Open Problem: Can Local Regularization Learn All Multiclass Problems?

Julian Asilis ASILIS@USC.EDU
Siddartha Devic DEVIC@USC.EDU
Shaddin Dughmi SHADDIN@USC.EDU
Vatsal Sharan VSHARAN@USC.EDU
Shang-Hua Teng SHANGHUA@USC.EDU
University of Southern California

Editors: Shipra Agrawal and Aaron Roth

Abstract

Multiclass classification is the simple generalization of binary classification to arbitrary label sets. Despite its simplicity, it has been remarkably resistant to study: a characterization of multiclass learnability was established only two years ago by Brukhim et al. (2022), and the understanding of optimal learners for multiclass problems remains fairly limited. We ask whether there exists a simple algorithmic template — akin to empirical risk minimization (ERM) for binary classification — which characterizes multiclass learning. Namely, we ask whether *local regularization*, introduced by Asilis et al. (2024), is sufficiently expressive to learn all multiclass problems possible. Towards (negatively) resolving the problem, we propose a hypothesis class which may not be learnable by any such local regularizer.

1. Introduction

Classification refers to supervised learning under the 0-1 loss function ℓ_{0-1}, in which a predicted label incurs loss 0 if equal to the true label and loss 1 otherwise. More precisely, a classification problem is defined by a domain $\mathcal{X}$, label set $\mathcal{Y}$, and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. At training time, a learner receives a sample $S = \big((x_1, y_1), \dots, (x_n, y_n)\big)$ of labeled datapoints $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, each of which are drawn i.i.d. from a distribution D over $\mathcal{X} \times \mathcal{Y}$. The distribution D is required to be *realizable*, i.e., to satisfy the property that some $h^* \in \mathcal{H}$ has $\mathbb{P}_{(x,y)\sim D}[h^*(x) = y] = 1$. A learner $\mathcal{A}$ is a function from samples $S \in (\mathcal{X} \times \mathcal{Y})^{<\omega}$ to hypotheses[1] $\mathcal{A}(S) \in \mathcal{Y}^{\mathcal{X}}$ whose purpose is to emit a hypothesis such that

$$L_D(\mathcal{A}(S)) \leq \epsilon, \quad \text{where} \quad L_D(h) = \mathop{\mathbb{E}}_{(x,y)\sim D}[\ell_{0-1}(h(x), y)].$$

A related quantity is the *empirical risk* on a sample S, i.e., $L_S(h) = \sum_{i=1}^{n} \ell_{0-1}(h(x_i), y_i))$.

The case in which $|\mathcal{Y}| = 2$ is referred as *binary classification*, while larger label sets $\mathcal{Y}$ (perhaps of infinite size) fall under *multiclass classification*. Our open problem concerns the learnability of multiclass classification problems, as well as the sample complexity of learning such problems. We employ Valiant's celebrated PAC learning model (Valiant, 1984).

1. Note that the output of a learner need not be an element of the underlying hypothesis class $\mathcal{H}$.

Definition 1 *A learner $\mathcal{A}$ is a* **PAC learner** *for $\mathcal{H}$ if there exists a sample function $m : (0,1)^2 \to \mathbb{N}$ such that the following condition holds: for any realizable distribution D and $\epsilon, \delta \in (0,1)$, a D-i.i.d. sample S with $|S| \geq m(\epsilon, \delta)$ is such that, with probability at least $1-\delta$ over the choice of S, $L_D(\mathcal{A}(S)) \leq \epsilon$. The minimal such m is the* **sample complexity** *$m_{\mathcal{A}}$ of $\mathcal{A}$. The sample complexity of $\mathcal{H}$ is the pointwise minimal sample complexity attained by any of its learners, i.e., $m_{\mathcal{H}}(\epsilon, \delta) = \min_{\mathcal{A}} m_{\mathcal{A}}(\epsilon, \delta)$.*

In binary classification, it is well-known that a hypothesis class $\mathcal{H}$ is learnable precisely when it has finite VC dimension (Vapnik and Chervonenkis, 1971; Blumer et al., 1989), in which case the simple learning rule of empirical risk minimization (ERM) — which selects any hypothesis $h \in \mathcal{H}$ with perfect sample performance — attains nearly-optimal sample complexity. We ask whether such a simple and performant learning rule can be found for multiclass classification.

Notably, the work of Daniely and Shalev-Shwartz (2014) demonstrates that any such learner must be qualitatively different, and perhaps more complex, than ERM. In particular, they demonstrate the existence of a learnable hypothesis class $\mathcal{H}$ in classification with the property that any learner for $\mathcal{H}$ must emit hypotheses outside of $\mathcal{H}$ on some samples. Such learners are referred to as *improper*, and by the work of Daniely and Shalev-Shwartz (2014), improper learning is necessary for multiclass classification. Thus ERM, which is phrased as an optimization problem over $\mathcal{H}$, must fail on some learnable multiclass problems.

It is natural, then, to ask whether a more powerful version of ERM is sufficiently expressive to learn all multiclass problems. The most obvious candidate is structural risk minimization (SRM), in which one selects a regularizer $\psi : \mathcal{H} \to \mathbb{R}_{\geq 0}$ encoding an "inductive bias" over the hypotheses in $\mathcal{H}$ at the outset, and then given a sample S outputs a hypothesis in $\operatorname{argmin}_{\mathcal{H}} L_S(h) + \psi(h)$. It is easy to see, however, that SRM learners are proper and thus must fail on multiclass problems.

Our open problem concerns the expressive power of *local regularization*, a variant of classical SRM which captures non-uniform inductive biases and, crucially, permits its learners to be improper (Asilis et al., 2024).

Definition 2 *A* **local regularizer** *for a hypothesis class $\mathcal{H}$ is a function $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. A learner $\mathcal{A}$ is said to be* **induced** *by a local regularizer ψ if for all samples S and datapoints $x \in \mathcal{X}$,*

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

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

Intuitively, a local regularizer ψ for a class $\mathcal{H}$ has the flexibility to encode how the "complexity" of hypotheses in $\mathcal{H}$ varies across the domain $\mathcal{X}$. For instance, one can imagine a case in which a hypothesis $h_0 \in \mathcal{H}$ is complex on a region $A \subseteq \mathcal{X}$ and simple on $B \subseteq \mathcal{X}$, while $h_1 \in \mathcal{H}$ has the opposite behavior. In this case, classical SRM may be ill-suited in attempting to compare h_0 and h_1 with a single output. Local regularization, in contrast, is able to express this behavior by setting $\psi(h_0, a) > \psi(h_1, a)$ for $a \in A$ and $\psi(h_0, b) < \psi(h_1, b)$ for $b \in B$. Furthermore, the dependence of local regularizers on the test point permits its induced learners to be improper, as their outputs can "stitch together" the behaviors of hypotheses in $\mathcal{H}$ across different regions of the domain.

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

We ask whether local regularization has the ability to learn all multiclass problems possible, i.e., whether it plays an analogous role in multiclass classification as ERM does in binary classification. Let us briefly describe the significance of the problem:

1. A positive resolution to the problem would provide a simple and flexible algorithmic template for multiclass learning. Existing learners for multiclass problems are primarily mathematical in nature, relying upon orientations of prohibitively large *one-inclusion graphs* (OIGs) associated to the class $\mathcal{H}$ (Daniely and Shalev-Shwartz, 2014; Brukhim et al., 2022). Notably, such OIGs can be infinite when $\mathcal{Y}$ is infinite, in which case the mere existence of favorable orientations employs the axiom of choice, and can be exponentially large in the dimension of $\mathcal{H}$ even when $\mathcal{Y}$ is finite. Local regularization, in contrast, would promise a considerably simpler learning rule.

2. Furthermore, a proof that local regularization can learn all multiclass problems would reveal a fundamental redundancy to OIGs as learners. In particular, OIGs are defined using the underlying class $\mathcal{H}$, the collection of unlabeled datapoints in the training sample S, and the test point x. A positive resolution to the conjecture would reveal that the unlabeled data in S is unnecessary, i.e., a certain redundancy to the information encoded in OIGs for learnable classes. Given the recent volume of work on OIGs, this finding may come as a surprise (see, e.g., Montasser et al. (2022); Aden-Ali et al. (2023b,a); Attias et al. (2023); Charikar and Pabbaraju (2023); Dughmi et al. (2024) and references therein).

3. Conversely, a negative resolution to the problem would suggest that the structure of OIGs is vital for learning, vindicating and offering new perspective on existing work (see point 2.).

One structural result which may be of use in resolving the open problem is the equivalence between learning and sublinear sample compression, as described by David et al. (2016) and recently exploited by Brukhim et al. (2022) to characterize multiclass learnability using the DS dimension.

We close the section with two remarks on local regularization. First, Asilis et al. (2024) studied a strengthened form of local regularization in which the regularizer can additionally use the unlabeled datapoints in S. This form of regularization, termed *unsupervised local regularization*, was shown to learn all multiclass problems possible with near-optimal sample complexity; the open problem asks whether this result can be improved.[2] Second, we note that local regularization may be an important concept beyond the setting of multiclass classification. In linear regression with squared loss, Vaškevičius and Zhivotovskiy (2023) recently demonstrated that a modified Vovk-Azoury-Warmuth (VAW) forecaster surpasses the performance of any learner emitting linear functions (i.e., of any proper learner). Interestingly, this VAW predictor employs precisely a local regularizer favoring functions whose output at the test point has small norm.

2. Candidate counterexample

Towards resolving the open problem, we now describe a learnable hypothesis class which may serve as a counterexample to Open Problem 1. Let $\mathcal{X}$ be an infinite set, say $\mathcal{X} = \mathbb{N}$, and let $\mathcal{Y} = \{*\} \cup 2^{\mathcal{X}}$, where $2^{\mathcal{X}}$ denotes the power set of $\mathcal{X}$. Before defining the hypothesis class $\mathcal{H}_{\triangle}$, let $\mathcal{J} \subseteq (2^{\mathcal{X}})^3$ be the collection of all triples of subsets $(A \subseteq \mathcal{X}, B \subseteq \mathcal{X}, C \subseteq \mathcal{X})$ such that:

2. We also note that this open problem is briefly mentioned in Asilis et al. (2024) using slightly different language, i.e., that of *local size-based regularization*. We opt for a simpler presentation.

1. $|A| = |B| = |C| =: k < \infty$.
2. $|A \cap B| = |A \cap C| = |B \cap C| = k/2$. In particular, $A \cap B \cap C = \emptyset$.

For each $(A, B, C) \in \mathcal{J}$, we will define 8 hypotheses in $\mathcal{H}_{\triangle}$. Namely, all those functions $h \in \mathcal{Y}^{\mathcal{X}}$ satisfying:

1. $h(x) = *$ for all $x \notin A \cup B \cup C$.
2. h is constant on each of $A \cap B$, $A \cap C$, and $A \cap B$.
3. For $x \in A \cap B$, $h(x) \in \{A, B\}$; for $x \in A \cap C$, $h(x) \in \{A, C\}$; and for $x \in B \cap C$, $h(x) \in \{B, C\}$.

Informally, each such h is simply the constant function $\_ \mapsto *$ outside of $A \cup B \cup C$, and on the regions $A \cap B$, $A \cap C$, and $B \cap C$ has the choice of acting as a constant function taking a value in $\{A, B\}$, $\{A, C\}$, or $\{B, C\}$ respectively. Geometrically, one can think of such a function as choosing how to layer the regions A, B, and C ontop of one another. (I.e., imagine sheets of paper over each of A, B, and C bearing the names of their corresponding sets; a function h is equivalent to a choice of layering the sheets of paper with respect to each other. The output of h at input x is the label at x seen from above, i.e., of the topmost sheet of paper.)

We define the class $\mathcal{H}_{\triangle}$ to be the union of all such functions over all triples $(A, B, C) \in \mathcal{J}$. One can see that that $\mathcal{H}_{\triangle}$ is PAC learnable by the learner which defaults to outputting $*$ at $x \in \mathcal{X}$ unless the label $A \ni x$ has been observed in the sample, in which case it outputs A.[3] Informally, consider any function h arising from an $(A, B, C) \in \mathcal{J}$ and a realizable distribution D with marginal $D_{\mathcal{X}}$ over $\mathcal{X}$. Then the previous learner incurs true error 0 once unlabeled datapoints have been seen in each of $A \cap B$, $A \cap C$, and $B \cap C$. Any such region either has negligible mass under $D_{\mathcal{X}}$ or will quickly be observed in a training set.

Let us now argue why we suspect $\mathcal{H}_{\triangle}$ not to be learnable by a local regularizer. First note that ERM learners fail to learn $\mathcal{H}_{\triangle}$; fix a large (finite) set $A \subseteq \mathcal{X}$ and let D be the uniform distribution over $\{(x, A) : x \in A\}$. Then D is a realizable distribution, and consider the output of an ERM learner on a training set $S \sim D$ with $|S| < |A|/2$. With probability $> 1/2$, a test point $(x_{\text{test}}, A) \sim D$ will be such that x_{test} was not seen in S. In this case, there exists a hypothesis $h \in \mathcal{H}_{\triangle}$ with empirical error 0 such that $h(x_{\text{test}}) \neq A$. Namely, an h arising from a triple of sets (A, B, C) such that $x_{\text{test}} \in B$ and S does not contain any unlabeled data in B. An ERM learner is free to predict the label B at x_{test}, thereby incurring constant test error. As the original set A may be chosen to be arbitrarily large, the problem affects ERM learners trained on arbitrarily large training sets S.

Informally, any learner $\mathcal{A}$ equipped with only the information of x_{test} and the empirical errors of all $h \in \mathcal{H}$ would seem to suffer from such a problem on uniform distributions over $\{(x, A) : x \in A\}$. That is, the great amount of symmetry inherent in $\mathcal{H}_{\triangle}$ prevents $\mathcal{A}$ from recognizing that A is the most "natural" prediction for x_{test}, in contrast to any of the sets B which contain x_{test} yet avoid S. In short, it seems that a learner must peek into the training set S in order to learn the geometry of the underlying distribution. Local size-based regularizers, however, cannot do so.

Open Problem 2 *Can the class $\mathcal{H}_{\triangle}$ be learned by a local regularizer? If so, with optimal (or nearly optimal) sample complexity?*

3. If *two* such labels $A \ni x$ and $B \ni x$ have been seen in the training sample, and this information reveals the true label of x (i.e., one label was seen on $A \cap B$), then simply predict this label. If two such labels were seen but this does not reveal the true label of x, arbitrarily choose either of A or B.

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. The one-inclusion graph algorithm is not always optimal. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 72–88. PMLR, 2023a.

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, 2023b.

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*. PMLR, 2024.

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.

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.

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

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

Ofir David, Shay Moran, and Amir Yehudayoff. On statistical learning via the lens of compression. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2792–2800, 2016.

Shaddin Dughmi, Yusuf Kalayci, and Grayson York. Is transductive learning equivalent to pac learning? *arXiv preprint arXiv:2405.05190*, 2024.

Omar Montasser, Steve Hanneke, and Nati Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. *Advances in Neural Information Processing Systems*, 35:37458–37470, 2022.

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

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

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