}}\} \setminus \{x_{\text{fool}}\}$. Thus $T_2 = \big[\mathcal{A}\big(S_2 \setminus \{x_{\text{test}}\}, h_2^*\big)(x_{\text{test}}) \neq h_2^*(x_{\text{test}})\big]$.

(3.) Let $h_3^* = h_{\overline{A}, C \oplus \overline{A}}$ where $\overline{A}$ denotes A with its x_{test}-th and x_{fool}-th entries each flipped. Let $S_3 = \sigma_0(C) = S_1$. Thus $T_3 = \big[\mathcal{A}\big(S_3 \setminus \{x_{\text{test}}\}, h_3^*\big)(x_{\text{test}}) \neq h_3^*(x_{\text{test}})\big]$.

(4.) Let $h_4^* = h_{\overline{A}, \overline{C} \oplus \overline{A}}$, $S_4 = \sigma_1(\overline{C}) = S_2$. Thus $T_4 = \big[\mathcal{A}\big(S_4 \setminus \{x_{\text{test}}\}, h_4^*\big)(x_{\text{test}}) \neq h_4^*(x_{\text{test}})\big]$.

It remains to prove the following lemma.

5. Note that $\overline{C}$ is balanced, as x_{test} corresponds to a 0 entry in C and x_{fool} to a 1 entry.

Lemma 19 *Each random variable T_i, for $i \in [4]$, is distributed as the error ψ incurs on the (randomly-chosen) transductive learning instance (S_i, h_i^*). Furthermore, for any value of the variables C, A, m_0, and m_1, the local regularizer ψ must err at x_{test} in at least one of the instances $(S_i, h_i^*)_{i\in[4]}$. (That is, $\max_i T_i = 1$.)*

Proof The first claim amounts to demonstrating that for each $i \in [4]$, conditioned upon h_i^* and S_i, x_{test} is distributed uniformly randomly across S_i. For $i \in \{1, 3\}$, this is immediate: x_{test} is selected as a uniformly random element of $\sigma_0(C)$ and $S_1 = S_3 = \sigma_0(C)$. For $i \in \{2, 4\}$, recall that $S_2 = S_4 = \sigma_1(C) \cup \{x_{\text{test}}\} \setminus \{x_{\text{fool}}\}$, and that x_{test} and x_{fool} are independent and uniformly random elements of $\sigma_0(C)$ and $\sigma_1(C)$, respectively. Further, C is chosen uniformly at random from the set of all balanced $2d$-strings. Then, conditioned upon $\overline{C}$, the probability that x_{test} takes the value of a given $s \in S = \sigma_1(\overline{C})$ is proportional to the cardinality of

$$\begin{aligned}
&\left\{(\widehat{C}, \widehat{x}_{\text{fool}}) : s \in \sigma_0(\widehat{C}), \widehat{x}_{\text{fool}} \in \sigma_1(\widehat{C}), \widehat{C} \text{ is balanced}, \widehat{C} \oplus e(s) \oplus e(\widehat{x}_{\text{fool}}) = \overline{C}\right\} \\
&\qquad = \left\{(\widehat{C}, \widehat{x}_{\text{fool}}) : s \in \sigma_0(\widehat{C}), \widehat{x}_{\text{fool}} \in \sigma_1(\widehat{C}), \widehat{C} \text{ is balanced}, \widehat{C} = \overline{C} \oplus e(s) \oplus e(\widehat{x}_{\text{fool}})\right\} \\
&\qquad = \left\{(\widehat{C}, \widehat{x}_{\text{fool}}) : s \in \sigma_1(\overline{C}), \widehat{x}_{\text{fool}} \in \sigma_0(\overline{C}), \widehat{C} \text{ is balanced}, \widehat{C} = \overline{C} \oplus e(s) \oplus e(\widehat{x}_{\text{fool}})\right\} \\
&\qquad = \left\{(\widehat{C}, \widehat{x}_{\text{fool}}) : \widehat{x}_{\text{fool}} \in \sigma_0(\overline{C}), \widehat{C} = \overline{C} \oplus e(s) \oplus e(\widehat{x}_{\text{fool}})\right\} \\
&\qquad \cong \left\{\widehat{x}_{\text{fool}} : \widehat{x}_{\text{fool}} \in \sigma_0(\overline{C})\right\}.
\end{aligned}$$

The first equality follows from the observation that $\widehat{C} \oplus e(s) \oplus e(\widehat{x}_{\text{fool}}) = \overline{C}$ is symmetric in $\widehat{C}$ and $\overline{C}$. The second equality rephrases the conditions on s and $\widehat{x}_{\text{fool}}$ in terms of $\overline{C}$, rather than $\widehat{C}$. The third equality employs the fact that $s \in \sigma_1(\overline{C})$ by definition, and that any $\widehat{C}$ satisfying the remaining conditions is automatically balanced. Thus the cardinality of this set does not depend upon s, meaning the posterior probability of x_{test} taking the value of any $s \in \sigma_1(C)$ is uniform, as desired.

For the second claim, note that $h_1^*(x_{\text{test}}) \neq h_2^*(x_{\text{test}})$, as $C(x_{\text{test}}) = 0$ yet $\overline{C}(x_{\text{test}}) = 1$. However, h_2^* attains zero empirical error on $S_1 \setminus \{x_{\text{test}}\}$, as for any $s \in S_1 \setminus \{x_{\text{test}}\}$,

$$h_1^*(s) = h_{A, C\oplus A}(s) = (0, A) = h_{A, \overline{C}\oplus A}(s) = h_2^*(s),$$

where the first and third equalities use the fact that $s \in S_1 \setminus \{x_{\text{test}}\} \subseteq \sigma_0(C) \cap \sigma_0(\overline{C})$. Thus, in order for ψ to correctly classify task (1.), it must be that $\psi(h_1^*, x_{\text{test}}) < \psi(h_2^*, x_{\text{test}})$. Likewise, $h_2^*(x_{\text{test}}) \neq h_3^*(x_{\text{test}})$ because $C(x_{\text{test}}) = 0 \neq 1 = \overline{C}(x_{\text{test}})$, yet h_3^* attains zero empirical error on $S_2 \setminus \{x_{\text{test}}\}$, as $S_2 \setminus \{x_{\text{test}}\} \subseteq \sigma_1(C) \cap \sigma_1(\overline{C})$ and $\overline{C} \oplus A = C \oplus \overline{A}$. Thus, for ψ to correctly classify task (1.) would require that $\psi(h_2^*, x_{\text{test}}) < \psi(h_3^*, x_{\text{test}})$.

Invoke this reasoning twice more with h_3^* and h_4^*; see Figure 1. In particular, $h_3^*(x_{\text{test}}) \neq h_4^*(x_{\text{test}})$ as $C(x_{\text{test}}) \neq \overline{C}(x_{\text{test}})$, yet $S_3 \setminus \{x_{\text{test}}\} = \sigma_0(C) \setminus \{x_{\text{test}}\} \subseteq \sigma_0(C) \cap \sigma_0(\overline{C})$. Thus ψ's success on T_3 relies upon $\psi(h_3^*, x_{\text{test}}) < \psi(h_4^*, x_{\text{test}})$. Finally, $h_4^*(x_{\text{test}}) \neq h_1^*(x_{\text{test}})$, but $S_4 \setminus \{x_{\text{test}}\} = \sigma_1(\overline{C}) \setminus \{x_{\text{test}}\} \subseteq \sigma_1(C) \cap \sigma_1(\overline{C})$ and $C \oplus A = \overline{C} \oplus \overline{A}$. Thus success of ψ on T_4 imposes the final requirement that $\psi(h_4^*, x_{\text{test}}) < \psi(h_1^*, x_{\text{test}})$. It follows immediately that ψ cannot succeed on all at once. ■

By the second claim of Lemma 19, the error ψ incurs at x_{test}, on average over the instances $(S_i, h_i^*)_{i\in[4]}$, is $\frac{1}{4}\sum_{i=1}^{4}\mathbb{E}[T_i] \geq \frac{1}{4}$. By the first claim of Lemma 19 and a use of the probabilistic method, this implies the existence of a single transductive learning instance (on d points) for which ψ incurs transductive error at least $\frac{1}{4}$. Conclude by recalling that d was chosen arbitrarily. ■

5. Challenges in Extending to the PAC Model

Let us now describe by some of the challenges involved in extending Theorem 18 to the *Probably Approximately Correct* (PAC) learning model of Valiant (1984). We begin by recalling the model.

Definition 20 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. A probability measure $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is $\mathcal{H}$-**realizable** if there exists an $h \in \mathcal{H}$ for which*

$$L_{\mathcal{D}}(h) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\big(\ell_{0-1}(h(x), y)\big) = 0.$$

Definition 21 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and $\mathcal{A}$ a learner. $\mathcal{A}$ is said to be a **PAC learner** for $\mathcal{H}$ if there exists a function $m\colon (0,1)^2 \to \mathbb{N}$ with the following property: for any $\mathcal{H}$-realizable distribution $\mathcal{D}$ and any $\varepsilon, \delta \in (0,1)$, if $S \sim \mathcal{D}^n$ is a sample of size $n \geq m(\varepsilon, \delta)$ drawn i.i.d. from $\mathcal{D}$, then*

$$L_{\mathcal{D}}(\mathcal{A}(S)) \leq \varepsilon$$

with probability at least $1 - \delta$ over the random choice of S.

In short, the PAC model differs from the transductive model by considering underlying probability distributions $\mathcal{D}$ and requiring learners to attain favorable performance on $\mathcal{D}$-i.i.d. test points when trained on $\mathcal{D}$-i.i.d. training sets. Notably, the condition of learnability (by an arbitrary learner) is equivalent for both the PAC and transductive models. Furthermore, there are efficient reductions for converting optimal PAC learners into nearly-optimal transductive learners (and vice versa), demonstrating an equivalence between sample complexities in both models, up to only a logarithmic factor in the error parameter (Dughmi et al., 2025). Converting a PAC learner $\mathcal{A}$ into a transductive learner, however, requires randomly sampling from the training set S and calling $\mathcal{A}$ on the resulting dataset. This procedure does *not* preserve the algorithmic form of transductive learners (as the resulting learner may even misclassify a point in S), and thus Theorem 18 does not imply — at a black-box level — the failure of local regularizers in the PAC model.

We now describe three primary sources of difficulty in porting Theorem 18 to the PAC model.

- **Favoring non-ground-truth hypotheses.** The proof of Theorem 18, and in particular of Lemma 19, makes use of a train/test setting in which there are exactly two hypotheses attaining zero empirical error, only one of which — the ground truth hypothesis — is correct for the test point. Using this fact, we are able to deduce precise inequalities relating the various candidate ground truth hypotheses considered, which collectively lead to a contradiction. (I.e., to a cycle in the preferences of the local regularizer.) When learning in the PAC model, one immediately loses such tight control over the train/test setup, as both datasets are drawn i.i.d. from the underlying distribution $\mathcal{D}$. In this setting, there will typically be various hypotheses attaining zero empirical error which are also correct at the test point, offering considerably more freedom to a successful local regularizer and (seemingly) prohibiting the detection of simple cycles.

- **Incomparable version spaces.** Another approach may be as follows: Note that a successful local regularizer ψ for $\mathcal{H}_{\text{otp}}$ must, at a minimum, perform well on average when $A \in \{0,1\}^n$ and $b \in \{0,1\}$ are selected uniformly at random, the training set S consist of $m < n$ points drawn uniformly at random from $[n]$ whose labels are all (b, A), and the test point x_{test} is likewise a uniformly random point in $[n]$. (Whose correct label is (b, A).) To deduce a contradiction from this strong condition imposed on ψ requires considering its behavior on sets of ERM hypotheses $L_S^{-1}(0)$ for varying training sets S. (These sets are often referred to as the *version spaces* of S (Mitchell, 1977).) In almost all cases, distinct version spaces will be incomparable as sets (i.e., neither subsets nor supersets of one another), rendering it difficult to derive contradictions from such conditions.

- **Error measurement.** One may note that by drawing m points uniformly at random from a set $S = \big((x_i, y_i)\big)_{i \in [n]}$ for carefully chosen $m = \Theta(n \log n)$, it will occur with constant probability that exactly one point in S is *not* drawn. Naïvely, this would seem to recover transductive learning as a special instance of PAC learning. Crucially, however, the transductive model places full weight upon the learner's performance at the (unseen) test point, whereas the PAC model averages a learner's performance across the entire distribution $\mathcal{D} = \text{Unif}(S)$. In the PAC model, then, simply memorizing the training set suffices to learn $\text{Unif}(S)$ to small error in $\Theta(n \log n)$ samples.

6. Conclusion

We study perhaps the simplest candidate template for multiclass learning: local regularization (Asilis et al., 2024a). As our primary result, we demonstrate that there exists a learnable hypothesis class which cannot be transductively learned by any local regularizer. The hypothesis class $\mathcal{H}_{\text{otp}}$ which we employ for this result is based upon techniques from *secret-sharing*, such as the one-time pad, and generalizes the *first Cantor class* of Daniely and Shalev-Shwartz (2014). We conjecture that the same class also cannot be learned by any local regularizer in the PAC model, though we highlight some of the difficulties involved in extending our result in Section 5.

Acknowledgments

The authors thank – in alphabetical order – Siddartha Devic, Vatsal Sharan, and Shang-Hua Teng for many useful conversations regarding the role of regularization in learning. Part of this work was done during the course of a research program administered by Pioneer academics; Sky Jafar and Shaddin Dughmi thank Pioneer and its staff for their support. Julian Asilis was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1842487, and completed this work in part while visiting the Simons Institute for the Theory of Computing. Shaddin Dughmi was supported by NSF Grant CCF-2432219.

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 (FOCS)*, pages 1203–1223. IEEE, 2023.

Ishaq Aden-Ali, Mikael Møller Høandgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy. Majority-of-three: The simplest optimal learner? In *The Thirty Seventh Annual Conference on Learning Theory*, pages 22–45. PMLR, 2024.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Open problem: Can local regularization learn all multiclass problems? In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 5301–5305. PMLR, 30 Jun–03 Jul 2024a. URL https://proceedings.mlr.press/v247/asilis24b.html.

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Regularization and optimal multiclass learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 260–310. PMLR, 2024b.

Julian Asilis, Mikael Møller Høgsgaard, and Grigoris Velegkas. Understanding aggregations of proper learners in multiclass classification. In *36th International Conference on Algorithmic Learning Theory*, 2025.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: Pac learning and online learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine learning*, 43:211–246, 2001.

Amos Beimel. Secret-sharing schemes: A survey. In *International conference on coding and cryptology*, pages 11–46. Springer, 2011.

George Robert Blakley. Safeguarding cryptographic keys. In *Managing requirements knowledge, international workshop on*, pages 313–313. IEEE Computer Society, 1979.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

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 (FOCS)*, pages 943–955. IEEE, 2022.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Shaddin Dughmi. Pac learning is just bipartite matching (sort of). *arXiv preprint arXiv:2502.00607*, 2025.

Shaddin Dughmi, Yusuf Kalayci, and Grayson York. Is transductive learning equivalent to pac learning? In *36th International Conference on Algorithmic Learning Theory*, 2025.

Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. In *The Twelfth International Conference on Learning Representations*.

Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

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

Mikael Møller Høgsgaard, Kasper Green Larsen, and Markus Engelund Mathiasen. The many faces of optimal weak-to-strong learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Mitsuru Ito, Akira Saito, and Takao Nishizeki. Secret sharing scheme realizing general access structure. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 72(9):56–64, 1989.

Stephan Krenn and Thomas Lorünser. An introduction to secret sharing: A systematic overview and guide for protocol selection. 2023.

Tom M Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, pages 305–310, 1977.

Jean Prost, Antoine Houdard, Andrés Almansa, and Nicolas Papadakis. Learning local regularization for variational image restoration. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 358–370. Springer, 2021.

Benjamin Rubinstein, Peter Bartlett, and J Rubinstein. Shifting, one-inclusion mistake bounds and tight multiclass expected risk bounds. *Advances in Neural Information Processing Systems*, 19, 2006.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir Vapnik. Estimation of dependences based on empirical data: Springer series in statistics (springer series in statistics), 1982.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

Tomas Vaškevičius and Nikita Zhivotovskiy. Suboptimality of constrained least squares and improvements via non-linear predictors. *Bernoulli*, 29(1):473–495, 2023.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Lior Wolf and Yoni Donner. Local regularization for multiclass classification facing significant intraclass variations. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, pages 748–759. Springer, 2008.