Regularization and Optimal Multiclass Learning

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

The quintessential learning algorithm of empirical risk minimization (ERM) is known to fail in various settings for which uniform convergence does not characterize learning. Relatedly, the practice of machine learning is rife with considerably richer algorithmic techniques, perhaps the most notable of which is regularization. Nevertheless, no such technique or principle has broken away from the pack to characterize optimal learning in these more general settings.

The purpose of this work is to precisely characterize the role of regularization in perhaps the simplest setting for which ERM fails: multiclass learning with arbitrary label sets. Using *one-inclusion graphs (OIGs)*, we exhibit optimal learning algorithms that dovetail with tried-and-true algorithmic principles: *Occam's Razor* as embodied by *structural risk minimization (SRM)*, the *principle of maximum entropy*, and *Bayesian* inference. We also extract from OIGs a combinatorial sequence we term the *Hall complexity*, which is the first to characterize a problem's *transductive error rate* exactly.

Lastly, we introduce a generalization of OIGs and the *transductive learning* setting to the agnostic case, where we show that optimal orientations of Hamming graphs — judged using nodes' outdegrees minus a system of node-dependent *credits* — characterize optimal learners exactly. We demonstrate that an agnostic version of the Hall complexity again characterizes error rates exactly, and exhibit an optimal learner using maximum entropy programs.

Keywords: Regularization, PAC Learning, Classification, One-inclusion Graphs

1. Introduction

The purpose of a machine learning algorithm is to generalize sample information into an effective model for future prediction. The poster-child of learning algorithms, empirical risk minimization (ERM), proceeds by simply selecting an element of the underlying hypothesis class with best fit to the training data. Despite its success over an impressive array of learning problems, ERM is known to fail catastrophically for many learnable problems, including even mild generalizations of binary classification (Shalev-Shwartz et al., 2010; Alon et al., 2022).

Relatedly, machine learning has a rich algorithmic history of formulating learning as an optimization problem with a more carefully chosen objective function expressing the interplay between sample performance and generalization, bias and variance, or various other learning desiderata. In particular, a heavily used algorithmic principle is that of *regularization*, the most familiar and explicit form of which is *structural risk minimization* (SRM). SRM adds a regularization term to the empirical risk when defining the objective to be minimized over the underlying hypothesis class.

Informally, the regularizer is often meant to encode some notion of *hypothesis complexity*, such that the combined objective function can be viewed as an implementation of the principle of *Occam's Razor*. The practice of machine learning is chock full of such regularization techniques for controlling model capacity, with impressive scientific and societal impact. Nonetheless, no such algorithmic technique or principle has broken away from the pack to characterize optimal learning in these more general settings. In particular, it is not known whether SRM or any particular regularization technique is sufficiently powerful to characterize optimal learning. It is this gap between theory and practice which motivates our work.

The purpose of this paper is to characterize the power of regularization in possibly the simplest setting for which ERM fails: multiclass learning. We are inspired by the result of Daniely and Shalev-Shwartz (2014) that there are learnable problems in multiclass learning that are not learnable by any proper learner (i.e., a learner that always emits an element of the underlying hypothesis class). This impossibility result has direct ramifications to the framework of structural risk minimization. In particular, it implies that any learner witnessed as an optimization problem over the underlying hypothesis class $\mathcal{H}$ is obligated to fail on some multiclass problems. Notably, this includes the standard toolkit of SRM.

This motivates us to address the following fundamental question:

> *What is the minimal augmentation to classical SRM that allows it to learn all (learnable) multiclass problems?*

Our first augmentation, necessary in order to bypass the properness obstruction of Daniely and Shalev-Shwartz (2014), is to grant regularizers access to the test point as input. Formally, such a *local regularizer* is a map $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ for $\mathcal{X}$ the domain set. We demonstrate, however, that local regularizers remain insufficiently expressive to learn certain multiclass problems. We proceed to consider regularizers that furthermore receive as input the sequence of unlabeled datapoints in the training set. We call these regularizers *local unsupervised regularizers* — formally, they are defined as maps $\psi : \mathcal{H} \times \mathcal{X}^n \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ for training sets of size n. Intuitively, the *local* and *unsupervised* modifications to the regularizer allow its induced learner to (1) be improper; and (2) perform an unsupervised pre-training step in which it uses the unlabeled sample data in order to establish local preferences over hypotheses in $\mathcal{H}$.

Contributions. As our primary technical contribution, we demonstrate that local unsupervised regularizers characterize multiclass learnability and, in fact, optimal learning. That is, a multiclass problem is learnable if and only if there exists a local unsupervised regularizer whose SRM learners all learn the problem. Furthermore, in this event there is guaranteed to exist an unsupervised regularizer whose SRM learners are all optimal. Importantly, we also demonstrate that our modification of SRM is *minimal*: disallowing the regularizer access to either the test point or the unlabeled training set leads to the existence of learnable problems which cannot be learned by SRM using either of these weaker regularizers.

Techniques and broad connections. Our characterization of local unsupervised regularization is enabled by the beautiful and insightful *one-inclusion graphs* (OIGs), which have been used to model classification in the transductive setting starting with the work of Haussler et al. (1994). Through Hall's theorem and its generalization to infinite graphs, we derive from OIGs a complexity measure which exactly characterizes the optimal transductive error of a classification problem — we aptly

name this the *Hall complexity*. We then use OIGs to derive learners which follow tried-and-true algorithmic principles: *Occam's razor* as embodied by structural risk minimization (SRM), the *principle of maximum entropy*, and *Bayesian* inference.

We provide two instantiations of an optimal algorithm within the local unsupervised SRM framework. The first is a deterministic learner implicit in the work of Daniely and Shalev-Shwartz (2014), whose regularization function we show to exist abstractly. We derive the second optimal learner, which is randomized, as the dual of a maximum-entropy convex program for orienting the OIG. Its regularization function is the *relative entropy* to a prior over hypotheses obtained through unsupervised learning. In addition to validating the maximum entropy principle in learning, this second algorithm can also be interpreted as a Bayesian learner which samples from a posterior distribution over labels, relative to a prior derived through unsupervised learning.

Most of our contributions extend from the realizable to the agnostic setting, including the characterization of optimal transductive errors through the (agnostic) Hall complexity and the design of an optimal Bayesian learner. To enable this extension, we adapt OIGs to the agnostic case so as to characterize agnostic multiclass learning in the transductive model.

1.1. Related Work

We give a brief overview of related work and defer a full discussion to Appendix F. The importance of multiclass learners beyond the classical paradigm of empirical risk minimization (ERM) was discussed by Shalev-Shwartz et al. (2010), who exhibited a learnable class that is not learnable by any ERM learner. The history of one-inclusion graphs in learning theory commences with the seminal work of Haussler et al. (1994), who employed the transductive learning setting and the OIG algorithm proposed by Alon et al. (1987) to obtain error guarantees for VC classes. Daniely and Shalev-Shwartz (2014) advanced the theory of multiclass learning on several fronts by demonstrating that proper learners fail on learnable hypothesis classes, improving the analysis of transductive error rates, and introducing the DS dimension.

More recently, Brukhim et al. (2022) used OIGs to prove that the DS dimension characterizes multiclass learnability, whereas the Natarajan dimension does not. OIGs also form a key ingredient in the study of learnability for *partial* concept classes (Alon et al., 2022; Kalavasis et al., 2022) as well as recent advances in PAC bounds for various problems (Aden-Ali et al., 2023a), a characterization of learnability for realizable regression (Attias et al., 2023), and the design of optimal learners in the robust setting (Montasser et al., 2022), to name only a few contributions.

Regarding regularization, perhaps most related to our theoretical formalization of regularizers is Hopkins et al. (2022), who consider the task of transforming realizable learners into agnostic learners. In particular, the agnostic learners they produce can be seen as a type of unsupervised regularization, though not described as so in their work. Their learners use unlabeled sample data to restrict focus to a collection of hypotheses F on which they perform ERM. When $F \subseteq \mathcal{H}$, this restriction can be seen as a "hard" regularizer assigning value ∞ to hypotheses in F and zero otherwise. Note, however, that $F \subseteq \mathcal{H}$ only if the original realizable learner is proper. (And, as we have seen, there exist learnable multiclass problems without any proper learners.) Furthermore, they use distinct datasets for regularization and risk minimization, while we do not. Lastly, their work begins with a realizable learner, whereas we are primarily concerned with the design of learners "from scratch."

2. Preliminaries

2.1. Notation

For $m \in \mathbb{N}$, we use $[m]$ to denote the set $\{1, \ldots, m\}$. For a predicate P, we let $[P]$ denote the Iverson bracket of P, i.e., $[P] = 1$ if P is true and 0 if P is false. When Z is a set, we use $Z^{<\omega}$ to denote the collection of all finite sequences in Z, i.e., $Z^{<\omega} = \bigcup_{i=1}^{\infty} Z^i$. When Z is finite, we use Δ_Z to denote the set of all probability measures over Z,

$$\Delta_Z = \left\{ P \colon Z \to \mathbb{R}_{\geq 0} : \sum_Z P(z) = 1 \right\}.$$

For a tuple $S = (z_1, \ldots, z_n)$, we let S_{-i} denote S with its ith entry removed, as in $S_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the greatest integer weakly less than x.

2.2. Learning Theory

We first recall the standard toolkit of supervised learning. A learning problem is determined by a **domain** $\mathcal{X}$, a **label set** $\mathcal{Y}$, and a **hypothesis class** $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. We will refer to any function $\mathcal{X} \to \mathcal{Y}$ as a **hypothesis** or **predictor**. Learning also requires a **loss function** $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ to quantify a predictor's quality, perhaps the most fundamental of which is the 0-1 loss: $\ell_{0-1}(y, y') = [y \neq y']$.

We refer to learning problems with the 0-1 loss as **multiclass classification** problems when $|\mathcal{Y}| > 2$ and **binary classification** problems when $|\mathcal{Y}| = 2$. A **labeled datapoint** is a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and an **unlabeled datapoint** is an element of $\mathcal{X}$. We will occasionally refer to an (un)labeled datapoint merely as a datapoint when clear from context. A **training set** S is a finite tuple of labeled datapoints, i.e., $S \in (\mathcal{X} \times \mathcal{Y})^{<\omega}$. We may refer to them as *training samples* or simply *samples*.

A **learner** A is a (possibly randomized) function from training sets to hypotheses, i.e., $A : (\mathcal{X} \times \mathcal{Y})^{<\omega} \to \mathcal{Y}^{\mathcal{X}}$. The **true error**, or simply **error**, incurred by a hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ with respect to a distribution D over $\mathcal{X} \times \mathcal{Y}$ is the average loss its predictions incur on labeled datapoints drawn from D, i.e., $L_D(h) = \mathbb{E}_{(x,y)\sim D}\, \ell(h(x), y)$. The **empirical error** incurred by a hypothesis h on a sample $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ is the average loss it suffers over datapoints in S, as in

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

Definition 1 *A learner A is an* **empirical risk minimizer** *(ERM) with respect to $\mathcal{H}$ if for all samples S, we have that $A(S) \in \operatorname{argmin}_{\mathcal{H}} L_S(h)$.*

A related technique is structural risk minimization (SRM), which amounts to empirical risk minimization along with an inductive bias favoring certain hypotheses in $\mathcal{H}$ over others.

Definition 2 *A* **regularizer** *for a hypothesis class $\mathcal{H}$ is a function $\psi : \mathcal{H} \to \mathbb{R}_{\geq 0}$. A learner A is a* **structural risk minimizer** *(SRM) with respect to $\mathcal{H}$ if there exists a regularizer ψ for $\mathcal{H}$ such that, for all samples S, we have that*

$$A(S) \in \operatorname*{argmin}_{\mathcal{H}} L_S(h) + \psi(h).$$

A learning problem is also defined by a criterion for success, the most celebrated of which is certainly Valiant's PAC learning framework (Valiant, 1984).

Definition 3 *Let $\mathbb{D}$ be a collection of probability measures over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. A learner A is a* **PAC learner for $\mathcal{H}$ with respect to $\mathbb{D}$** *if there exists a* **sample function** $m : (0,1)^2 \to \mathbb{N}$ *such that the following condition holds: for any $D \in \mathbb{D}$ 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 and any internal randomness in A,*

$$L_D(A(S)) \leq \inf_{\mathcal{H}} L_D(h) + \epsilon.$$

In Definition 3, when $\mathbb{D}$ consists of all probability measures over $\mathcal{X} \times \mathcal{Y}$, one says that A is an **agnostic PAC learner** for $\mathcal{H}$. When $\mathbb{D}$ consists of all probability measure D such that $L_D(h) = 0$ for some $h \in \mathcal{H}$, one says that A is a **realizable PAC learner** for $\mathcal{H}$. Hereafter, we will suppress dependence on the family $\mathbb{D}$ and trust that it be established clearly in the surrounding context.

Definition 4 *The* **sample complexity** *of a learner A with respect to a hypothesis class $\mathcal{H}$, $m_{\mathrm{PAC},A} : (0,1)^2 \to \mathbb{N}$, is the minimal sample function it attains as a learner for $\mathcal{H}$. The* **sample complexity** *of a class $\mathcal{H}$ is the pointwise minimal sample complexity attained by any of its learners, i.e., $m_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) = \min_A m_{\mathrm{PAC},A}(\epsilon, \delta)$.*

It will be useful to note that ERM and SRM learners are *proper* learners, as we now define.

Definition 5 *A learner A is said to be* **proper with respect to $\mathcal{H}$** *if the guarantee $A(S) \in \mathcal{H}$ holds for all samples S.*

We remark briefly that we take a rather hands-off approach to measure-theoretic details throughout, adopting standard assumptions from e.g. Shalev-Shwartz and Ben-David (2014). See, e.g., Blumer et al. (1989), (Pollard, 2012, Appendix C) for a more precise treatment.

2.3. Error: High-probability, Expected, and Transductive

The PAC learning framework demands high-probability guarantees of its learners, as presented in Definition 3. Though certainly the most standard framework for assessing learners, it is by no means the only one. Perhaps the most straightforward measure of a learner's quality is simply its expected error. This is the *prediction model* of learning proposed by Haussler et al. (1994). Here, we require learners to attain vanishingly small expected error when trained on increasingly large samples, and endow such learners with corresponding sample complexities $m_{\mathrm{Exp}}(\epsilon)$. A third notion of error is that of the transductive error, to be described precisely shortly, which captures a learner's performance in the *transductive model* of learning. This is a somewhat more adversarial version of the prediction model, and is in fact also described and used by Haussler et al. (1994).[1] Once again, this framework requires vanishing transductive error of its learners, thereby equipping them with sample complexities $m_{\mathrm{Trans}}(\epsilon)$.

A natural question to ask of these learning criteria is as follows: How do the sample complexities they induce on classes $\mathcal{H}$ differ? Does there exist a class $\mathcal{H}$ whose sample complexities $m_{\mathrm{Exp}}(\epsilon)$

1. See also prior discussion of transductive learning by Vapnik and Chervonenkis (1974) and Vapnik (1982).

and $m_{\mathrm{PAC}}(\epsilon, \delta)$ scale considerably differently with ϵ (or δ) than its transductive sample complexity $m_{\mathrm{Trans}}(\epsilon)$? For the case of realizable learning with bounded loss, the three frameworks turn out to be essentially equivalent. Notably, this places our study of one-inclusion graphs — which are tailored to minimizing transductive error — on firmer theoretical footing.

Definition 6 *The* **transductive learning** *setting is that in which the following steps take place:*

1. *An adversary chooses a collection of n unlabeled datapoints $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$, along with a hypothesis $h \in \mathcal{H}$.*
2. *The unlabeled data S is revealed to the learner.*
3. *One datapoint x_i is selected uniformly at random from S. The remaining datapoints S_{-i} and their labels under h are displayed to the learner. That is, the learner receives $(x_j, h(x_j))_{j \neq i}$.*
4. *The learner is prompted to predict the label of x_i, i.e., $h(x_i)$.*

We refer to x_i as the **test datapoint***, and the remaining S_{-i} as the* **training datapoints***. The* **transductive error** *incurred by a learner A on the instance (S, h) is its expected error over the uniformly random choice of x_i. That is,*

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

where $A(S_{-i}, h)$ denotes the output of A on the sample of datapoints in S_{-i} labeled by h.

Intuitively, transductive error can be thought of as a fine-grained form of expected error that demands favorable performance on each individual sample S, and that furthermore "hard-codes" a uniform distribution over the datapoints of S. In particular, note the lack of an underlying distribution D in the transductive setting.

The **transductive error rate** of a learner A or class $\mathcal{H}$ is defined respectively as

$$\epsilon_{A,\mathcal{H}}(n) = \max_{S \in \mathcal{X}^n, h \in \mathcal{H}} L^{\mathrm{Trans}}_{S,h}(A), \quad \text{and} \quad \epsilon_{\mathcal{H}}(n) = \min_{A} \epsilon_{A,\mathcal{H}}(n).$$

We show that the sample complexities of learning in the PAC, expected, and transductive settings differ by at most logarithmic factors in the realizable case. We emphasize that the content of this claim is neither novel nor particularly profound. Nevertheless, we believe that the community may benefit from a singular, organized treatment of the topic, which — to our knowledge — does not at present appear in the literature. See Appendix A for further detail, along with Dughmi et al. (2024).

Proposition 7 (Informal Proposition 32) *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and a loss function taking values in $[0, 1]$. Let $m_{\mathrm{PAC},\mathcal{H}}$, $m_{\mathrm{Exp},\mathcal{H}}$, and $m_{\mathrm{Trans},\mathcal{H}}$ be the sample complexities of learning $\mathcal{H}$ in the realizable PAC, expected, and transductive settings, respectively. Then the following inequalities hold for all $\epsilon, \delta \in (0, 1)$ and the constant $e \approx 2.718$.*

1. $m_{\mathrm{Exp},\mathcal{H}}(\epsilon + \delta) \leq m_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) \leq O\left(m_{\mathrm{Exp},\mathcal{H}}(\epsilon/2) \cdot \log(1/\epsilon)\right)$.
2. $m_{\mathrm{Exp},\mathcal{H}}(\epsilon) \leq m_{\mathrm{Trans},\mathcal{H}}(\epsilon) \leq m_{\mathrm{Exp},\mathcal{H}}(\epsilon/e)$.

So far, we have restricted attention to realizable learning. We describe agnostic analogues of transductive learning and of Proposition 7 in Section 5.

3. One-inclusion Graphs and the Hall Complexity

One-inclusion graphs (OIGs) are powerful combinatorial objects that capture the structure of realizable learning under the 0-1 loss. They are particularly well-suited for analyzing transductive error, as defined in Definition 6. We begin the section by briefly reviewing OIGs, and refer the reader to Appendix B for a more complete (and self-contained) treatment. We then introduce the *Hall complexity* derived from OIGs, which we show exactly characterizes a problem's optimal transductive error rate. Throughout the section, we remain in the setting of realizable multiclass classification.

Definition 8 *Let $\mathcal{X}$ be a domain, $\mathcal{Y}$ a label set, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. The* **one-inclusion graph** *of $\mathcal{H}$ with respect to $S \in \mathcal{X}^n$, denoted $G(\mathcal{H}|_S) = (V, E)$, is the following hypergraph:*

- $V = \mathcal{H}|_S$, *and*
- $E = \bigcup_{i=1}^{n} \mathcal{H}|_{S_{-i}}$, *where $e = h \in \mathcal{H}|_{S_{-i}}$ is incident to all $g \in \mathcal{H}|_S$ such that $g|_{S_{-i}} = h$.*

Intuitively, each edge e in $G(\mathcal{H}|_S)$ corresponds to a labeled training sample and an unlabeled test point x_{test}. A learner chooses a manner of completing e into a fully labeled dataset by predicting a label for x_{test}; this is precisely a choice of node incident to e (i.e., an *orientation* of e). In this view, the transductive error incurred on an instance (S, h) equals the number of edges incident to $h|_S \in V$ which were *not* oriented towards $h|_S$ — that is, the number of times the learner "should have" completed the label of x_{test} in accordance with the ground truth $h|_S$ but did not. This is simply the outdegree of the node $h|_S$. Let us formalize these concepts in the following definition.

Definition 9 *An* **orientation** *of a hypergraph $G = (V, E)$ is a function $f : E \to V$ such that $f(e)$ is incident to e for all $e \in E$. The* **outdegree** *of $v \in V$ in orientation f is the number of edges e incident to v with $f(e) \neq v$. The* **indegree** *of v in f is the number of edges e with $f(e) = v$. We say G is α-***orientable** *if it can be oriented so that the in-degree of each vertex is at least α. Similarly, G is α-***coorientable** *if it can be oriented so that the out-degree of each vertex is at most α. We refer to orientations satisfying these conditions as α-***orientations** *and α-***coorientations***, respectively.*

We will also consider randomized orientations of OIGs, in which case we naturally extend Definition 9 so that a randomized α-orientation is one which satisfies expected in-degree requirements, and likewise for coorientations. The following lemma formalizes the equivalence between learners attaining low transductive error and orientations of the OIG with low outdegree.

Lemma 10 *Let $\mathcal{A}$ be a transductive learner for $\mathcal{H}$. The following conditions are equivalent:*

1. *$\mathcal{A}$ incurs transductive error at most ϵ on all samples of size n;*
2. *For each $h \in \mathcal{H}$ and $S \in \mathcal{X}^n$, $\mathcal{A}$ induces an $(\epsilon \cdot n)$-coorientation on $G(\mathcal{H}|_S)$; and*
3. *For each $h \in \mathcal{H}$ and $S \in \mathcal{X}^n$, $\mathcal{A}$ induces an $((1-\epsilon) \cdot n)$-orientation on $G(\mathcal{H}|_S)$.*

Proof sketch Conditions (2.) and (3.) are equivalent as $G(\mathcal{H}|_S)$ is regular; each node has degree $|S| = n$. Conditions (1.) and (2.) are equivalent by our previous reasoning, i.e., the outdegree of a node $h \in G(\mathcal{H}|_S)$ (with respect to the orientation induced by $\mathcal{A}$) counts the number of errors made by A when h is the underlying transductive instance. See Lemma 42 for the formal proof. ■

We now define the *Hall complexity*, which characterizes transductive error rates exactly. When $G = (V, E)$ is an undirected hypergraph and $U \subseteq V$, we let $E[U] \subseteq E$ denote the collection of edges with at least one incident node in U.

Definition 11 *The* Hall density *of a graph* $G = (V, E)$ *is* $\mathsf{Hall}(G) = \inf_{\substack{U \subseteq V, \\ |U| < \infty}} \frac{|E[U]|}{|U|}$.

Definition 12 *The* **Hall complexity** *of a hypothesis class* $\mathcal{H}$ *is the function* $\pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ *defined*

$$\pi_{\mathcal{H}}(n) = \max_{S \in \mathcal{X}^n} n - \mathsf{Hall}\big(G(\mathcal{H}|_S)\big).$$

Proposition 13 (Informal Proposition 46) *The Hall complexity of a class* $\mathcal{H}$ *exactly characterizes the optimal transductive error rate of learning* $\mathcal{H}$*. That is,* $\epsilon_{\mathcal{H}}(n) = \frac{\pi_{\mathcal{H}}(n)}{n}$ *for all* $n \in \mathbb{N}$.[2]

The proof is essentially an application of Hall's theorem to the (bipartite) edge-vertex incidence graph of the one-inclusion graph. For a discussion of the Hall complexity's relation to other dimensions and sequences related to transductive error, we refer the reader to Remark 47.

4. Structural Risk Minimization

In this section, we establish our main results in Theorems 20 and 21: realizable multiclass learnability is characterized by generalized regularizers granted access to the test point and unlabeled training points as input. That is, a multiclass problem is learnable precisely when it can be learned by such a regularizer, in which case one such regularizer is guaranteed to produce optimal learners. We term these regularizers *local unsupervised regularizers*, and support our central theorem using results from Sections 4.1, 4.2, and 4.3, which establish both sufficiency and — in a precise sense — minimality of this generalized regularizer. Our proofs also help to illustrate the role of unsupervised methods in multiclass learning. Notably, our insights rely crucially on the machinery of one-inclusion graphs. Throughout the section, we remain in the setting of realizable classification.

4.1. Impossibility Results

We begin by collecting impossibility results concerning learning techniques which learn all multiclass classification problems possible. The first result, due to Daniely and Shalev-Shwartz (2014), establishes that proper learners are in general insufficient for multiclass classification problems.

Proposition 14 (Daniely and Shalev-Shwartz (2014, Lemma 2)) *There exists a PAC learnable hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$*, for infinite label set* $\mathcal{Y}$*, that is not learnable by any proper learner.*

Notably, this eliminates the possibility that ERM or SRM learners are sufficiently expressive to learn all multiclass problems possible. To see why a proper learner may be insufficient, imagine a setting in which a class $\mathcal{H}$ contains hypotheses whose "complexities" vary considerably over the domain $\mathcal{X}$. For instance, $h_0 \in \mathcal{H}$ acts "simply" on $X_0 \subseteq \mathcal{X}$ and with complexity on $X_1 \subseteq \mathcal{X}$, while $h_1 \in \mathcal{H}$ does the opposite. Then, intuitively, a regularizer ψ would like to alternate between favoring the functions h_0 and h_1, depending upon the location being queried. See Figure 1.

2. We note briefly that the Hall complexity exactly characterizes the error rate of learning with arbitrary learners, which are permitted to use internal randomness. Introducing a floor in the definition of the Hall density begets a version of the Hall complexity which characterizes error rates with respect to deterministic learners.

This reasoning naturally gives rise to the notion of a *local regularizer* that is non-uniform with respect to $\mathcal{X}$.

Definition 15 *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 A is a* **local structural risk minimizer** *with respect to $\mathcal{H}$ if there exists a local regularizer ψ for $\mathcal{H}$ such that, for all samples S and $x \in \mathcal{X}$,*

$$A(S)(x) \in \left\{ h(x) : h \in \operatorname*{argmin}_{\mathcal{H}} L_S(h) + \psi(h, x) \right\}.$$

In this case, we say that A is a learner **induced** *by ψ. We say that ψ* **learns** *the class $\mathcal{H}$ if all learners it induces are PAC learners for $\mathcal{H}$.*

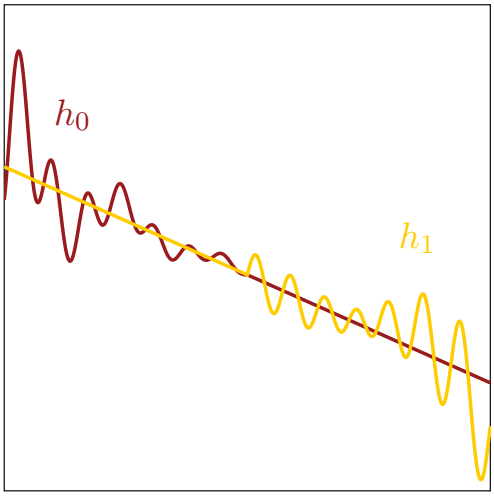

Figure 1 : *Hypotheses h_0 and h_1, depicted in red and yellow respectively. A local regularizer may favor the simplicity of h_0 on test points drawn from the right region of the domain, and the simplicity of h_1 on test points drawn from the left region.*

We note that local regularizers (or similar ideas) have been previously considered in computer vision (Wolf and Donner, 2008; Prost et al., 2021) and in the learning theory community (Bottou and Vapnik, 1992). Nonetheless, we demonstrate that this augmention is insufficient: even local regularizers are required to fail on some learnable classification problems.

Proposition 16 *There exists a PAC learnable class $\mathcal{H}$ which no local regularizer can learn.*

We defer the proof to Appendix D.1, and note that it employs the *first Cantor class* of Daniely and Shalev-Shwartz (2014). Having already equipped regularizers with the information of the test point, there remains only one additional source of information with which to empower regularizers: the sample S itself. If we were to grant local regularizers full access to S, it is straightforward to see that they could induce any learner, rendering the characterization meaningless (see Appendix C.1). We should ask, then, what is the weakest summary statistic of S which can be supplied to regularizers in order to increase their power? Perhaps the most simple is $|S|$, the cardinality of S. Equipping regularizers with this information is both a powerful tool in the practice of machine learning[3] and, notably, would allow them to learn the class $\mathcal{H}_\infty$ used in the proof of Proposition 16.

Definition 17 *A* **local size-based regularizer** *for a hypothesis class $\mathcal{H}$ is a function $\psi : \mathcal{H} \times \mathbb{N} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. A learner A is said to be* **induced** *by a local size-based regularizer ψ if for all samples S and datapoints $x \in \mathcal{X}$,*

$$A(S)(x) \in \left\{ h(x) : h \in \operatorname*{argmin}_{\mathcal{H}} L_S(h) + \psi(h, |S|, x) \right\}.$$

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

3. There is evidence in both theory and practice which suggests that sample size plays a crucial role in calibrating regularizers; see e.g. the sample-size dependent regularizer from Shalev-Shwartz et al. (2010) or the SVM regularization in Wainer and Cawley (2017); Shalev-Shwartz and Srebro (2008).

We now articulate a conjecture: local size-based regularizers are insufficient for classification.

Conjecture 18 *There exists a PAC learnable class $\mathcal{H}$ that is not learned by any local size-based regularizer.*

In Appendix C.2 we provide a learnable hypothesis class which we suspect cannot be learned by any local size-based regularizer, towards justifying the conjecture. Furthermore, we note that a negative resolution to the conjecture would somewhat undermine the structure of OIGs themselves. That is, there would exist learners for all multiclass problems which use strictly less information than OIGs (i.e., $|S|$ rather than all its unlabeled data). Given the volume of work on OIGs and their insights for learning, such an outcome could be considered surprising.

4.2. Deterministic Learning with Acyclic Orientations

Given the collection of impossibility results above and our suspicion about the insufficiency of size-based regularizers, we turn to providing a regularizer which utilizes not only the cardinality of $|S|$, but indeed the entire unlabeled sample set. For a sample $S = ((x_i, y_i))_{i\in[n]}$, we let $S_{\mathcal{X}}$ denote the sequence of unlabeled datapoints in S, i.e, $S_{\mathcal{X}} = (x_i)_{i\in[n]}$.

Definition 19 *A* **local unsupervised regularizer***, or simply* unsupervised regularizer*, for a hypothesis class $\mathcal{H}$ is a function $\psi : \mathcal{H} \times \mathcal{X}^{<\omega} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. A learner A is a* **local unsupervised structural risk minimizer** *with respect to $\mathcal{H}$ if there exists a local unsupervised regularizer ψ such that the following guarantee holds for all samples S and datapoints $x \in \mathcal{X}$:*

$$A(S)(x) \in \left\{ h(x) : h \in \operatorname*{argmin}_{\mathcal{H}} L_S(h) + \psi(h, S_{\mathcal{X}}, x) \right\}.$$

In this case, we say that A is a learner **induced** *by ψ. We say that ψ* **learns** *the class $\mathcal{H}$ if all learners it induces are PAC learners for $\mathcal{H}$.*

The central result of this section is that local unsupervised regularizers are indeed sufficiently expressive to optimally learn all multiclass problems with the 0-1 loss. We note that the proof is constructive when $\mathcal{Y}$ is finite and employs a compactness argument for infinite $\mathcal{Y}$.

Theorem 20 *Let $\mathcal{Y}$ be a finite or countable label set and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. Then $\mathcal{H}$ has an unsupervised local regularizer ψ whose induced learners all attain optimal transductive error up to a constant factor of 2.*

Proof sketch The full proof is given in Appendix D.2. First suppose that $\mathcal{Y}$ is finite. It suffices to demonstrate that for each $S \in \mathcal{X}^n$ there exists an acyclic orientation of $G(\mathcal{H}|_S)$ that is optimal to within a factor of 2. (From the acyclic orientation, one can topologically sort $G(\mathcal{H}|_S)$ and define an unsupervised local regularizer that is decreasing on layers.) The acyclic orientation, implicit in the work of Daniely and Shalev-Shwartz (2014), arises from the following k-*core* algorithm: repeatedly remove the vertex of lowest degree from $G(\mathcal{H}|_S)$ and place it in the last layer of a topological ordering. The outdegree of any vertex in this ordering is precisely its degree in the undirected (sub)graph of $G(\mathcal{H}|_S)$ before it was removed. This is bounded above by the maximum subgraph density of $G(\mathcal{H}|_S)$, which in turns bounds the error of its best learner up to a factor of 2. The case of infinite $\mathcal{Y}$ requires a compactness argument. ■

Note that local unsupervised regularizers are *minimal* in the following sense: If we were to disallow the test point x from a regularizer's input, its induced learners would be proper and thus fail on learnable problems by Proposition 14. If we restricted access of the regularizer to $S_{\mathcal{X}}$, its corresponding learners would again fail on learnable problems by Proposition 16.

4.3. Randomized Learning with Maximum Entropy Distributions

The deterministic learner from the previous section serves as a tool for implementing principled tie-breaking between hypotheses attaining zero empirical risk. In this section, we propose a tie-breaking rule for *randomized learners* based upon the maximum entropy principle. This rule also has the property of being Bayesian in nature: it corresponds to learning a Bayesian prior using the unlabeled samples $S_{\mathcal{X}}$ and test datapoint x_{test}, which is then updated to a suitable posterior distribution over $\mathcal{H}|_{S_{\mathcal{X}} \cup \{x_{\text{test}}\}}$ once the labeled data is revealed to the learner. We derive this learner from the solution of a certain maximum-entropy convex program. Hereafter, we let $S^+_{\mathcal{X}}$ denote the collection of unlabeled training datapoints alongside the test datapoint, i.e., $S^+_{\mathcal{X}} = S_{\mathcal{X}} \cup \{x_{\text{test}}\}$.

We restrict attention to finite but arbitrarily large label sets $\mathcal{Y}$ in this section, as a consequence of certain measurability issues. However, none of our results are parameterized by the size of the underlying label set, suggesting that they may admit extension to infinite label spaces (perhaps via compactness arguments, as in (Asilis et al., 2024; Brukhim et al., 2022; Daniely and Shalev-Shwartz, 2014)).

Theorem 21 *There exists an optimal randomized learner which can be summarized as follows.*

1. *Upon receiving the unlabeled datapoints $S^+_{\mathcal{X}} = (x_1, \ldots, x_n)$, including the test point, use a convex program to compute an optimal randomized orientation of $G(\mathcal{H}|_{S^+_{\mathcal{X}}})$ with maximum entropy, then derive a prior distribution ρ over $\mathcal{H}|_{S^+_{\mathcal{X}}}$ by normalizing the dual variables.*
2. *Given the index i of the test point, and labels y_j for all datapoints $x_{j \neq i}$, apply a Bayes update to ρ in order obtain a posterior ρ'. This posterior corresponds to restricting the prior to hypotheses consistent with the provided labels, and rescaling accordingly.*
3. *Sample a hypothesis h from ρ' and output $h(x_i)$ as the predicted label of x_i.*

Proof sketch The full proof is given in Appendix D.3. First, we draw an equivalence between the OIG $G(\mathcal{H}|_{S^+_{\mathcal{X}}})$ and a certain bipartite representation of the OIG, G_{BP}. The graph G_{BP} has the property that assignments from its left-hand side are equivalent to orientations in $G(\mathcal{H}|_{S^+_{\mathcal{X}}})$. We define a convex program akin to Singh and Vishnoi (2014) which (randomly) assigns the left-hand side of G_{BP} so as to guarantee optimal error while maximizing entropy subject to this constraint. Due to the fact that dual variables enforce local decisions in convex matching programs, we are able to back out a Bayesian update step from the learner while preserving optimal error guarantees. ■

By Theorem 21, we have that when the true labels for the training datapoints are revealed, the randomized optimal (maximum entropy) learner is Bayesian in that it updates its posterior over hypotheses to have minimum relative entropy to the prior ρ, constrained on outputting hypotheses consistent with the training data. We argue that this is a local unsupervised SRM in a natural generalized sense. In particular, we expand the space of hypotheses to include distributions over hypotheses, and have our regularizer assign a complexity to each such randomized hypothesis.

Definition 22 *Let $\psi : \Delta_{\mathcal{H}} \times \mathcal{X}^{<\omega} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a local unsupervised regularizer for randomized hypotheses.*[4] *A (randomized) learner A is said to be induced by the regularizer ψ if for all samples S and datapoints $x \in \mathcal{X}$,*

$$A(S)(x) = h(x) \text{ where } h \sim D \text{ for } D \in \operatorname*{argmin}_{\Delta_{\mathcal{H}}} \mathop{\mathbb{E}}_{h \sim D} \left[L_S(h)\right] + \psi(D, S, x).$$

Corollary 23 *The learner from Theorem 21 can be realized as a local unsupervised regularizer for randomized hypotheses.*

We defer the proof to Appendix D.4 and note that the regularizer of Corollay 23 depends upon the relative entropy (i.e., KL divergence) between the distribution D and the Bayesian prior ρ. In addition to being an SRM in the generalized sense just described, our learner can also be interpreted as an instantiation of the maximum entropy principle. In particular, if the prior ρ were uniform, then indeed our learner would sample from the maximum entropy distribution over hypotheses consistent with the training data. More generally, sampling from the distribution which hues most closely to the prior subject to the provided labels, as measured by relative entropy, is the natural generalization of the maximum entropy principle to incorporate prior knowledge. That is, our learner deviates as little as possible from the prior subject to consistency with the provided labels.

Discussion. A priori, the value of a randomized learner may be unclear given that the previous section derives a deterministic one. The pioneering work of Jaynes (1957) states the *principle of maximum entropy*: one should choose prior probabilities consistent with available information so as to maximize the entropy of the system. Jaynes (1957) further states: "it is *unreasonable* to assign zero probability to any event unless our data really rules out the case." Coupled with the fact that the previous deterministic learners are highly contingent upon the choice of optimal orientation, it is natural to ask for a (randomized) learner making fewer arbitrary choices. Yet another advantage of the randomized learner from this section is that it generalizes to the agnostic setting, as we will see shortly, whereas the deterministic strategy of Section 4.2 does not appear to.

5. Agnostic Learning

Our discussion of learning and one-inclusion graphs has thus far pertained to the realizable case. Indeed, the structure of one-inclusion graphs and the transductive learning setting depends crucially upon guarantees provided by the realizability assumption.

We devote this section to the generalization of the one-inclusion graph and its accompanying insights to the agnostic case. In particular, we define an agnostic version of the OIG and demonstrate that our previous results holds for the agnostic OIG, with the exception of the deterministic learner from Theorem 20. A more complete discussion of the agnostic OIG, including precise statements of results and their proofs, is deferred to Appendix E.

Definition 24 (Informal Definition 57) Transductive learning in the agnostic case *is defined as in the realizable case (Definition 6), except that the adversary can arbitrarily label their datapoints.*

4. Technically, for any $(S_{\mathcal{X}}, x_{\text{test}}) \in \mathcal{X}^{<\omega} \times \mathcal{X}$, we need only consider distributions over $\mathcal{H}|_{S'}$ for $S' = S_{\mathcal{X}} \cup \{x_{\text{test}}\}$, i.e., elements of $\Delta_{\mathcal{H}|_{S'}}$. As $\mathcal{Y}$ is taken to be finite in this section, $\mathcal{H}|_{S'} \subseteq \mathcal{Y}^{S'}$ will always be finite.

To compensate for the increased difficulty of the agnostic case, and in accordance with the PAC definition of agnostic learning, an agnostic transductive learner A is only judged relative to the best-in-class performance across $\mathcal{H}$:

$$L_S^{\text{Trans}}(A) = \frac{1}{n}\sum_{i\in[n]} \ell(A(S_{-i})(x_i), y_i) - \inf_{h\in\mathcal{H}} \frac{1}{n}\sum_{i\in[n]} \ell(h(x_i), y_i).$$

Unfortunately, the sample complexities of transductive and PAC learning are not known to be as closely related in the agnostic case as they are in the realizable case. With a simple of simple use of Markov's inequality and a repetition argument, one can show that agnostic PAC sample complexities exceed their transductive counterparts by at most a factor of $1/\epsilon$. This is, of course, not a lower order factor, but it may be a loose bound. See Dughmi et al. (2024) for further discussion of the relationship between agnostic PAC and transductive learning.

Definition 25 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. The* **agnostic one-inclusion graph** *of $\mathcal{H}$ with respect to $S \in \mathcal{X}^n$, denoted $G_{\text{Ag}}(\mathcal{H}|_S) = (V, E)$, is the hypergraph with:*

- *$V = \mathcal{Y}^n$, one node for each possible labeling of the n datapoints, and*
- *$E = \bigcup_{i=1}^n \mathcal{Y}^n|_{S_{-i}}$, where $e \in \mathcal{Y}^n|_{S_{-i}}$ is incident to each $v \in \mathcal{Y}^n$ such that $v|_{S_{-i}} = e$.*

The agnostic OIG is sometimes referred to as the *Hamming graph* (Brouwer and Haemers, 2011). We note that Long (1998) use a similarly expanded OIG to study binary classification under distribution shift, but to our knowledge no previous work has fully expanded upon the idea to analyze learning in the agnostic case (see Appendix F for further discussion).

Analagously to the realizable case, agnostic learners are in close correspondence with orientations of the agnostic OIG (c.f. Lemma 10). A crucial difference, however, is that each vertex v in an agnostic OIG is endowed with a number of *Hamming credits* reflecting its distance from the underlying class $\mathcal{H}$. Informally, the Hamming credits are subtracted from the outdegree of v in any orientation of the OIG, so that learners (equivalently, orientations) are judged only upon the "excess outdegree" they induce on vertices. Semantically, the Hamming credits reflect the fact that learners for the agnostic case need only compete with the performance of the best hypothesis in $\mathcal{H}$. In Definition 66 we use Hamming credits to define a suitable generalization of the Hall complexity which retains the exact transductive error characterization of its realizable counterpart.

Proposition 26 (Informal Proposition 67) *The agnostic Hall complexity of a class $\mathcal{H}$ exactly characterizes the optimal agnostic transductive error rate of learning $\mathcal{H}$.*

Lastly, we demonstrate that the randomized learner of Section 4.3 generalizes to the agnostic case. Informally, the convex program used to produce the maximum entropy learner in Theorem 21 applies to the agnostic case with nearly identical reasoning, as it is robust to the addition of nodes to the OIG and to non-uniform out-degree requirements (i.e., to the presence of Hamming credits).

Proposition 27 (Informal Proposition 70) *The Bayesian randomized learner from Theorem 21 and its associated randomized regularizer from Corollary 23 can be extended to the agnostic case.*

6. Conclusion

In pursuit of an algorithmic template for multiclass classification, we study the role of regularization in multiclass learning. We first observe that classical regularizers $\psi : \mathcal{H} \to \mathbb{R}_{\geq 0}$ are insufficient to learn multiclass problems owing to the work of Daniely and Shalev-Shwartz (2014), and extend this impossibility result to the more powerful *local regularizers* which are given access to the test datapoint. We then consider *unsupervised local regularizers*, which are regularizers granted access to both the unlabeled test datapoint and the collection of all unlabeled training points. By exploiting the connection between unsupervised local regularizers and acyclic orientations of one-inclusion graphs (OIGs), we provide deterministic transductive learners that are nearly optimal (up to a factor of 2) for all multiclass problems in the realizable case. We then demonstrate an optimal randomized transductive learner for both the realizable and agnostic settings by way of a certain maximum entropy program, and show it to be an unsupervised local SRM as well as a pre-trained Bayesian learner. As part of our efforts, we also generalize the one-inclusion graph to the agnostic case and define the *Hall complexity* associated to a class $\mathcal{H}$, which is the first to provide an exact combinatorial characterization of transductive error rates. Future work includes resolving our conjecture that local size-based regularizers, which are of intermediate power between local regularizers and unsupervised local regularizers, are insufficient to learn multiclass problems. It would also be of interest to design optimal (or nearly optimal) multiclass learners which are computationally efficient, and to study the role of regularization (local, unsupervised, and otherwise) beyond classification.

Acknowledgments

Julian Asilis was supported by the USC Viterbi School of Engineering Graduate School Fellowship and by the Simons Foundation. Siddartha Devic was supported by the Department of Defense through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program. Shaddin Dughmi was supported by NSF Grant CCF-2009060. Vatsal Sharan was supported by NSF CAREER Award CCF-2239265 and an Amazon Research Award. Shang-Hua Teng was supported in part by NSF Grant CCF-2308744 and the Simons Investigator Award from the Simons Foundation. We thank Li Han for discussions that inspired the questions considered in this work. We thank Jim Ferry for pointing to the connection between Proposition 32 and Stirling numbers of the second kind. We thank Yusuf Hakan Kalayci for finding a typo in the definition of the agnostic Hall complexity from a previous draft. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the sponsors such as the NSF.

References

I. Aden-Ali, Y. Cherapanamjeri, A. Shetty, and N. Zhivotovskiy. Optimal pac bounds without uniform convergence. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1203–1223, 2023a.

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, 2023b.

Noga Alon, David Haussler, and Emo Welzl. Partitioning and geometric embedding of range spaces of finite vapnik-chervonenkis dimension. In *Proceedings of the third annual symposium on Computational geometry*, pages 331–340, 1987.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of pac learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671. IEEE, 2022.

Julian Asilis, Siddartha Devic, Shaddin Dughmi, Vatsal Sharan, and Shang-Hua Teng. Transductive sample complexities are compact. *arXiv preprint arXiv:2402.10360*, 2024.

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

Léon Bottou and Vladimir Vapnik. Local learning algorithms. *Neural computation*, 4(6):888–900, 1992.

Andries E Brouwer and Willem H Haemers. *Spectra of graphs*. Springer Science & Business Media, 2011.

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.

Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

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

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

AJ Dobson. A note on stirling numbers of the second kind. *Journal of Combinatorial Theory*, 5(2): 212–214, 1968.

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

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pretraining help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. *arXiv preprint arXiv:2303.01566*, 2023.

P. Hall. On representatives of subsets. *Journal of the London Mathematical Society*, s1-10(1):26–30, 1935. doi: https://doi.org/10.1112/jlms/s1-10.37.26.

Marshall Hall Jr. Distinct representatives of subsets. *Bulletin of the American Mathematical Society*, 54(10):922–926, 1948.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

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

Max Hopkins, Daniel M Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Conference on Learning Theory*, pages 3015–3069. PMLR, 2022.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Alkis Kalavasis, Grigoris Velegkas, and Amin Karbasi. Multiclass learnability beyond the pac framework: Universal rates and partial concept classes. *Advances in Neural Information Processing Systems*, 35:20809–20822, 2022.

Jason D Lee, Ben Recht, Nathan Srebro, Joel Tropp, and Russ R Salakhutdinov. Practical large-scale optimization for max-norm regularization. *Advances in neural information processing systems*, 23, 2010.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Philip M Long. The complexity of learning according to two models of a drifting environment. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 116–125, 1998.

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.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

Saharon Rosset, Grzegorz Swirszcz, Nathan Srebro, and Ji Zhu. L1 regularization in infinite dimensional feature spaces. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings 20*, pages 544–558. Springer, 2007.

Benjamin IP Rubinstein, Peter L Bartlett, and J Hyam Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59, 2009.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language models help solve downstream tasks. *International Conference on Learning Representations*, 2021.

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

Shai Shalev-Shwartz and Nathan Srebro. Svm optimization: inverse dependence on training set size. In *Proceedings of the 25th international conference on Machine learning*, pages 928–935, 2008.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Mohit Singh and Nisheeth K Vishnoi. Entropy, optimization and counting. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 50–59, 2014.

Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. *Advances in neural information processing systems*, 21, 2008.

A. Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

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.

Jacques Wainer and Gavin C. Cawley. Empirical evaluation of resampling procedures for optimising SVM hyperparameters. *J. Mach. Learn. Res.*, 18:15:1–15:35, 2017.

Manfred K Warmuth. The optimal pac algorithm. In *International Conference on Computational Learning Theory*, pages 641–642. Springer, 2004.

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.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *ACM/JMS Journal of Data Science*, 1(2):1–51, 2024.

Contents

1 Introduction **1**
1.1 Related Work . . . 3

2 Preliminaries **4**
2.1 Notation . . . 4
2.2 Learning Theory . . . 4
2.3 Error: High-probability, Expected, and Transductive . . . 5

3 One-inclusion Graphs and the Hall Complexity **7**

4 Structural Risk Minimization **8**
4.1 Impossibility Results . . . 8
4.2 Deterministic Learning with Acyclic Orientations . . . 10
4.3 Randomized Learning with Maximum Entropy Distributions . . . 11

5 Agnostic Learning **12**

6 Conclusion **14**

A Equivalence of Errors **20**
A.1 A Simple Equivalence . . . 21

B One-inclusion Graphs and the Hall Complexity **24**
B.1 One-inclusion Graphs (with a Bipartite Perspective) . . . 24
B.2 The Hall Complexity . . . 27

C Structural Risk Minimization: Supplement **30**
C.1 Regularizer Using S Can Induce Any Learner . . . 30
C.2 Support for Conjecture 18 . . . 30

D Omitted proofs **31**
D.1 Proof of Proposition 16 . . . 31
D.2 Proof of Theorem 20 . . . 32
D.3 Proof of Theorem 21 . . . 35
D.4 Proof and Discussion of Corollary 23 . . . 40

E Extension to Agnostic Learning **41**
E.1 Transductive Learning . . . 42
E.2 Agnostic One-inclusion Graphs . . . 43
E.3 Equivalence of Errors and Orienting the Agnostic OIG . . . 46

F Related Work **49**

Appendix Organization

Appendices A and B are self-contained sections expanding upon claims already mentioned in the main body: the equivalence between the PAC, transductive, and expected learning frameworks (Appendix A), and the Hall complexity of one-inclusion graphs (Appendix B). Appendix C briefly expands upon two points mentioned in Section 4.1: the triviality of local regularizers permitted access to all of S, and our conjecture concerning the insufficiency of local size-based regularizers.

Appendix D is devoted to proofs which were omitted from the main text. Appendix E extends the machinery of one-inclusion graphs and Hall complexity to the agnostic case in a self-contained manner, as advertised briefly in Section 5. We conclude in Appendix F with an expanded discussion of related work.

Appendix A. Equivalence of Errors

The PAC learning framework demands high-probability guarantees of its learners, as presented in Definition 3. Though certainly the most standard framework for assessing learners, it is by no means the only one. Perhaps the most straightforward measure of a learner's quality is simply its expected error. This is the *prediction model* of learning proposed by Haussler et al. (1994). Here, we require learners to attain vanishingly small expected errors when trained on increasingly large samples, and endow such learners with corresponding sample complexities $m_{\mathrm{Exp}}(\epsilon)$. A third notion of error is that of the transductive error, to be described precisely shortly, which captures a learner's performance in the *transductive model* of learning. This is a somewhat more adversarial version of the prediction model, and is in fact also described and used by Haussler et al. (1994). Once again, this framework requires vanishing transductive error of its learners, thereby equipping them with sample complexities $m_{\mathrm{Trans}}(\epsilon)$.

The most natural question to ask of these criteria for learning is whether they determine the same collection of learnable classes. That is, does there exist any hypothesis class $\mathcal{H}$ learnable under one such framework but not another? In the case of realizable learning with a bounded loss function — as we primarily consider — it can be shown with little difficulty that the frameworks coincide in this sense. A central concern of learning theory, however, is not merely whether a given hypothesis class $\mathcal{H}$ can be learned, but moreover the sample complexity with which it can be learned. Consequently, one should ask of the three learning criteria: How do the sample complexities they induce on classes $\mathcal{H}$ differ? Does there exist a class $\mathcal{H}$ whose sample complexities $m_{\mathrm{Exp}}(\epsilon)$ and $m_{\mathrm{PAC}}(\epsilon, \epsilon)$ scale considerably differently with ϵ than its transductive sample complexity $m_{\mathrm{Trans}}(\epsilon)$? For a given learning problem, how do guarantees at the level of one error correspond to guarantees for the others, if at all?

The purpose of this section is to review these concepts and study the conditions under which one can favorably transform a learner with guarantees in one such error regime into a learner with guarantees in another (i.e., with only a modest effect on sample complexity). This is a topic which has received relatively little attention from the learning theory community, and which we believe would benefit from a clear collection of existing results. Furthermore, it will place our study of one-inclusion graphs — which are tailored to minimizing transductive error — on firmer theoretical footing. In particular, we will show for the case of realizable learning with bounded loss that the three learning frameworks are essentially equivalent, by providing modest bounds (at most logarithmic) on the extent to which their sample complexities may differ.

A.1. A Simple Equivalence

Throughout the section, we fix an arbitrary domain $\mathcal{X}$, label set $\mathcal{Y}$, hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, and a bounded loss function ℓ, which we normalize to take values in $[0,1]$. We also direct our attention to learning in the realizable case, and let $\mathbb{D}_{\mathcal{H}}$ denote the collection of all $\mathcal{H}$-realizable distributions over $\mathcal{X} \times \mathcal{Y}$. That is, $\mathbb{D}_{\mathcal{H}}$ consists of those distributions D for which some $h \in \mathcal{H}$ satisfies $L_D(h) = 0$.

Definition 28 *The* sample complexity $m_{\mathrm{Exp},A} : (0,1) \to \mathbb{N}$ *of a learner A in the expected error framework is the function mapping ϵ to the minimal m for which the following condition holds:*

$$(\forall D \in \mathbb{D}_{\mathcal{H}})(\forall m' \geq m) \mathop{\mathbb{E}}_{S \sim D^{m'}} L_D\big(A(S)\big) \leq \epsilon.$$

That is, $m_{\mathrm{Exp},A}$ tracks the minimal number of samples required by A to attain a desired level of expected error, with respect to any $D \in \mathbb{D}_{\mathcal{H}}$. Definition 28 forms an appropriate definition at the level of individual learners, but our interest ultimately lies in proving claims at the level of hypothesis classes. That is, we require a notion of sample complexity for a class $\mathcal{H}$.

Definition 29 *The* sample complexity *of a hypothesis class $\mathcal{H}$ in the expected error framework is the optimal sample complexity attained by its learners, i.e.,*

$$m_{\mathrm{Exp},\mathcal{H}}(\epsilon) = \min_{A} m_{\mathrm{Exp},A}(\epsilon),$$

where A ranges over all learners.

Definition 30 *The* **transductive learning** *setting is that in which the following steps take place:*

1. *An adversary chooses a collection of n unlabeled datapoints $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$, along with a hypothesis $h \in \mathcal{H}$.*
2. *The datapoints S are displayed to the learner.*
3. *One datapoint x_i is selected uniformly at random from S. The remaining datapoints*

   $$S_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

   and their labels under h are displayed to the learner. That is, the learner receives the data of $\big(x_j, h(x_j)\big)_{x_j \in S_{-i}}$.
4. *The learner is prompted to predict the label of x_i, i.e., $h(x_i)$.*

We refer to x_i as the **test datapoint***, and the remaining S_{-i} as the* **training datapoints***. The* **transductive error** *incurred by a learner A on the instance (S,h) is its expected error over the uniformly random choice of x_i. That is,*

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

where we use $A(S_{-i}, h)$ to denote the output of A on the sample consisting of datapoints in S_{-i} labeled by h. The **transductive error rate** *incurred by A is the function $\epsilon_{A,\mathcal{H}} : \mathbb{N} \to \mathbb{R}$ defined by*

$$\epsilon_{A,\mathcal{H}}(n) = \max_{S \in \mathcal{X}^n,\, h \in \mathcal{H}} L^{\mathrm{Trans}}_{S,h}(A).$$

Intuitively, transductive error can be thought of as a fine-grained form of expected error that demands favorable performance on each individual sample S, and that furthermore "hard-codes" a uniform distribution over the datapoints of S. In particular, note the lack of an underlying distribution D in the transductive setting.

Definition 31 *The* transductive sample complexity $m_{\mathrm{Trans},A} : (0,1) \to \mathbb{N}$ *of a learner A is the function mapping δ to the minimal m for which $\epsilon_{A,\mathcal{H}}(m') < \delta$ for all $m' \geq m$. That is,*

$$m_{\mathrm{Trans},A}(\delta) = \min\{m \in \mathbb{N} \,:\, \epsilon_{A,\mathcal{H}}(m') < \delta,\ \forall m' \geq m\}.$$

We are now equipped to present the central claim of the section: these sample complexities differ by at most logarithmic factors. We emphasize again that the content of the claim is neither novel nor particularly profound. Nevertheless, we believe that the community may benefit from a singular, organized treatment of the topic, which — to our knowledge — does not at present appear in the literature.

Proposition 32 *Fix an arbitrary domain $\mathcal{X}$, label set $\mathcal{Y}$, and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Use a loss function taking values in $[0,1]$. Then the following inequalities hold for all $\epsilon, \delta \in (0,1)$ and the constant $e \approx 2.718$.*

1. $m_{\mathrm{Exp},\mathcal{H}}(\epsilon + \delta) \leq m_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) \leq O\left(m_{\mathrm{Exp},\mathcal{H}}(\epsilon/2) \cdot \log(1/\epsilon)\right)$.
2. $m_{\mathrm{Exp},\mathcal{H}}(\epsilon) \leq m_{\mathrm{Trans},\mathcal{H}}(\epsilon) \leq m_{\mathrm{Exp},\mathcal{H}}(\epsilon/e)$.

Proof The first inequality of claim (1.) follows directly from the fact that the loss function is bounded above by 1. That is, any learner attaining error $\leq \epsilon$ with probability $\geq (1-\delta)$ on sample of size n automatically incurs an expected error of at most

$$\epsilon \cdot (1-\delta) + 1 \cdot \delta \leq \epsilon + \delta.$$

For the second inequality in (1.), let A be a learner attaining expected error $\leq \epsilon/2$ on samples of size n, from which we would like to extract a high-probability guarantee. The key observation is that A can be boosted to attain expected error $\leq \epsilon$ with probability $\geq 1 - \epsilon$ using only an additional factor of $O(\log(1/\epsilon))$ many examples. In particular, a sample of size $n \cdot O(\log(1/\epsilon))$ can be divided into units of size n, half of which are used to train A and produce candidate hypotheses $h_1, \ldots, h_\ell$, and half of which are used as a single validation set to select the best such h_i, as described in Daniely and Shalev-Shwartz (2014).

The first inequality of claim (2.) follows from the standard leave-one-out argument of Haussler et al. (1994). In particular, if A is a learner incurring transductive error $\leq \epsilon$, then the same can be said of its expected error. For a sample $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$, recall that S_{-i} denotes the sample consisting of all labeled examples in S other than (x_i, y_i).

$$\begin{aligned}
\mathop{\mathbb{E}}_{S \sim D^m} L_D\big(A(S)\big) &= \mathop{\mathbb{E}}_{\substack{S \sim D^m \\ (x,y) \sim D}} \ell\big(A(S)(x), y\big) \\
&= \mathop{\mathbb{E}}_{S \sim D^{m+1}} \ell\big(A(S_{-(m+1)})(x_{m+1}), y_{m+1}\big) \\
&= \mathop{\mathbb{E}}_{S \sim D^{m+1}} \mathop{\mathbb{E}}_{i \in_R [m+1]} \ell\big(A(S_{-i})(x_i), y_i\big) \\
&\leq \sup_S \mathop{\mathbb{E}}_{i \in_R [m+1]} \ell\big(A(S_{-i})(x_i), y_i\big) \\
&= \epsilon_{A,\mathcal{H}}(m+1).
\end{aligned}$$

The second inequality in (2.) is claimed without proof by Daniely and Shalev-Shwartz (2014); in the interest of completeness, we provide a proof in Lemma 34. ■

We note briefly that Theorem 2.2 of Aden-Ali et al. (2023a) provides tighter bounds than Proposition 32 for transforming a learner with optimal transductive error into one with PAC error guarantees. Notably, however, the theorem is phrased only for finite label sets, in contrast to the generality of Proposition 32. The ability to quantify over arbitrary label sets is crucial for our purposes, as it is primarily over infinite label sets that the theory of multiclass classification departs from binary classification (e.g., uniform convergence fails to characterize learnability, ERM learners fail for learnable problems, etc.). As such, we are best served with the slightly looser but considerably more general statement of Proposition 32.

Remark 33 *It was recently shown in Aden-Ali et al. (2023b) that one-inclusion graphs, which attain optimal transductive error, do not always provide optimal high-probability guarantees. We note that this is compatible with Proposition 32, which quantifies over all learners for a given class $\mathcal{H}$ and does* not *claim that a learner attaining optimal error in one regime need do so for the others as well. Proposition 32 instead demonstrates that the levels of performance attained by the (possibly distinct) optimal learners for the three notions of error are comparable. This suffices to justify a focus on any of the three errors — in our case, transductive — as the sample complexities enjoyed by optimal learners in the other regimes will be comparable. (And in fact, the proof of Proposition 32 provides a simple recipe for transforming optimal transductive learners into near-optimal high-probability learners.)*

Lemma 34 *Fix an arbitrary domain $\mathcal{X}$, label set $\mathcal{Y}$, and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Then for all $\epsilon \in (0,1)$, $m_{\mathrm{Trans},\mathcal{H}}(\epsilon) \leq m_{\mathrm{Exp},\mathcal{H}}(\epsilon/e)$.*

Proof Fix ϵ, set $n = m_{\mathrm{Exp},\mathcal{H}}(\epsilon)$, and let A be a learner attaining expected error $\leq \epsilon$ on samples of size n. We will extract from A a learning attaining transductive error at most $e \cdot \epsilon$, completing the proof. First, an intermediate result.

> **Lemma 35** *For each $n \in \mathbb{N}$, there exists an $m_n \in \mathbb{N}$ such that m_n independent draws from a uniform distribution over n items results in seeing exactly $n-1$ unique elements with probability at least $\frac{1}{e}$.*
>
> **Proof** Let $\left\{ {a \atop b} \right\}$ denote Stirling numbers of the second kind. Then for m many draws from the uniform distribution over n elements, the probability of seeing exactly $n-i$ unique elements is
>
> $$\frac{1}{n^m} \left\{ {m \atop n-i} \right\} \frac{n!}{i!}.$$
>
> We refer to this event as E_i, when n and m are clear from context. For the moment let us shift our perspective so that m is the variable of interest and our asymptotics are taken with respect to m. That is, for each given m, can we find an n such that m uniform draws result in exactly $n-1$ unique elements with constant probability?
>
> Given m, let k_m be a value maximizing $\left\{ {m \atop i} \right\}$ over $i \in [m]$. Then take $n_m = k_m + 1$. For each $i \in [m]$, we have that $P(E_i) \propto \left\{ {m \atop n_m - i} \right\} \cdot \frac{1}{i!}$. By definition of n_m, the first

factor in this product is maximized for $i = 1$. Therefore, we can lower bound $P(E_i)$ by imagining instead that $P(E_i) \propto \frac{1}{i!}$. And in this case, clearly, $P(E_i) \geq \frac{1}{e}$.

Lastly, by Dobson (1968), each $n \in \mathbb{N}$ appears as k_m for some m, completing the argument. ■

We will now design a randomized learner B that attains transductive error $\leq e \cdot \epsilon$ on samples of size n. In particular, B acts as follows upon receiving sample $S = \big((x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\big)$ and test datapoint x_n: it randomly generates a uniform sample of size m_n from the uniform distribution over S, call it S', and returns $A(S')$.

To analyze the transductive error of B, fix a transductive learning instance (S, h). Let $S = (x_1, \ldots, x_n)$, and have D denote the uniform distribution over $(x_1, h(x_1)), \ldots, (x_n, h(x_n))$. Call a sample drawn from D *good* if it contains exactly $n - 1$ unique elements. The crucial observation is as follows: the probability that B errs, averaged over a uniformly random test point $x_i \in S$ and the randomness internal to B, equals precisely the probability that A errs conditioned on receiving good samples of size m_n. By Lemma 35, the latter quantity is at most $e \cdot \epsilon$. This completes the argument. ■

Appendix B. One-inclusion Graphs and the Hall Complexity

One-inclusion graphs (OIGs) are powerful combinatorial objects that capture the structure of realizable learning under the 0-1 loss. They are particularly well-suited for analyzing transductive error, as defined in Definition 30. In particular, it was demonstrated in Daniely and Shalev-Shwartz (2014) that for a given hypothesis class $\mathcal{H}$, a combinatorial sequence $\mu_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ associated to the one-inclusion graphs of $\mathcal{H}$ provides a constant factor approximation to the optimal transductive error of its learners. The central result of this section is the introduction of a new sequence associated to the OIGs of $\mathcal{H}$, which we term the *Hall complexity*, that characterizes optimal transductive error exactly. To this end, we recall the appropriate definitions concerning one-inclusion graphs in Section B.1, and present the Hall complexity in Section B.2.

Throughout the section we restrict focus to realizable learning under the 0-1 loss, over arbitrary domain and label sets $\mathcal{X}, \mathcal{Y}$.

B.1. One-inclusion Graphs (with a Bipartite Perspective)

Recall the basic structure of transductive learning: a learner is presented with n unlabeled datapoints $S = (x_1, \ldots, x_n)$, one such datapoint i is removed uniformly at random from S, and the learner is asked to guess $h(i)$ from the data of $h|_{S_{-i}}$, for some $h \in \mathcal{H}$. Finally, the learner is judged on its average performance over the randomness of $i \in [n]$, with respect to the 0-1 loss function. Note that, equipped with this loss function, the transductive error incurred by learner A on the instance (S, h) is

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

Now let us take the perspective of a transductive learner for the moment, and imagine that we have just been given the data of the n datapoints $S = (x_1, \ldots, x_n)$, including the yet-to-be-selected test point. Our objective is to minimize the worst-case transductive error we incur over

any possible "ground truth" $h \in \mathcal{H}|_S$. First, a simple observation: upon observing $h|_{S_{-i}}$, we ought to output $g(x_i)$ for $g \in \mathcal{H}|_S$ a function such that $g|_{S_{-i}} = h|_{S_{-i}}$. Otherwise, it is guaranteed that $g(x_i) \neq h(x_i)$ and thus that we incur a loss of 1, the maximal possible loss. That is, in realizable learning with the 0-1 loss, any sensible learner ought to be an ERM. Hereafter, we will only consider (transductive) learners obeying this mild property.

We now introduce some notation. Let us represent each $g \in \mathcal{H}|_S$ as a *fully labeled dataset* $\big((x_1, g(x_1)), \ldots, (x_n, g(x_n))\big)$, i.e., by its graph. Similarly, we can represent each $h \in \mathcal{H}|_{S_{-i}}$ as a *partially labeled dataset* $\big((x_1, h(x_1)), \ldots, (x_i, ?), \ldots, (x_n, h(x_n))\big)$, i.e., by its graph augmented by a "?" accompanying the test datapoint x_i. For $g \in \mathcal{H}|_S$ a fully labeled dataset and $h \in \mathcal{H}|_{S_{-i}}$ a partially labeled dataset, we say that g *completes* h (or is *compatible* with h) when they agree on S_{-i}.

In this light, by our previous reasoning, the task of a transductive learner is to complete each partially labeled dataset into a fully labeled dataset from among $\mathcal{H}|_S$. Note also the following observations:

1. Each fully labeled dataset (i.e., ground truth $g \in \mathcal{H}|_S$) is compatible with exactly n partially labeled datasets, each corresponding to one location for the "?".

2. Upon making a choice of fully labeled dataset for each partially labeled dataset, the transductive error incurred on the ground truth $g \in \mathcal{H}|_S$ is proportional to the number of compatible partially labeled datasets that are *not* assigned to g. Equivalently, n minus the number of partially labeled datasets that are assigned to g.

These insights are perhaps best expressed in the form of a bipartite graph $G_{\mathrm{BP}} = (\mathcal{A}, \mathcal{B}, E)$. Let the partially labeled datasets and fully labeled datasets form $\mathcal{A}$ and $\mathcal{B}$, respectively. We then have an edge $(u, v) \in E$ precisely when v is a fully labeled dataset completing u, as depicted in Figure 2. The following claims follow immediately from our previous reasoning: a transductive learner on S is precisely a choice of assignment in G_{BP} (mapping each $u \in \mathcal{A}$ to an incident $v \in \mathcal{B}$), and the worst-case transductive error it incurs over $h \in \mathcal{H}|_S$ is determined by the minimal indegree of a node in $\mathcal{B}$ under this assignment.

In short, transductive learning devolves to finding assignments in bipartite graphs that maximize minimal indegrees in $\mathcal{B}$. With only a slight change in perspective, we will arrive at one-inclusion graphs. Namely, given $G_{\mathrm{BP}} = (\mathcal{A}, \mathcal{B}, E)$, consider the hypergraph G whose vertex set is $\mathcal{B}$ and edge set is $\mathcal{A}$, such that $a \in \mathcal{A}$ is incident to precisely those vertices in G (as an edge) with which it is incident in G (as a node). Simply put, view each $a \in \mathcal{A}$ as an edge rather than a node!

This is formalized by the following definition.

Definition 36 *Let $\mathcal{X}$ be a domain, $\mathcal{Y}$ a label set, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. The* **one-inclusion graph** *of $\mathcal{H}$ with respect to $S \in \mathcal{X}^n$, denoted $G(\mathcal{H}|_S)$, is the hypergraph defined by the following vertex and edge sets:*

- *$V = \mathcal{H}|_S$, and*
- *$E = \bigcup_{i=1}^{n} \mathcal{H}|_{S_{-i}}$, where $e = h \in \mathcal{H}|_{S_{-i}}$ is incident to all $g \in \mathcal{H}|_S$ such that $g|_{S_{-i}} = h$.*

We will sometimes write such an edge e as $e_{(h,i)}$, with $h \in \mathcal{H}|_S$ and $i \in [n]$. Under this representation, $e_{(h,i)}$ is incident to all nodes $g \in \mathcal{H}|_S$ such that $g|_{S_{-i}} = h|_{S_{-i}}$.

Let us now formally define the bipartite view of one-inclusion graphs, G_{BP}, which we have informally discussed. We will return frequently to this view of one-inclusion graphs throughout the work.

Definition 37 *Let $\mathcal{X}$ be a domain, $\mathcal{Y}$ a label set, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. The* **bipartite view of the one-inclusion graph** *of $\mathcal{H}$ with respect to $S \in \mathcal{X}^n$, denoted $G_{\mathrm{BP}}(\mathcal{H}|_S)$, is defined as follows. Let $G(\mathcal{H}|_S) = (V, E)$ be the usual one-inclusion graph of $\mathcal{H}$ with respect to S. Then $G_{\mathrm{BP}}(\mathcal{H}|_S) = (L, R, E')$ is the bipartite graph in which $L = E$, $R = V$, and $(e, v) \in E'$ precisely when e is incident to v in $G(\mathcal{H}|_S)$. In other words, $G_{\mathrm{BP}}(\mathcal{H}|_S)$ is precisely the edge-vertex incidence graph of $G(\mathcal{H}|_S)$. We may denote it simply by G_{BP} when $\mathcal{H}$ and S are clear from context.*

Remark 38 *There are various non-equivalent definitions of one-inclusion graphs in the literature. Ours is equivalent to that of (Brukhim et al., 2022, Definition 9), and allows for hyperedges to have size 1 (i.e., to be self-loops), which crucially results in each node having degree exactly $n = |S|$. Originally, OIGs were defined by Haussler et al. (1994) with the requirement that edges have size at least 2 (see also Alon et al. (1987)). For our purposes, this has the unfavorable consequence of permitting nodes to have different degrees, which prevents us from establishing an equivalence between maximizing nodes' indegrees and minimizing nodes' outdegrees (to be seen shortly). In addition to the usual graph-theoretic definitions of OIGs, it will be often be useful to retain the bipartite interpretation of Figure 2 as a supplementary perspective. In particular, it will form the basis for analyzing the Hall complexity in a few moments, and for generalizing OIGs to the agnostic case in Appendix E.*

Recall now the notion of an orientation of a hypergraph.

Definition 39 *An* **orientation** *of a hypergraph $G = (V, E)$ is a function $f : E \to V$ such that $f(e)$ is incident to e for all $e \in E$. The* **outdegree** *of $v \in V$ in orientation f is the number of edges e incident to v with $f(e) \neq v$. Similarly, the* **indegree** *of v in f is the number of edges e with $f(e) = v$.*

The following is immediate from our previous reasoning.

Lemma 40 *There is a one-to-one correspondence between (deterministic) transductive learners for $\mathcal{H}$ and orientations of $G(\mathcal{H}|_S)$ for all $S \in \mathcal{X}^{<\omega}$. Furthermore, a transductive learner A incurs transductive error $\leq \epsilon$ on the instance (h, S) if and only if $h|_S$ has outdegree $\leq \epsilon \cdot |S|$ in the graph $G(\mathcal{H}|_S)$ oriented by A.*

Proof Fix a learner A and $S \in \mathcal{X}^n$. The action of A on partially labeled datasets induces an orientation on $G(\mathcal{H}|_S)$, such that each hyperedge $e = \big((x_1, y_1), \ldots, (x_i, ?), \ldots, (x_n, y_n)\big)$ is oriented toward $\big((x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\big)$ for y_i the output of A on input e. We write $e \to h$ if e is oriented towards h and $e \not\to h$ otherwise. Now, for any $h \in \mathcal{H}|_S$ we have:

$$\begin{aligned} L^{\mathrm{Trans}}_{S,h}(A) &= \frac{1}{n} \sum_{i \in [n]} \big[A(S_{-i}, h)(x_i) \neq h(x_i)\big] \\ &= \frac{1}{n} \sum_{i \in [n]} [e_{(h,i)} \not\to h] \\ &= \frac{1}{n} \cdot \mathrm{outDeg}(h). \end{aligned}$$

■

The previous lemma justifies a focus on orientations that minimize nodes' outdegrees. For one-inclusion graphs, in which every node has undirected degree exactly $|S|$, this amounts precisely to maximizing nodes' indegrees.

Definition 41 *Let G be an undirected hypergraph. We say G is α-**orientable** if it can be oriented so that all its vertices' indegrees are at least α. G is α-**coorientable** if it can be oriented so that all its vertices' outdegrees are at most α. We will refer to orientations satisfying these conditions as α-**orientations** and α-**coorientations**, respectively.*

We may sometimes use *randomized* orientations to describe randomized learners. In these cases, we naturally extend Definition 41 so that a randomized α-orientation is one which satisfies *expected* in-degree requirements, and likewise for coorientations.

Lemma 42 *Let A be a transductive learner for $\mathcal{H}$. The following conditions are equivalent.*

1. *A incurs transductive error at most ϵ on all samples of size n.*
2. *For each $h \in \mathcal{H}$ and $S \in \mathcal{X}^n$, A induces an $(\epsilon \cdot n)$-coorientation on $G(\mathcal{H}|_S)$.*
3. *For each $h \in \mathcal{H}$ and $S \in \mathcal{X}^n$, A induces an $((1-\epsilon) \cdot n)$-orientation on $G(\mathcal{H}|_S)$.*

Proof Conditions (2.) and (3.) are patently equivalent, as any node in the undirected graph $G(\mathcal{H}|_S)$ has degree exactly $n = |S|$, one for each point in S that can be omitted. The equivalence between (1.) and (2.) follows immediately from Lemma 40. ■

B.2. The Hall Complexity

Given a framework for learning and a hypothesis class $\mathcal{H}$, perhaps the most pressing question is: How quickly can $\mathcal{H}$ be learned, if at all? For transductive learning, progress was first made on the issue by Haussler et al. (1994), who demonstrated an upperbound on the transductive error rate of learning $\mathcal{H}$ based upon the maximum subgraph density of its one-inclusion graphs. (See Remark 47 for further detail). Subsequently, Daniely and Shalev-Shwartz (2014) introduced a sequence characterizing optimal transductive errors up to constant factors. We now introduce a combinatorial sequence that characterizes the errors of optimal transductive learners exactly.

Definition 43 *Let $G = (V, E)$ be an undirected hypergraph. For a set of nodes $U \subseteq V$, let $E[U] \subseteq E$ denote the collection of edges with at least one incident node in U. The **Hall density** of G is*

$$\mathsf{Hall}(G) = \inf_{\substack{U \subseteq V, \\ |U| < \infty}} \frac{|E[U]|}{|U|}.$$

Proposition 44 *Let G be an undirected hypergraph in which each node has finite degree. Then $\mathsf{Hall}(G)$ is the supremum of all α for which G is α-orientable.*

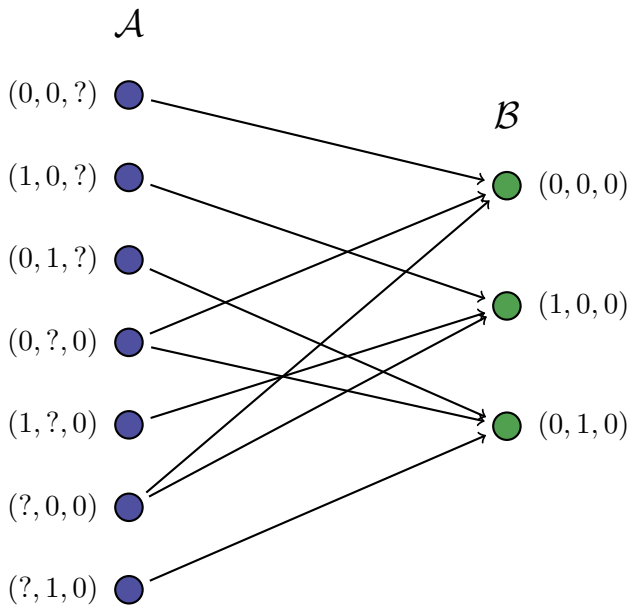

Figure 2: Bipartite graph G_{BP} representing transductive learning on a traning set S of three datapoints, for which $\mathcal{H}$ can express the binary strings $(0,0,0)$, $(1,0,0)$, and $(0,1,0)$. We fix an ordering of the unlabeld data in S and represent each fully or partially labeled dataset using only its labels. Note that each fully labeled dataset has degree exactly 3 = $|S|$. An optimal transductive learner amounts precisely to an assignment of G (i.e., choice of incident edge for each vertex in $\mathcal{A}$) that maximizes the minimal indegree in $\mathcal{B}$. Here, the best indegree that can be attained is 2.

Proof An orientation of $G = (V, E)$ is precisely a (possibly randomized) assignment $E \to V$ in which each edge is assigned to an incident node. In order for there to exist such an assignment in which each $v \in V$ receives α in-degree, it is clearly necessary that for each finite $U \subseteq V$, $|E[U]| \geq \alpha \cdot |U|$. We now demonstrate sufficiency through cases.

Case 1. When $\mathsf{Hall}(G) \in \mathbb{N}$, and G is finite, sufficiency follows from the classical statement of Hall's theorem Hall (1935) combined with a standard splitting argument. (I.e., creating $\mathsf{Hall}(G)$ many copies of each right-hand side node in the bipartite view of G.)

Case 2. When $\mathsf{Hall}(G) \in \mathbb{N}$ and G is infinite, sufficiency follows from the generalization of Hall's theorem to infinite collections of finite sets, as described in (Hall Jr, 1948, Theorem 1).[5] In particular, each set in the set system corresponds to the collection of edges incident to a given node in G, and is thus finite as a consequence of each node's degree being finite.

Case 3. When $\mathsf{Hall}(G) = \frac{p}{q} \in \mathbb{Q}$, let G_{BP} be the bipartite edge-incidence graph of G, i.e., $G = (L, R, E')$ with $L = E$, $R = V$, and $(e, v) \in E'$ when e is incident to v in G. (See Definition 37.) Then create a graph G'_{BP} resembling G_{BP} but with q copies of each left-hand side

5. We remark briefly that (Hall Jr, 1948, Theorem 1) appears to be phrased for countable collections of finite sets, but that its proof nevertheless holds for arbitrary collections of finite sets obeying Hall's condition.

vertex, each of which has the same incidence relations as in G_{BP}. By design of q, the resulting graph has an assignment such that each right-hand side node receives indegree at least $q \cdot \mathsf{Hall}(G) = p \in \mathbb{N}$. Then interpret each left-hand side node in G'_{BP} as being in charge of a $\frac{1}{q}$-fraction of an edge in G_{BP}. This gives a fractional/randomized assignment in G_{BP} for which each right-hand side receives at least $\mathsf{Hall}(G)$ quantity of edges. Interpreting this assignment as an orientation in the original graph G, perhaps randomized, yields the desired result.

Case 4. When $\mathsf{Hall}(G) \notin \mathbb{Q}$, for any $\epsilon > 0$ there exists $\frac{p}{q} \in \mathbb{Q}$ such that $0 < \mathsf{Hall}(G) - \frac{p}{q} < \epsilon$. Take $\epsilon \to 0$: for any given $\frac{p}{q}$, apply the argument from Case 3. We then obtain a sequence of α-orientations for which $\alpha \to \mathsf{Hall}(G)$. ■

Definition 45 *The **Hall complexity** associated to a hypothesis class $\mathcal{H}$ is the function $\pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ defined*

$$\pi_{\mathcal{H}}(n) = \max_{S \in \mathcal{X}^n} n - \mathsf{Hall}\big(G(\mathcal{H}|_S)\big).$$

Recall now the transductive error rate of a learner A, as defined in Definition 30, i.e.,

$$\epsilon_{A,\mathcal{H}}(n) = \max_{S \in \mathcal{X}^n, h \in \mathcal{H}} L^{\mathrm{Trans}}_{S,h}(A).$$

We analagously define the **transductive error rate** of a class $\mathcal{H}$ as the pointwise minimal error rate attained by any of its learners, i.e.,

$$\epsilon_{\mathcal{H}}(n) = \min_{A} \epsilon_{A,\mathcal{H}}(n).$$

Proposition 46 *Fix any domain $\mathcal{X}$, label set $\mathcal{Y}$, and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. We have that $\epsilon_{\mathcal{H}}(n) = \frac{\pi_{\mathcal{H}}(n)}{n}$ for all $n \in \mathbb{N}$.*

Proof First note that $\pi_{\mathcal{H}}(n)$ is the smallest α for which the graphs $G(\mathcal{H}|_S)$ are all α-coorientable, by Proposition 44 and the fact that each $G(\mathcal{H}|_S)$ is α-coorientable if and only if it is $(n-\alpha)$-orientable. Now, by Lemma 42, $\mathcal{H}$ has a learner attaining transductive error $\leq \epsilon$ on samples of size n if and only if $G(\mathcal{H}|_S)$ is $(\epsilon \cdot n)$-coorientable for all $S \in \mathcal{X}^n$. The claim follows. ■

Remark 47 *The astute reader may notice the similarity between the Hall complexity, the maximum subgraph density from Haussler et al. (1994), and the maximum average degree from Daniely and Shalev-Shwartz (2014). In binary classification, the Hall complexity and the maximum subgraph density are equal, and hence both exactly characterize the transductive error, whereas the maximum average degree is a factor of 2 larger. When there are k labels, the maximum subgraph density serves as a loose lowerbound on the transductive error, and can be up to a factor of $k-1$ smaller. The maximum average degree, on the other hand, serves as an upperbound of the transductive error, and can be up to a factor of 2 larger. The Hall complexity, in exactly characterizing the transductive error, is sandwiched between the maximum subgraph density and the maximum average degree. We also point out that using Hall's theorem seems to be implicit in the proof of Lemma 57 in Rubinstein et al. (2009), although they are still focused on maximum density, and do not indicate that Hall's theorem permits an exact characterization of transductive error.*

Appendix C. Structural Risk Minimization: Supplement

C.1. Regularizer Using S Can Induce Any Learner

If regularizers are permitted to access the full data of S, it is easy to see that the picture degenerates completely: any learner can be witnessed as SRM with respect to a regularizer of this form. In particular, an arbitrary learner A can be recovered as SRM with respect to the following regularizer:

$$\psi : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y})^{<\omega} \times \mathcal{X} \longrightarrow \mathbb{R}_{\geq 0}$$
$$(h, S, x) \longmapsto \begin{cases} 0 & A(S)(x) = h(x), \\ \infty & A(S)(x) \neq h(x). \end{cases}$$

In particular, given a sample S and test point x, the regularizer can simply deduce $A(S)$ and output extremely large values for hypotheses disagreeing with $A(S)$ on x. Note that this argument relies crucially upon the regularizer's access to the test point in $\mathcal{X}$ as an input; any regularizer which is uniform with respect to $\mathcal{X}$ (even if granted full information of S) begets proper learners, which are insufficient by Proposition 14.

Note also that the guarantee $A(S)(x) \in \mathcal{H}(x)$ holds for all $x \in \mathcal{X}$ when A is any sensible learner, as we are learning in the realizable case with respect to the 0-1 loss. (In particular, any learner A violating this condition of "local properness" can be improved by obligating it to be locally proper, in any way.) Further, as our losses are bounded by 1, a regularizer's output of $\infty \notin \mathbb{R}_{\geq 0}$ in this example is equivalent to simply outputting c for any $c > 1$.

C.2. Support for Conjecture 18

We now present a candidate hypothesis class $\mathcal{H}$ which may justify our Conjecture 18. That is, we will define a class $\mathcal{H}$ which is PAC learnable, and for which we suspect that $\mathcal{H}$ *cannot* be learned by a local size-based regularizer. Let $\mathcal{X}$ be an infinite set, say $\mathcal{X} = \mathbb{N}$, and $\mathcal{Y} = \{*\} \cup 2^{\mathcal{X}}$, where we use $2^{\mathcal{X}}$ to denote the power set of $\mathcal{X}$.

Before defining $\mathcal{H}$, let $\mathcal{J} \subseteq (2^{\mathcal{X}})^3$ be the collection of all triples of subsets (R, S, T) such that:

1. $|R| = |S| = |T| =: k \in 2\mathbb{N}$
2. $|R \cap S| = |R \cap T| = |S \cap T| = k/2$. In particular, $R \cap S \cap T = \emptyset$.

For each $(R, S, T) \in \mathcal{J}$, we will define 8 hypotheses in $\mathcal{H}$. Namely, all those hypotheses $h \in \mathcal{Y}^{\mathcal{X}}$ satisfying:

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

Informally, each such h is simply the constant function $\_ \mapsto *$ outside of $R \cup S \cup T$, and on the regions $R \cap S$, $R \cap T$, and $S \cap T$ has the choice of acting as a constant function taking a value in $\{R, S\}$, $\{R, T\}$, or $\{S, T\}$ respectively. Geometrically, one can think of such a function as choosing how to layer the regions S, T, and R ontop of each other. (I.e., sheets of paper over each of S, T, and

R 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 $\mathcal{H}$ to be the union of all such functions over all triples $(R, S, T) \in \mathcal{J}$. One can see that that $\mathcal{H}$ is PAC learnable by the learner which defaults to outputting $*$ at $x \in \mathcal{X}$ unless the label $S \ni x$ has been observed in the sample, in which case it outputs S.[6] Informally, consider any function h arising from an $(R, S, T) \in \mathcal{J}$ 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 $R \cap S$, $S \cap T$, and $T \cap R$. 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 $\mathcal{H}$ not to be learnable by a local size-based regularizer. First note that ERM learners fail miserably to learn $\mathcal{H}$: fix a large set $S \subseteq \mathcal{X}$ and let D be the uniform distribution over points of the form $\{(x, S) : x \in S\}$. Then D is a realizable distribution, and consider the output of an ERM learner on a training set $S_{\text{train}} \sim D$ with $|S_{\text{train}}| < |S|/2$. With probability $> 1/2$, a test point $(x_{\text{test}}, S) \sim D$ will be such that x_{test} was not seen in S_{train}. In this case, there exists a hypothesis $h \in \mathcal{H}$ with empirical error 0 such that $h(x_{\text{test}}) \neq S$. Namely, an h arising from a triple of sets (R, S, T) such that $S_{\text{train}} \cap \big(R \cap S\big) = \emptyset$ and $x_{\text{test}} \in R \cap S$. An ERM learner is free to predict label R at such an x_{test}, there incurring constant error at test time. As the original set S may be chosen to be arbitrarily large, the problem affects ERM learners trained on arbitrarily large training sets S_{train}.

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 $\{(x, S) : x \in S\}$. That is, the great amount of symmetry inherent in $\mathcal{H}$ prevents $\mathcal{A}$ from recognizing that S is the most "natural" prediction for x_{test}, in contrast to any of the sets R which contain x_{test} yet avoid S_{train}. In short, it seems that a learner must peek into the training set S_{train} in order to learn the geometry of the underlying distribution and have a chance of learning. Local size-based regularizers, however, are not permitted to peek into the training set.

Appendix D. Omitted proofs

D.1. Proof of Proposition 16

Proof We use the *first Cantor class* of (Daniely et al., 2015; Daniely and Shalev-Shwartz, 2014). In particular, let $\{\mathcal{X}_d\}_{d \in 2\mathbb{N}}$ be a disjoint collection of sets with $|\mathcal{X}_d| = d$. Furthermore, let $\mathcal{Y}_d = 2^{\mathcal{X}_d} \cup \{*\}$ for each $d \in 2\mathbb{N}$. For each $A \subseteq \mathcal{X}_d$, define $h_A : \mathcal{X}_d \to \mathcal{Y}_d$ by

$$h_A(x) = \begin{cases} A & x \in A, \\ * & x \notin A. \end{cases}$$

Now let $\mathcal{X}_\infty = \cup_{d \in 2\mathbb{N}} \mathcal{X}_d$ and $Y_\infty = \big(\cup_{d \in 2\mathbb{N}} 2^{\mathcal{X}_d}\big) \cup \{*\}$. We can extend each $h_A : \mathcal{X}_d \to \mathcal{Y}_d$ to a function $\mathcal{X}_\infty \to \mathcal{Y}_\infty$ by simply definining it to return $*$ outside of $\mathcal{X}_d$. Lastly, set

$$\mathcal{H}_\infty = \Big\{h_A : A \subseteq \mathcal{X}_d \text{ for some } d, |A| = \frac{d}{2}\Big\}.$$

6. If *two* such labels $S \ni x$ and $R \ni x$ have been seen in the training sample, and this information reveals the true label of x (i.e., one label was seen on $S \cap T$), 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 S or T).

Then $\mathcal{H}_\infty$ is PAC learnable as a consequence of (Daniely and Shalev-Shwartz, 2014, Lemma 20), i.e., by the learner that returns the constant function $*$ if it attains zero empirical risk and the unique h_A attaining zero empirical risk otherwise. Note that this constant function is not in $\mathcal{H}_\infty$.

We now show that $\mathcal{H}_\infty$ cannot be learned by any local SRM learner. Fix a local regularizer $\psi : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. Now define $\psi_{\mathcal{H}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ as follows,

$$\psi_{\mathcal{H}}(x, y) = \inf_{\substack{h \in \mathcal{H}, \\ h(x)=y}} \psi(h, x).$$

Informally, $\psi_{\mathcal{H}}$ captures the local preferences over labels, rather than entire hypotheses, induced by ψ. We will say that ψ weakly prefers y to y' at x if $\psi_{\mathcal{H}}(x, y) \leq \psi_{\mathcal{H}}(x, y')$.

Suppose now that the following property P holds of ψ: for each $x \in \mathcal{X}_\infty$, ψ weakly prefers every set A containing x to $*$. We show that there is a learner $\mathcal{A}$ induced by ψ that is not a PAC learner for $\mathcal{H}_\infty$. Pick some $A \subseteq \mathcal{X}_d$ of size $\frac{d}{2}$. Let D_A be the uniform measure over the finite set $\big\{(x, *) : x \in \mathcal{X}_d \setminus A\big\}$. Note that D_A is an $\mathcal{H}_\infty$-realizable distribution, as $h_A \in \mathcal{H}_\infty$ attains a true error of zero with respect to it. Now let $S \sim D_A^m$ be a sample of size $m < d/4$. As $|\mathcal{X}_d \setminus A| = \frac{d}{2}$, S does not contain the full support of D_A. In particular, for any $x \in \mathcal{X}_d \setminus S$, there exists an $A_x \subseteq \mathcal{X}_d$ of size $\frac{d}{2}$ such that $x \in A_x$ and $A_x \cap S = \emptyset$.

For such x, $L_S(A_x) = 0$, as A_x avoids S. Furthermore, by definition of property P, $\psi_{\mathcal{H}}(x, A_x) \leq \psi_{\mathcal{H}}(x, *)$, meaning $\psi(h_{A_x}, x) \leq \psi(h_B, x)$ for every B such that $x \notin B$. Thus $\mathcal{A}$ can be taken such that $\mathcal{A}(S)(x) \neq *$. As this holds for all $x \in \mathcal{X}_d \setminus S$, $\mathcal{A}$ misclassifies all points in $\mathcal{X}_d \setminus A$ that are not in S. Consequently, A incurs an expected true error of at least $\frac{1}{2}$. And d can be taken to be arbitrarily large by ranging across $\{\mathcal{X}_d\}_{d \in 2\mathbb{N}} \subseteq \mathcal{X}_\infty$, meaning $\mathcal{A}$ is not a PAC learner.

Suppose now that ψ does not satisfy property P. Then there exists an $x \in \mathcal{X}_d$ and $A \ni x$ such that $\psi_{\mathcal{H}}(x, A) > \psi_{\mathcal{H}}(x, *)$. In particular, as h_A is the only function in $\mathcal{H}_\infty$ with A in its image, we have $\psi(h_A, x) > \psi(h_B, x)$ for any B such that $x \notin B$. Set $c_2 = \psi_{\mathcal{H}}(x, A)$ and $c_1 = \psi_{\mathcal{H}}(x, *)$. Now let D be the distribution placing $\min(\frac{c_2 - c_1}{2}, 1)$ mass on the point (x, A) and its remaining mass uniformly across datapoints in $\mathcal{X}_d \setminus A$ with label $*$. For increasingly large samples S drawn from D, the proportion p_S of sample points taking the form (x, A) will satisfy $p < c_2 - c_1$ with high probability. Consequently, at point x, h_A will attain greater structural risk than some h_B with $x \notin B$. That is, $\mathcal{A}(S)(x) \neq A$, and thus $\mathcal{A}$ incurs constant true error over arbitrarily large samples. This completes the proof. ■

D.2. Proof of Theorem 20

Before commencing with the proof, it will be useful to establish a basic fact: local regularizers of any kind should serve only to "tie-break" between hypotheses attaining zero empirical risk.

Lemma 48 *Let $\psi : \mathcal{H} \times \mathcal{X}^{<\omega} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a local unsupervised regularizer. Then, without loss of generality, ψ can be assumed to obey the following property for all $S \in \mathcal{X}^n$ and $x \in \mathcal{X}$:*

$$\psi(h, S, x) < \frac{1}{|S|} \quad \forall h \in \mathcal{H}.$$

Proof As we are learning in the realizable case with respect to the 0-1 loss, any learner A can be assumed to satisfy $A(S)(x) \in \big\{h(x) : h \in \operatorname{argmin}_{\mathcal{H}} L_S(h)\big\}$. In particular, for the underlying

distribution D, there exists an $h \in \mathcal{H}$ with $L_D(h) = 0$. Then on any sample S, $h \in \operatorname{argmin}_{\mathcal{H}} L_S(h)$ with probability 1. If A violates the previous property, then with probability 1, $A(S)(x) \neq h(x)$ and thus $\ell_{0-1}\big(A(S)(x), h(x)\big) = 1$, meaning A incurs the maximal loss possible (for any underlying distribution D).

Consequently, any regularizer ψ should be such that

$$\Big\{h(x) : h \in \operatorname*{argmin}_{\mathcal{H}} L_S(h) + \psi(h, S_{\mathcal{X}}, x)\Big\} \subseteq \Big\{h(x) : h \in \operatorname*{argmin}_{\mathcal{H}} L_S(h)\Big\} = \Big\{h(x) : L_S(h) = 0\Big\}.$$

In particular, the action of ψ should only help compare hypotheses attaining equal empirical risk. As $L_S(h) \in \{\frac{1}{|S|}, \ldots, \frac{|S|}{|S|}\}$ for any $h \in \mathcal{H}$, it suffices to show that an arbitrary ψ can be compressed to the interval $[0, \frac{1}{|S|})$ by a strictly increasing function. And indeed this can be achieved, by such a function as $x \mapsto \frac{2}{\pi|S|} \tan^{-1}(x)$. ■

Perhaps the most important ingredient to our central result is the relationship between unsupervised regularizers and acyclic orientations of OIGs. In particular, to demonstrate that an OIG can be oriented favorably by an unsupervised regularizer (i.e., by its induced learners), it suffices to demonstrate that it can be oriented favorably in an acyclic manner.

Proposition 49 *Let $\mathcal{Y}$ be a finite or countable label set, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class, and A a learner for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, corresponding to a collection of orientations for the one-inclusion graphs $\{G(\mathcal{H}|_S)\}_{S \in \mathcal{X}^{<\omega}}$. If A begets acyclic orientations on all the one-inclusion graphs of $\mathcal{H}$, then it is a local unsupervised SRM learner for some ψ.*

Proof Suppose that $\mathcal{Y}$ is finite and that A gives rise to acylic orientations of all one-inclusion graphs for $\mathcal{H}$. We claim that a local unsupervised regularizer for $\mathcal{H}$ can be thought of as an arbitrary function $\phi : V_\infty \to \mathbb{R}_{\geq 0}$, for V_∞ the union of vertices across all the one-inclusion graphs of $\mathcal{H}$. This can easily be seen by observing that a local unsupervised regularizer ψ can choose to judge each $h \in \mathcal{H}$ based only upon the information of its restriction to $S_{\mathcal{X}} \cup \{x_{\text{test}}\}$. Now fix one such graph $G(\mathcal{H}|_S) = (V, E)$ for $S \in \mathcal{X}^n$, directed by the action of A. As it is acyclic, it can be topologically ordered, so that each vertex $v \in V$ lies in layer $\ell_v \in \mathbb{N}$.

It thus suffices to exhibit a function $\phi : V \to \mathbb{R}_{\geq 0}$ such that A orients each $e \in E$ toward an incident vertex with maximal output under ϕ. This is achieved by, for instance,

$$\phi : v \mapsto \frac{1}{2|S|} \cdot \left(1 - \frac{1}{\ell_v}\right).$$

Notably, ϕ satisfies each of the two properties we require of it: that it be strictly increasing with nodes' layers in the topological ordering of $G(\mathcal{H}|_S)$, and that it be bounded above by $\frac{1}{2|S|}$, in accordance with Lemma 48.

The case for countable $\mathcal{Y}$ is slightly more involved. Let $\mathcal{Y}$ be countably infinite and A a learner for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ that induces acyclic orientations on all the one-inclusion graphs of $\mathcal{H}$, $\{G(\mathcal{H}|_S)\}_{S \in \mathcal{X}^{<\omega}}$. Fix one such $S \in \mathcal{X}^n$ and $G(\mathcal{H}|_S) = (V, E)$ the accompanying one-inclusion graph. Note that V is at most countable, as it has cardinality at most $|\mathcal{Y}^n| = |\mathcal{Y}|$. The acyclic orientation on G endows its vertex set V with a partial order P, defined by $u \leq v$ if $u = v$ or if there exists a path $p : u \to v$.

Appealing to Zorn's lemma, P can be completed into a total order on V. Furthermore, V then embeds into $\mathbb{R}$ as a totally ordered set owing to its countability. That is, one can embed $V = \{v_i\}_{i \in \mathbb{N}}$ into $\mathbb{R}$ inductively, beginning with $v_0 \mapsto 0$. Once $(v_i)_{i<k}$ have been embedded as $(r_i)_{i<k}$, v_k can be mapped to $\max_{i<k} r_i + 1$ if $v_k > \max_{i<k} v_i$, to $\min_{i<k} r_i - 1$ if $v_k < \min_{i<k} v_i$, and otherwise to $\frac{r_i + r_j}{2}$ for v_i and v_j the least upper bound and greatest lower bounds of v_k in $(v_i)_{i<k}$.

Lastly, by post-composing such an embedding with an isomorphism of ordered sets $\mathbb{R} \to (0, \frac{1}{|S|})$, such as $x \mapsto \frac{2}{\pi |S|} \tan^{-1}(x)$, we have that V embeds into $(0, \frac{1}{S})$ as a totally ordered set. Call this embedding ϕ. As in the proof for finite $\mathcal{Y}$, ϕ gives rise to a local unsupervised regularizer ψ that recovers precisely the action of A. That is, $\psi(h, S, x)$ chooses to judge h based upon its restriction to $S \cup \{x\}$, at which point it acts as the embedding ϕ.

Notably, ψ satisfies the two crucial properties required of it: that it be bounded above by $\frac{1}{|S|}$, in accordance with Lemma 48, and that it be compatible with the acyclic orientation of $G(\mathcal{H}|_S)$ induced by A. ■

We are now ready to prove Theorem 20.

Proof By Proposition 49, it suffices to demonstrate the following claim: for any $S \in \mathcal{X}^n$, there exists an acylic orientation of $G(\mathcal{H}|_S)$ that is optimal within a factor of 2, i.e., that minimizes nodes' maximal outdegrees to within a factor of 2. In particular, note that the unsupervised regularizer ψ described in Proposition 49 induces a unique learner. We will use the *density* of a finite undirected graph to refer to the average degree of its nodes. Now suppose that $G(\mathcal{H}|_S)$ is α-coorientable, as described in Definition 41. Then the maximal density of any finite subgraph of $G(\mathcal{H}|_S)$ is at most 2α, by (Daniely and Shalev-Shwartz, 2014, Theorem 2). When $G(\mathcal{H}|_S) = (V, E)$ is finite, this allows us to design an acyclic (2α)-coorientation by employing the following *k-core* algorithm implicit in (Daniely and Shalev-Shwartz, 2014, Lemma 3): for $i \in [|V|]$, compute the vertex $v \in G(\mathcal{H}|_S)$ of lowest degree, remove it from $G(\mathcal{H}|_S)$, and place it in the ith layer of a topological ordering of $G(\mathcal{H}_S)$. The outdegree of any vertex in this topological ordering is precisely its degree in the undirected graph $G(\mathcal{H}|_S)$ before it was removed, which is at most 2α owing to the maximal subgraph density of $G(\mathcal{H}|_S)$.

When $G(\mathcal{H}|_S)$ is infinite, the claim follows from the compactness theorem of propositional calculus (see Daniely and Shalev-Shwartz (2014); Brukhim et al. (2022)), which, for completeness, we detail as follows.

Let $G(\mathcal{H}|_S) = (V, E)$ be infinite and β the maximal density of any of its finite subgraphs (i.e., the average degree of nodes). By (Daniely and Shalev-Shwartz, 2014, Theorem 2), it suffices to show that $G(\mathcal{H}|_S)$ can be β-cooriented. Let $|S| = n$, and for any $(v, e) \in V \times E$ let E_v denote the set of all edges incident to v and V_e the set of all vertices incident to e. For vertices $u, v \in V$, let $\mathcal{P}_{u,v}$ denote the set of all paths from u to v, i.e., of finite sequences $p = \big((e_1, v_1), \ldots, (e_\ell, v_\ell)\big)$ such that e_1 is incident to u, e_i is incident to each of v_i and v_{i-1} (when they exist), and $v_\ell = v$.

Now consider the set of propositional variables $\{P_{e,v} : e \in E, v \in V_e\}$. Intuitively, $P_{e,v}$ will be true if the edge e is oriented to v and false otherwise. We define a set of sentences Σ as follows.

1. $\neg(P_{e,v} \wedge P_{e,v'})$ for all $e \in E$ and pairs of nodes v, v' both incident to e.
2. $\displaystyle\bigvee_{\substack{E'_v \subseteq E_v, \\ |E'_v| \geq n - \beta}} \bigwedge_{e' \in E'_v} P_{e',v}$ for all $v \in V$.

3. $\left(\neg \bigwedge_{(e,s)\in p} P_{e,s}\right) \vee \left(\neg \bigwedge_{(e',s')\in p'} P_{e',s'}\right)$ for all $u, v \in V$ and $(p, p') \in \mathcal{P}_{u,v} \times \mathcal{P}_{v,u}$.

Sentences from (1.) correspond to the requirement that each edge be oriented to at most one incident vertex, those from (2.) demand that each node have indegree at least $n - \beta$ (equivalently, outdegree at most β), and those from (3.) demand that $G(\mathcal{H}|_S)$ never contain both a path $u \to v$ and $v \to u$ (i.e., a cycle). Note that sentences from (2.) are finite because each vertex v has finite degree exactly n in the undirected graph $G(\mathcal{H}|_S)$, and sentences from (3.) are finite because paths in $\mathcal{P}_{u,v}$ are finite.

Let us now demonstrate that Σ is finitely satisfiable. Suppose we impose a finite collection of sentences $\Sigma' \subseteq \Sigma$, involving only the variables $P_{e,v}$ for the finite set $V' \subseteq V$. Let $G[V'] \subseteq G(\mathcal{H}|_S)$ be the subgraph of $G(\mathcal{H}|_S)$ with vertex set V' and edge set $E' \subseteq E$ consisting of those edges with at least two incident nodes in V'. As $G[V']$ is a finite subgraph of $G(\mathcal{H}|_S)$, its density is at most β. By our previous work from the proof of Theorem 20 for finite $\mathcal{Y}$, $G[V']$ can β-cooriented acyclically. Such an acyclic β-coorientation amounts precisely to a choice of $P_{e,v}$ for each $(e, v) \in E' \times V'$ satisfying sentences (1.), (2.), and (3.) restricted to $G[V']$. Setting all the remaining variables to false continues to satisfy Σ', thus demonstrating that Σ is finitely satisfiable.

Then, by compactness, Σ is satisfiable, meaning there exists a partial orientation σ of $G(\mathcal{H}|_S)$ such that each node has outdegree at most β and there are no directed cycles. In particular, there may be an $e \in E$ such that $P_{e,v}$ is false for all $v \in V_e$. Such a partial orientation on $G(\mathcal{H}|_S)$ endows its vertex set V with a partial order P, defined by $u \leq v$ if $u = v$ or if there exists a directed path $p : u \to v$. Appealing to Zorn's lemma, P can be completed into a total order on V. Embedding V into $(0, \frac{1}{|S|})$ as a totally ordered set — as in the proof of Proposition 49 — defines a regularizer ψ. Notably, ψ recovers the action of σ on the respective edges, and any manner of completing σ into a total orientation can only increase nodes' indegrees (i.e., reduce their outdegrees), completing the argument. ■

Remark 50 *The unsupervised regularizer employed in Propositions 49 and 20 does not distinguish between x_{test} and the elements of $S_{\mathcal{X}}$. In particular, it is symmetric with respect to x_{test} and any element of $S_{\mathcal{X}}$. This has two central consequences:*

1. *Propositions 49 and 20 hold for a regularizer that factors through $S_{\mathcal{X}} \cup \{x_{\text{test}}\}$, i.e., that may as well be defined to receive $S_{\mathcal{X}} \cup \{x_{\text{test}}\}$ as input. This demonstrates sufficiency of a regularizer receiving even less information than the local unsupervised regularizers of Definition 19.*

2. *Semantically, this implies the existence of near-optimal transductive learners based on regularizers that decide their values in the transductive setting merely after observing the collection of unlabeled datapoints (including test point). Furthermore, it demonstrates that an OIG $G = (V, E)$ can always be oriented near-optimally by assigning a value to each node $v \in V$, rather than by assigning an incident node to each $e \in E$. This achieves another central aim of this section: to describe optimal orientations of OIGs more parsimoniously, using global structure rather than local structure.*

D.3. Proof of Theorem 21

Before proving Theorem 21, we introduce the notion of an *assignment* for a bipartite graph G, which will replace our discussion of orientations in $G(\mathcal{H}|_S)$, as the two are equivalent.

Definition 51 *Let $G = (\mathcal{A}, \mathcal{B}, E)$ be a bipartite graph. An* **assignment** *in G is a function $\sigma : \mathcal{A} \to \mathcal{B}$ such that $(a, \sigma(a)) \in E$ for all $a \in \mathcal{A}$.*

The following is immediate from the definitions of OIGs and their bipartite counterparts.

Lemma 52 *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and sequence $S \in \mathcal{X}^n$. Let $G(\mathcal{H}|_S)$ be the OIG of $\mathcal{H}$ with respect to S and $G_{\text{BP}} = (\mathcal{A}, \mathcal{B}, E)$ its bipartite counterpart, as defined in Definition 37. Then the following are equivalent for any $\alpha \in \mathbb{N}$.*

1. *There exists an $(n - \alpha)$-orientation of $G(\mathcal{H}|_S)$; and*
2. *There exists an assignment σ of G_{BP} such that each $v \in \mathcal{B}$ receives at least $n - \alpha$ degree, i.e., $|\sigma^{-1}(v)| \geq n - \alpha$.*

With the language of assignments, we can begin discussion of our maximum entropy convex program.

Proposition 53 *Let $\mathcal{H}$ be a hypothesis class, and recall the optimal transductive error rate $\epsilon_{\mathcal{H}}(n) = \frac{\pi_{\mathcal{H}}(n)}{n}$. For any $S \in \mathcal{X}^n$, let $G(\mathcal{H}|_S)$ be the one-inclusion graph of $\mathcal{H}$ on S and $\mathcal{D}$ the collection of all orientations of $G(\mathcal{H}|_S)$ (equivalently, the collection of assignments in $G_{\text{BP}} = (\mathcal{A}, \mathcal{B}, E)$, the bipartite analogue of $G(\mathcal{H}|_S)$). Then there is a unique distribution over assignments $P^* \in \Delta_{\mathcal{D}}$ such that each $v \in \mathcal{B}$ receives at least $n \cdot (1 - \epsilon_{\mathcal{H}}(n)) = n - \pi_{\mathcal{H}}(n)$ in-degree in expectation, and the following conditions hold simultaneously:*

1. *P^* achieves maximum entropy among all $P \in \Delta_{\mathcal{D}}$ subject to the degree requirements.*
2. *P^* induces a randomized transductive learner $\mathcal{A}$ which, in expectation over the randomness of the learner, achieves optimal error rate $\epsilon_{\mathcal{H}}(n)$.*

Proof (2.) follows immediately from the connection between in-degree and transductive error, as established in Lemma 42.

(1.) is proven using the maximum entropy convex program of Singh and Vishnoi (2014), with a slight modification owing to the fact that we are in a more general multiset setting. Let $\mathcal{D}$ be the set of all assignments in G_{BP}. Recall that such assignments can be thought of as learners, as discussed in Section B.1. We will think of each $d \in \mathcal{D}$ as a $|\mathcal{A}| \times |\mathcal{B}|$ matrix with 0-1 entries, where all row sums are exactly 1 (that is, every partially labeled dataset in $\mathcal{A}$ is oriented toward one fully labeled dataset in $\mathcal{B}$). We will index entries of this matrix as $d(u, v)$ for $u \in \mathcal{A}, v \in \mathcal{B}$.

Let n be fixed. Our goal is to find a maximum entropy distribution over assignments $P^* \in \Delta_{\mathcal{D}}$ such that each $v \in \mathcal{B}$ is allocated at least $c = (1 - \epsilon_{\mathcal{H}}(n))n$ in-degree in expectation. By Lemma 42, this would suffice to define a (randomized) learner induced by P^* which achieves error at most $\epsilon_{\mathcal{H}}(n)$.

Therefore, we want the solution to the following maximum entropy optimization problem, which we term `MaxEnt`.

$$\begin{aligned}
\texttt{MaxEnt} = \max_{p_d:\, d\in\mathcal{D}} \quad & \sum_{d\in\mathcal{D}} p_d \ln \frac{1}{p_d} \\
\text{s.t.} \quad & \sum_{d\in\mathcal{D}} p_d \sum_{u\in\mathcal{A}} d(u,v) \geq c && \forall v \in \mathcal{B} \\
& \sum_{d\in\mathcal{D}} p_d = 1 \\
& p_d \geq 0 && \forall d \in \mathcal{D}
\end{aligned}$$

Notice that `MaxEnt` is a convex program with a concave objective and linear constraints. The number of variables is equal to the number of assignments in the graph G_{BP}.

Let us confirm that the program is feasible and, in fact, enjoys a unique optimal solution. Feasibility arises as a consequence of the fact that some randomized learner attains error rate $\epsilon_{\mathcal{H}}(n)$, corresponding to a randomized orientation of the OIG that satisfies the degree bounds required by our program. Uniqueness of the solution follows from strict concavity of the objective. This completes the proof of (1.). ■

Theorem 21 demonstrates that the maximum entropy learner can furthermore be interpreted as *Bayesian* in the following sense: First, it learns a prior over hypotheses from the unlabeled data. (In particular, over the projection of $\mathcal{H}$ to the unlabeled data.) Second, given the labels of all but the test point, it performs a Bayes update to form a posterior over the hypotheses consistent with these labels. Third, it samples a hypothesis from this posterior, which it then uses to predict a label for the test point.

The proof of Theorem 21 will build upon the dual characterization of the maximum entropy convex program `MaxEnt` defined in Proposition 53. To ensure that this program (and its dual) are well-defined, we need to check three things: feasibility, uniqueness of the optimal solution, and strong duality. The proof of Proposition 53 already addresses feasibility and uniqueness. Strong duality follows from the following observation: Notice that the inequality constraints of `MaxEnt` are affine (linear) in the variables. Therefore, we can directly apply the weak version of Slater's condition to get that strong duality holds. Recall that weak Slater's condition requires strict feasibility only in non-affine inequality constraints (of which we have none). Then, feasibility is enough to prove strong duality.

The final key ingredient to the proof of Theorem 21 is the following lemma which relates the (optimal) primal and dual variables of `MaxEnt`.

Lemma 54 *Let $\lambda_v \in \mathbb{R}$ for all $v \in \mathcal{B}$ and $z \in \mathbb{R}$ be the optimal dual variables. Then, for each $d \in \mathcal{D}$, the associated optimal primal variable p_d can be written as $p_d^* = e^{-1-z} \prod_{(u,v)\in d} \exp(-\lambda_v)$.*

Proof Since strong duality of `MaxEnt` holds, we will derive the dual problem similar to Singh and Vishnoi (2014) (see their Appendix A.1). First, we find the Lagrangian.

$$L(p,\lambda,z) = \sum_{d\in\mathcal{D}} p_d \ln(\frac{1}{p_d}) + \sum_{v\in\mathcal{B}} \lambda_v (c - \sum_{d\in\mathcal{D}} p_d \sum_{u\in\mathcal{A}} d(u,v)) + z(1 - \sum_{d\in\mathcal{D}} p_d) \tag{1}$$

$$= \sum_{d\in\mathcal{D}} p_d \ln(\frac{1}{p_d}) + c\sum_{v\in\mathcal{B}} \lambda_v - \sum_{d\in\mathcal{D}} p_d \sum_{v\in\mathcal{B}} \lambda_v \sum_{u\in\mathcal{A}} d(u,v) + z - z\sum_{d\in\mathcal{D}} p_d \tag{2}$$

Now take partials and set to zero.

$$\frac{\partial L}{\partial p_d} = p_d \cdot \frac{1}{\frac{1}{p_d}} \cdot \frac{\partial}{\partial p_d}(\frac{1}{p_d}) - \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) - z + \ln(\frac{1}{p_d}) = 0 \tag{3}$$

$$-1 - \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) - z + \ln(\frac{1}{p_d}) = 0 \tag{4}$$

Therefore,

$$\ln(\tfrac{1}{p_d}) = 1 + \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) + z$$

$$p_d = \exp\left(-1 - \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) - z\right). \tag{5}$$

Summing over all $d \in \mathcal{D}$,

$$\sum_{d \in \mathcal{D}} p_d = e^{-1-z} \sum_{d \in \mathcal{D}} \exp\left(-\sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v)\right).$$

This obtains the characterization of the probability of an assignment in terms of the dual variables, as we required.

For completeness, we finish the derivation of the dual optimization problem. Multiply each (4) by p_d and sum over all A to get that

$$\sum_{d \in \mathcal{D}} \left(-p_d - p_d \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) - z p_d + p_d \ln(\tfrac{1}{p_d})\right) = 0,$$

implying that

$$\sum_{d \in \mathcal{D}} p_d = \sum_{d \in \mathcal{D}} \left(-p_d \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) - z p_d + p_d \ln(\tfrac{1}{p_d})\right).$$

Plug these two facts into (2):

$$L(p,\lambda,z) = c \sum_{v \in \mathcal{B}} \lambda_v + z + \sum_{d \in \mathcal{D}} p_d = c \sum_{v \in \mathcal{B}} \lambda_v + z + e^{-1-z} \sum_{d \in \mathcal{D}} \exp\left(-\sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v)\right).$$

Which is only a function of λ, z. Take partial with respect to z and set to zero:

$$\frac{\partial L}{\partial z} = 1 - e^{-1-z} \sum_{d \in \mathcal{D}} \exp\left(-\sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v)\right) = 0$$

$$z = -1 + \ln\left(\sum_{d \in \mathcal{D}} \exp\left(-\sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v)\right)\right).$$

Therefore, the dual optimization problem then becomes:

$$\min_{\forall v \in \mathcal{B},\ \lambda_v \in \mathbb{R}} c \sum_{v \in \mathcal{B}} \lambda_v + \ln \sum_{d \in \mathcal{D}} \exp\left(- \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) \right).$$

Notice that (5) gives us the optimal value for each p_d:

$$p_d = e^{-1-z} \exp\left(- \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) \right) = e^{-1-z} \prod_{v \in \mathcal{B}} \prod_{u \in \mathcal{A}} \exp\left(-\lambda_v d(u,v)\right) = e^{-1-z} \prod_{(u,v) \in d} \exp(-\lambda_v),$$

where the last equality follows from the fact that $d(u,v) \neq 0$ only for (u,v) which exist in the assignment d. ■

We now prove Theorem 21.

Proof Let $G_{\mathrm{BP}} = (\mathcal{A}, \mathcal{B}, E)$ be the bipartite OIG associated to the unlabeled datapoints $S^+_{\mathcal{X}} = (x_1, \ldots, x_n)$. Let $\mathcal{D}$ denote the collection of assignments in G_{BP}. For a given assignment $d \in \mathcal{D}$, we say that $(u,v) \in d$ if the edge $(u,v) \in E$ is selected in d, and write $d(u,v) = 1$. From the dual derivation of `MaxEnt` in Lemma 54, within the optimal randomized maximum entropy learner P^*, each $d \in \mathcal{D}$ has an associated probability given by:

$$p^*_d = e^{-1-z} \prod_{(u,v) \in d} \exp(-\lambda_v)$$

where λ_v are dual variables corresponding to each $v \in \mathcal{B}$, and the final dual variable z has identical value for each $d \in \mathcal{D}$ given by:

$$z = -1 + \ln \left(\sum_{d \in \mathcal{D}} \exp\left(- \sum_{v \in \mathcal{B}} \lambda_v \sum_{u \in \mathcal{A}} d(u,v) \right) \right).$$

For all $v \in \mathcal{B}$, let $\gamma_v = \exp(-\lambda_v)$, and define $\rho_v = \gamma_v / \sum_{v \in \mathcal{B}} \gamma_v$. Then we can rewrite p^*_d as follows.

$$p^*_d = e^{-1-z} \prod_{(u,v) \in d} \gamma_v = e^{-1-z} \left(\sum_{v \in \mathcal{B}} \gamma_v \right)^n \prod_{(u,v) \in d} \frac{\gamma_v}{\sum_{v \in \mathcal{B}} \gamma_v} = \underbrace{e^{-1-z} \left(\sum_{v \in \mathcal{B}} \gamma_v \right)^n}_{(A)} \cdot \prod_{(u,v) \in d} \rho_v \propto \prod_{v \in \mathcal{B}} (\rho_v)^{\deg_d(v)}$$

We let $\deg_d(v)$ denote the degree of vertex $v \in \mathcal{B}$ in the assignment d. The final proportionality claim holds because (A) is fixed and takes the same value for any d. We can interpret this as saying that the probability that P^* selects a certain assignment is exactly proportional to the product of the normalized dual variables ρ_v for the fully labeled datasets $v \in \mathcal{B}$ present in that assignment.

Using the optimal dual variables of the maximum entropy convex program, we have effectively defined a prior distribution ρ on hypotheses as $\rho_v = \gamma_v / \sum_{v \in \mathcal{B}} \gamma_v$ for all $v \in \mathcal{B}$.

We will argue that the optimal (randomized) learner implied by P^* takes the following special form. Consider the distribution over assignments P' generated by the following random process

where for each $u \in \mathcal{A}$, we sample an incident $v \in \mathcal{B}$, independently and with probability proportional to ρ_v. Then, for any assignment $d \in \mathcal{D}$, the probability of observing d under the random process is given by the following.

$$p'_d = \prod_{(u,v)\in d} \frac{\rho_v}{\sum_{(u,v')\in E} \rho_{v'}} = \prod_{v\in\mathcal{B}} (\rho_v)^{\deg_d(v)} \underbrace{\prod_{u\in\mathcal{A}} \frac{1}{\sum_{(u,v')\in E} \rho_{v'}}}_{\text{(B)}} \propto \prod_{v\in\mathcal{B}} (\rho_v)^{\deg_d(v)}$$

Where the proportional follows from the fact that (B) is identical for all assignments d. Notice that p^*_d and p'_d are proportional to the same quantity for all $d \in \mathcal{D}$, and therefore the distributions P' and P^* are identical. This shows that the optimal randomized learner is in reality, for a given partially labeled dataset u, sampling from the previously defined prior over hypotheses, subject to a restriction to only the consistent hypotheses.

Since each hypothesis consistent with u gives rise to u with equal probability $1/n$, a simple application of Bayes' theorem implies that the optimal learner is sampling from the posterior $\rho' = \rho|u$ induced by the partially labeled dataset u. ■

D.4. Proof and Discussion of Corollary 23

We now argue that the maximum entropy randomized learner can be viewed as an SRM, and also as an instantiation of the principle of maximum entropy. Recall that for two distributions P and Q supported on a finite set Ω, the *relative entropy* from Q *to* P is defined as

$$D_{KL}(P \mid Q) = \sum_{x\in\Omega} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

Lemma 55 *Given a distribution Q with finite support Ω, and a subset $\Omega' \subset \Omega$, the distribution P supported on Ω' which has minimum relative entropy from Q is exactly the restriction of Q to Ω'.*

Proof Consider the following optimization problem describing the distribution with minimum relative entropy from Q, subject to the constraint that it only have non-zero support on elements of Ω'.

$$\min_{P\in\Delta_{\Omega'}} \sum_{x\in\Omega'} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

The partial derivative of the objective with respect to $P(x)$ is $\log\left(\frac{P(x)}{Q(x)}\right) + 1$. Given the constraint to probability distributions on Ω', the KKT conditions state that all partial derivatives are equal. The unique distribution satisfying the KKT conditions is the restriction of Q to Ω'. ■

We are now equipped to prove Corollary 23.

Proof Using Lemma 55, we have that when the ground truth for $n-1$ datapoints is revealed as some $u^* \in \mathcal{A}$, the randomized optimal (maximum entropy) learner is Bayesian in that it updates its posterior over hypotheses to have minimum relative entropy to the prior ρ, constrained on outputting hypotheses consistent with u^*.

Now, consider the learner induced by the following regularizer:

$$\psi(H, S_{-i}, x) = \frac{1}{K} \tan^{-1}(D_{KL}(H \mid \rho)),$$

where $K > 0$ is a parameter, and ρ corresponds to the previously defined re-normalized dual variables of the maximum entropy convex program. As K grows large, this learner places more relative importance on empirical risk. Recall that our maximum entropy learner minimizes $D_{KL}(H \mid \rho)$ subject to empirical risk equaling exactly zero. (Equivalently, it minimizes $\tan^{-1} D_{KL}(H \mid \rho)$, which has the favorable property of being bounded above.) Therefore, as $K \to \infty$ the output of the learner induced by ψ converges in total variation distance to the output of our maximum entropy learner. ■

In addition to being an SRM in the generalized sense just described, our learner can also be interpreted as an instantiation of the maximum entropy principle. In particular, if the prior ρ were uniform, then indeed our learner would sample from the maximum entropy distribution over hypotheses consistent with the supervised training data. More generally, sampling from the distribution which hues most closely to the prior subject to the provided labels — as measured by relative entropy — is the natural generalization of the maximum entropy principle to incorporate prior knowledge. In other words, our learner deviates as little as possible from the prior subject to being consistent with the provided labels.

The fact that our learner simultaneously instantiates SRM and the maximum entropy principle may appear counter-intuitive or even contradictory. Indeed, SRM is the embodiment of *Occam's razor*, which prefers the most simple hypotheses consistent with the data. On the other hand, maximizing entropy might appear to be the opposite of this, as high entropy distributions are arguably more "complex." However, taking the perspective of relative entropy from the uniform distribution (or from a high entropy prior ρ), the maximum entropy principle can be viewed as making as few assumptions as possible beyond those substantiated by the data. This is, in fact, fully in accordance with the spirit of Occam's razor.

Appendix E. Extension to Agnostic Learning

Our discussion of learning and one-inclusion graphs has thus far been restricted to the realizable case. Indeed, the structure of one-inclusion graphs, and of the transductive learning setting, depends crucially upon the guarantees provided by learning in the realizable case. In the agnostic case, any notions analogous to the OIG or to transductive learning would be obligated to look considerably different: the fully and partially labeled datasets of Figure 2 would no longer be required to agree with a function in $\mathcal{H}$, and the adversary of Definition 30 would likewise be permitted to label its datapoints arbitrarily. But how exactly should an *agnostic* one-inclusion graph be defined, and how would its optimal orientations be judged? How would this relate to an agnostic notion of transductive learning, if at all?

We devote this section precisely to the development such concepts, and introduce an agnostic one-inclusion graph whose optimal orientations — judged using vertices' outdegrees minus their "credits" — correspond precisely to learners attaining optimal agnostic transductive error. We also demonstrate that an agnostic version of the Hall complexity again characterizes the optimal error rates of hypothesis classes exactly, and exhibit one such optimal learner using maximum entropy programs.

Throughout the section, we employ the 0-1 loss function.

E.1. Transductive Learning

A crucial tool in our endeavor will be the notion of Hamming distance between functions.

Definition 56 *Let $S \in (\mathcal{X} \times \mathcal{Y})^n$ be a sample and $h \in \mathcal{Y}^{\mathcal{X}}$ a hypothesis. The **Hamming distance** between S and h, denoted $||S - h||_0$, is the empirical error incurred by h on S, i.e.,*

$$||S - h||_0 = \sum_{i=1}^{n} \big[h(x_i) \neq y_i\big].$$

When $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a hypothesis class, the Hamming distance *between S and $\mathcal{H}$ is the minimal Hamming distance incurred between S and any $h \in \mathcal{H}$, i.e.,*

$$||S - \mathcal{H}||_0 = \min_{h \in \mathcal{H}} ||S - h||_0.$$

We are now equipped to define the problem of transductive learning in the agnostic case.

Definition 57 *The **agnostic transductive learning** setting is that in which the following steps take place:*

1. *An adversary chooses a collection of n labeled datapoints, $S = \big((x_1, y_1), \dots (x_n, y_n)\big)$.*
2. *The unlabeled datapoints in S are revealed to the learner, i.e., the data of $(x_1, \dots, x_n)$.*
3. *One datapoint x_i is selected uniformly at random from S as the test point. The information of*

   $$S_{-i} = \big((x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n)\big)$$

   is revealed to the learner.
4. *The learner is prompted to predict the label of x_i, i.e., y_i.*

Definition 58 *The **agnostic transductive error** incurred by a learner A on the labeled sample S is its expected error over the uniformly random choice of x_i, relative to the performance of the best hypothesis in $\mathcal{H}$, i.e.,*

$$L_S^{\text{Trans}}(A) = \left(\frac{1}{n} \sum_{i \in [n]} [A(S_{-i})(x_i) \neq y_i] \right) - \frac{1}{n} \|S - \mathcal{H}\|_0.$$

Let us briefly justify that Definitions 57 and 58 are the appropriate generalizations of their realizable counterparts. We impose two key alterations:

1. The data S selected by the adversary is no longer required to be labeled according to an $h \in \mathcal{H}$. This agrees precisely with the notion of agnostic PAC learning, in which a learner is required to perform well with respect to any distribution D over $\mathcal{X} \times \mathcal{Y}$.

2. The learner is judged based only on its performance relative to the best hypothesis in $\mathcal{H}$. This again corresponds to the PAC criterion for agnostic learners, in which learners are benchmarked with respect to hypotheses in $\mathcal{H}$.

We now define error rates in the standard manner.

Definition 59 *The **agnostic transductive error rate** incurred by a learner A is the function*

$$\epsilon^{\mathrm{Ag}}_{A,\mathcal{H}}(n) = \max_{S \in (\mathcal{X} \times \mathcal{Y})^n} L^{\mathrm{Trans}}_S(A).$$

The agnostic transductive error rate *of a hypothesis class $\mathcal{H}$ is the pointwise minimal error rate enjoyed by any of its learners, i.e.,*

$$\epsilon^{\mathrm{Ag}}_{\mathcal{H}}(n) = \inf_A \epsilon^{\mathrm{Ag}}_{A,\mathcal{H}}(n),$$

where the infimum ranges over all learners for $\mathcal{H}$.

E.2. Agnostic One-inclusion Graphs

We now present a simple modification of the one-inclusion graph that captures the problem of transductive learning in the agnostic case. We note that similar techniques were introduced in the work of Long (1998) to analyze binary classification for realizable learning with distribution shift. Our analysis, however, applies to multiclass classification in the agnostic case over arbitrary label sets, and demonstrates an equivalence between (optimal) transductive learners and (optimal) orientations of agnostic one-inclusion graphs. We also introduce an *agnostic Hall complexity*, akin to the Hall complexity introduced in Section 3, that exactly characterizes the errors of optimal transductive learners.

Definition 60 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. The **agnostic one-inclusion graph** of $\mathcal{H}$ with respect to $S \in \mathcal{X}^n$, denoted $G_{\mathrm{Ag}}(\mathcal{H}|_S)$, is the hypergraph given by the following vertex and edge sets:*

- *$V = \mathcal{Y}^n$, one node for each possible labeling of the n datapoints.*
- *$E = \bigcup_{i=1}^n \mathcal{Y}^n|_{S_{-i}}$, where $e \in \mathcal{Y}^n|_{S_{-i}}$ is incident to each $v \in \mathcal{Y}^n$ such that $v|_{S_{-i}} = e$.*

As in Definition 8, we will sometimes write such an edge e as $e_{(g,i)}$, with $g \in \mathcal{Y}^n$ and $i \in [n]$. Under this representation, $e_{(g,i)}$ is incident to all nodes $v \in \mathcal{Y}^n$ such that $g|_{S_{-i}} = v|_{S_{-i}}$.

In the setting of binary classification, with $\mathcal{Y} = \{0, 1\}$, $G_{ag}(\mathcal{H}|_S)$ is simply the $|S|$-dimensional boolean hypercube. Larger label sets $\mathcal{Y}$ give rise to analogues of the boolean hypercube sometimes referred to as *Hamming graphs* (Brouwer and Haemers, 2011, Section 12.3.1). Though the vertex and edge sets of the agnostic one-inclusion graph $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ do not explicitly depend on the class $\mathcal{H}$ itself — only S — we will think of $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ as containing the information of which vertices $v \in V$ are members of $\mathcal{H}|_S$ and which are not. This will be useful information to retain when handling such graphs and, for instance, allows one to deduce $||v - \mathcal{H}||_0$ for any vertex v using only the information in $G_{\mathrm{Ag}}(\mathcal{H}|_S)$.

Lemma 61 *Let $\mathcal{H}$ be a hypothesis class, $S \in (\mathcal{X} \times \mathcal{Y})^n$ a sample, and $S_{\mathcal{X}} \in \mathcal{X}^n$ the sequence of unlabeled datapoints in S. Then the following conditions are equivalent for any learner A.*

1. *A incurs agnostic transductive error at most ϵ on the instance S.*
2. *S, thought of as a vertex in $G_{\mathrm{Ag}}(\mathcal{H}|_{S_{\mathcal{X}}})$, has outdegree at most $n \cdot \epsilon + \|S - \mathcal{H}\|_0$ in the graph oriented by A.*

Proof Let $v \in V$ be the node of S in $G_{\mathrm{Ag}}(\mathcal{H}|_S)$. For an edge e incident to g, we write $e \to g$ if e is oriented towards g (by the action of A) and $e \not\to g$ otherwise. Then,

$$\begin{aligned} L_S^{\mathrm{Trans}}(A) &= -\frac{1}{n} \cdot \|S - \mathcal{H}\|_0 + \frac{1}{n} \sum_{i \in [n]} [A(S_{-i})(x_i) \neq y_i] \\ &= -\frac{1}{n} \cdot \|S - \mathcal{H}\|_0 + \frac{1}{n} \sum_{i \in [n]} [e_{(g,i)} \not\to g] \\ &= -\frac{1}{n} \cdot \|S - \mathcal{H}\|_0 + \frac{1}{n} \cdot \mathrm{outDeg}(g). \end{aligned}$$

■

The correspondence between (agnostic) transductive error and vertices' outdegrees again justifies a focus on orientations of one-inclusion graphs that control nodes' outdegrees. Note, however, that degree requirements should no longer be uniform, as they were in Definition 41. Each node is instead judged on the basis of its outdegree minus the *credits* it receives as compensation for being distant from $\mathcal{H}$. This is formalized by the following definition.

Definition 62 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and $S \in \mathcal{X}^n$. We say that $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ is $(\alpha, \mathcal{H})$-**agnostic-orientable** if it can be oriented so that all its vertices v have indegrees at least $\alpha - \|v - \mathcal{H}\|_0$. Similarly, G is $(\alpha, \mathcal{H})$-**agnostic-coorientable** if it can be oriented so that all its vertices v have outdegrees at most $\alpha + \|v - \mathcal{H}\|_0$. We will suppress $\mathcal{H}$ when it is clear from context, and write simply α-agnostic-(co)orientable.*

Lemma 63 *Let A be a learner for $\mathcal{H}$. The following conditions are equivalent.*

1. *A incurs agnostic transductive error at most ϵ on all samples of size n.*
2. *For each $h \in \mathcal{H}$ and $S \in \mathcal{X}^n$, A induces an $(n\epsilon)$-agnostic-coorientation on $G_{\mathrm{Ag}}(\mathcal{H}|_S)$.*
3. *For each $h \in \mathcal{H}$ and $S \in \mathcal{X}^n$, A induces an $((1-\epsilon) \cdot n)$-agnostic-orientation on $G_{\mathrm{Ag}}(\mathcal{H}|_S)$.*

Proof Conditions (2.) and (3.) are equivalent as a consequence of each node in the undirected graph $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ having degree exactly $n = |S|$, one for each point in S that can be omitted. In particular, for any fixed orientation of $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ we have:

$$\begin{aligned} (2.) \quad &\equiv \quad \mathrm{outDeg}(v) \leq n \cdot \epsilon + \|v - \mathcal{H}\|_0 \\ &\iff \quad n - \mathrm{inDeg}(v) \leq n \cdot \epsilon + \|v - \mathcal{H}\|_0 \\ &\iff \quad \mathrm{inDeg}(v) \geq n(1-\epsilon) - \|v - \mathcal{H}\|_0 \\ &\equiv \quad (3.). \end{aligned}$$

The equivalence between (1.) and (2.) follows immediately from Lemma 61, applied pointwise to all $S \in (\mathcal{X} \times \mathcal{Y})^n$. ■

We now generalize the Hall density and Hall complexity of Section 3, describing agnostic analogues that provide an exact combinatorial characterization of the optimal agnostic transductive error that can be attained on samples of size n. (That is, of the transductive error rate of $\mathcal{H}$.) The central ingredient is to incorporate the non-uniform degree requirements into the Hall arguments of Section 3. Crucially, Hall's theorem and its generalizations are robust to non-uniform degree requirements, allowing us to transfer our reasoning from Section 3 in a relatively routine manner.

Definition 64 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, $S \in \mathcal{X}^n$ a sequence of unlabeled datapoints, and $G_{\text{Ag}}(\mathcal{H}|_S) = (V, E)$ the agnostic one-inclusion graph of $\mathcal{H}$ with respect to S. For a set of nodes $U \subseteq V$, let $E[U] \subseteq E$ denote the collection of edges with at least one incident node in U. The* ***agnostic Hall density*** *of $G_{\text{Ag}}(\mathcal{H}|_S)$ is*

$$\begin{aligned}\mathsf{Hall}^{\text{Ag}}(G) &= \inf_{\substack{U \subseteq V, \\ |U| < \infty}} \frac{|E[U]| + \|U - \mathcal{H}\|_0}{|U|} \\ &:= \inf_{\substack{U \subseteq V, \\ |U| < \infty}} \frac{|E[U]| + \sum_{u \in U} ||u - \mathcal{H}||_0}{|U|}.\end{aligned}$$

Proposition 65 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class, $S \in \mathcal{X}^n$ a sequence of unlabeled datapoints, and $G_{\text{Ag}}(\mathcal{H}|_S)$ the agnostic one-inclusion graph of $\mathcal{H}$ on S. Then $\mathsf{Hall}^{\text{Ag}}\big(G_{\text{Ag}}(\mathcal{H}|_S)\big)$ is the greatest α for which $G_{\text{Ag}}(\mathcal{H}|_S)$ is α-agnostic-orientable.*

Proof The claim follows immediately from the proof of Proposition 44 and the observation that Hall's theorem is robust to non-uniform degree requirements. That is, an α-agnostic-orientation of $G_{\text{Ag}}(\mathcal{H}|_S) = (V, E)$ is a function $E \to V$ such that each edge is assigned to an incident node and each $v \in V$ receives at least $\alpha - ||v - \mathcal{H}||_0$ edges. In order for such a function to exist, it is clearly necessary that for each finite U, $|E[U]| \geq |U| \cdot \alpha - ||U - \mathcal{H}||_0$. The classical statement of Hall's theorem — along with a cloning argument to handle non-uniform degree requirements — demonstrates sufficiency when $G_{\text{Ag}}(\mathcal{H}|_S)$ is finite (Hall, 1935). When $G_{\text{Ag}}(\mathcal{H}|_S)$ is infinite, sufficiency follows from the generalization of Hall's theorem to infinite bipartite graphs in which all nodes on the right side have finite degree (and the recollection that each node in $G_{\text{Ag}}(\mathcal{H}|_S)$ has degree $|S|$) (Hall Jr, 1948). See the proof of Proposition 44 for further detail on the splitting argument. ■

Definition 66 *The* **agnostic Hall complexity** *of a hypothesis class $\mathcal{H}$ is the function $\pi^{\text{Ag}}_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ defined:*

$$\pi^{\text{Ag}}_{\mathcal{H}}(n) = \max_{S \in (\mathcal{X} \times \mathcal{Y})^n} n - \mathsf{Hall}^{\text{Ag}}(G_{\text{Ag}}(\mathcal{H}|_S)).$$

We now demonstrate that the agnostic Hall complexity exactly characterizes the agnostic transductive error rate, i.e., the transductive error attained by an optimal learner.

Proposition 67 *Fix any domain $\mathcal{X}$, label set $\mathcal{Y}$, and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Then $\epsilon^{\mathrm{Ag}}_{\mathcal{H}}(n) = \frac{\pi^{\mathrm{Ag}}_{\mathcal{H}}(n)}{n}$ for all $n \in \mathbb{N}$.*

Proof First note that $\pi^{\mathrm{Ag}}_{\mathcal{H}}(n)$ is the minimal α for which all $\{G(\mathcal{H}|_S)\}_{S \in \mathcal{X}^n}$ are α-coorientable, by Proposition 65 and the equivalence between conditions (2.) and (3.) of Lemma 63. Invoke the equivalence between conditions (1.) and (2.) of Lemma 63 to complete the proof. ■

E.3. Equivalence of Errors and Orienting the Agnostic OIG

A central feature of our work in the realizable case was the equivalence between high-probability, expected, and transductive errors, as established in Section 2.3. In particular, it permitted us to freely restrict focus to one-inclusion graphs and their optimal orientations, which are suited to the minimization of transductive error. For learning in the agnostic case, however, the equivalence between errors is not nearly as tight. Informally, the sensitivity of agnostic errors to multiplicative factors renders ineffective many of the arguments from the realizable case. (E.g., doubling the error of a learner for the realizable case is a benign operation, but lethal for an agnostic learner subject to an additive constraint.)

Using different arguments, however, we are able to demonstrate that the sample complexities corresponding to the expected and transductive errors differ by a factor of at most $3/\epsilon$. Let us first state the necessary definitions.

Definition 68 *The* **agnostic transductive sample complexity** $m^{\mathrm{Ag}}_{\mathrm{Trans},A} : (0,1) \to \mathbb{N}$ *of a learner A is the function mapping δ to the minimal m for which $\epsilon^{\mathrm{Ag}}_{A,\mathcal{H}}(m') < \delta$ for all $m' \geq m$. That is,*

$$m^{\mathrm{Ag}}_{\mathrm{Trans},A}(\delta) = \min\{m \in \mathbb{N} \,:\, \epsilon^{\mathrm{Ag}}_{A,\mathcal{H}}(m') < \delta,\ \forall m' \geq m\}.$$

The agnostic transductive sample complexity *of a hypothesis class $\mathcal{H}$ is the pointwise minimal sample complexity attained by any of its learners, i.e.,*

$$m^{\mathrm{Ag}}_{\mathrm{Trans},\mathcal{H}}(\epsilon) = \min_A m^{\mathrm{Ag}}_{\mathrm{Trans},A}(\epsilon),$$

where A ranges over all learners for $\mathcal{H}$.

Proposition 69 *Let $\mathcal{X}$ be an arbitrary domain, $\mathcal{Y}$ a finite label set, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. Then $m^{\mathrm{Ag}}_{\mathrm{Exp},\mathcal{H}}(\epsilon) \leq m^{\mathrm{Ag}}_{\mathrm{Trans},\mathcal{H}}(\epsilon) \leq m^{\mathrm{Ag}}_{\mathrm{Exp},\mathcal{H}}(\epsilon/2) \cdot \frac{3}{\epsilon}$.*

Proof In pursuit of the first inequality, let $n = m^{\mathrm{Ag}}_{\mathrm{Trans},\mathcal{H}}(\epsilon)$ and let A be a learner for $\mathcal{H}$ attaining this transductive error guarantee. Fix a distribution D over $\mathcal{X} \times \mathcal{Y}$. We show that A attains favorable

expected error on samples of size $n-1$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim D^{n-1}} L_D\big(A(S)\big) &= \mathop{\mathbb{E}}_{\substack{S\sim D^{n-1},\\(x,y)\sim D}} \big[A(S)(x)\neq y\big] \\
&= \mathop{\mathbb{E}}_{S\sim D^{n}} \big[A(S_{-n})(x_n)\neq y_n\big] \\
&= \mathop{\mathbb{E}}_{S\sim D^{n}} \mathop{\mathbb{E}}_{i\in[n]} \big[A(S_{-i})(x_i)\neq y_i\big] \\
&= \mathop{\mathbb{E}}_{S\sim D^{n}} \frac{1}{n}\operatorname{outDeg}(S) \\
&\leq \mathop{\mathbb{E}}_{S\sim D^{m}} \left(\epsilon + \frac{1}{n}\min_{h\in\mathcal{H}} \Big|\Big|S-h|_S\Big|\Big|_0\right) \\
&\leq \epsilon + \inf_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{S\sim D^{m}} \frac{1}{n}\Big|\Big|S-h|_S\Big|\Big|_0 \\
&= \epsilon + \inf_{h\in\mathcal{H}} L_D(h)
\end{aligned}$$

Conversely, let A be a learner attaining agnostic expected error at most ϵ on samples of size $\geq n$. We will extract from A a learner attaining agnostic transductive error at most 2ϵ on samples of size $n' = \frac{3n}{\epsilon}$. Fix an $S\in\mathcal{X}^{n'}$; we will design an $(n'\cdot\epsilon)$-agnostic-orientation of $G_{\mathrm{Ag}}(\mathcal{H}|_S)=(V,E)$. In fact, we will design a fractional orientation, i.e., an assignment from each edge $e\in E$ to several of its incident vertices, in non-negative amounts summing to 1.

Let e be incident to vertices $V_e=\{v_1,\ldots,v_k\}$. Let $x'\in\mathcal{X}$ be the unique datapoint on which the vertices V_e disagree, when thought of as functions $S\to\mathcal{Y}$. Furthermore, let D_{v_i} be the uniform distribution over the entries of v_i, thought of as a sequence of labeled datapoints. For vertices v_i, v_j and $S\sim D_{v_i}$, let $E_{x'}$ denote the event that S does not contain x'.

We now define our fractional orientation by assigning p_i units of e to v_i, where

$$p_i = \mathop{\mathbb{P}}_{S\sim D^n_{v_i}}\big(A(S)(x')=v_i(x') \;\big|\; E_{x'}\big).$$

Note that $\sum_i p_i = 1$ as a consequence of the fact that $A(S)(x')\in\mathcal{Y}=\{v_1(x'),\ldots,v_k(x')\}$. Now fix an arbitrary vertex $v\in G_{\mathrm{Ag}}(\mathcal{H}|_S)$. Let $N(v)$ denote the set of edges incident to v, corresponding to some datapoint x'_e on which v disagrees with the other nodes incident to e. Then,

$$\begin{aligned}
\frac{\operatorname{outDeg}(v)}{n} &= \frac{1}{n}\sum_{e\in N(v)} \mathop{\mathbb{P}}_{S\sim D^n_v}\big(A(S)(x'_e)\neq v(x'_e) \;\big|\; E_{x'_e}\big) \\
&\leq \frac{1}{\mathbb{P}(E_{x'_e})}\cdot\frac{1}{n}\sum_{e\in N(v)} \mathop{\mathbb{P}}_{S\sim D^n_v}\big(A(S)(x'_e)\neq v(x'_e)\big) \\
&\leq \left(1+\frac{\epsilon}{2}\right)\cdot \mathop{\mathbb{E}}_{S\sim D^n_v} L_{D_v}\big(A(S)\big) \\
&\leq \left(1+\frac{\epsilon}{2}\right)\cdot\left(\epsilon + \inf_{h\in\mathcal{H}} L_{D_v}(h)\right) \\
&\leq 2\epsilon + \inf_{h\in\mathcal{H}} L_{D_v}(h) \\
&= 2\epsilon + \inf_{h\in\mathcal{H}} ||h-v||_0.
\end{aligned}$$

In the third line, we make use of the fact that

$$\mathop{\mathbb{P}}_{S\sim D_v^n}(E_{x'_e}) = \left(1 - \frac{1}{n'}\right)^n \geq 1 - \frac{n}{n'} \geq 1 - \frac{\epsilon}{3}.$$

The second-to-last line uses the fact that our loss function is bounded above by 1, and thus $\inf_{h\in\mathcal{H}} L_{D_v}(h)$ as well. ■

Given the importance of agnostic OIGs, it is once again natural to ask for characterizations of — and computational insights on — their optimals orientations. We now demonstrate that the maximum entropy learner introduced in Section 4.3 generalizes to the agnostic case with analagous guarantees.

Proposition 70 *Let $\mathcal{X}$ be a domain, $\mathcal{Y}$ a finite label set, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. Recall the optimal agnostic transductive error rate $\epsilon_{\mathcal{H}}^{\mathrm{Ag}}(n) = \frac{\pi_{\mathcal{H}}^{\mathrm{Ag}}(n)}{n}$. For any $S \in \mathcal{X}^n$, let $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ be the agnostic one-inclusion graph of $\mathcal{H}$ on S and $\mathcal{D}$ the collection of all orientations in $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ (equivalently, the collection of assignments in the bipartite analogue $G_{\mathrm{BP}}^{\mathrm{Ag}} = (\mathcal{A}, \mathcal{B}, E)$, where $\mathcal{A}$ denotes partially labeled datasets and $\mathcal{B}$ fully labeled datasets[7].).*

Then there is a unique distribution over assignments $P^* \in \Delta_{\mathcal{D}}$ such that each $v \in \mathcal{B}$ receives at least $n - \pi_{\mathcal{H}}^{\mathrm{Ag}}(n) - \|v - \mathcal{H}\|_0$ in-degree in expectation, and the following conditions hold simultaneously:

1. *P^* achieves maximal entropy among all $P \in \Delta_{\mathcal{D}}$ satisfying the in-degree requirements.*
2. *P^* corresponds to a randomized transductive learner $\mathcal{A}$ which, in expectation over its internal randomness, attains optimal error rate $\epsilon_{\mathcal{H}}^{\mathrm{Ag}}(n)$.*
3. *$\mathcal{A}$ can be described as follows:*
   - *Upon receiving the unlabeled datapoints $S_{\mathcal{X}}^+ = (x_1, \ldots, x_n)$, including the test point, construct an appropriate prior distribution ρ over $\mathcal{H}|_{S_{\mathcal{X}}^+}$.*
   - *Given the index i of the test point, and labels y_j for all datapoints $x_{j\neq i}$, apply a Bayes update to ρ in order obtain a posterior ρ'. This posterior corresponds to restricting the prior to hypotheses consistent with the provided labels, and rescaling accordingly.*
   - *Sample a hypothesis h from ρ', and output $h(x_i)$ as the prediction for the label of x_i.*

Proof To prove (1.) and (2.), we define a feasible maximum entropy convex program which finds a randomized assignment equivalent to an $(n - \pi_{\mathcal{H}}^{\mathrm{Ag}}(n))$-agnostic-orientation. The program is identical to `MaxEnt` defined in Proposition 53, but rather than having a fixed c lower bound on the indegree for each fully labeled dataset $v \in \mathcal{B}$, we have a varying c_v on the RHS constraint in the

7. I.e., the edge-vertex incidence graph of $G_{\mathrm{Ag}}(\mathcal{H}|_S)$. See the analagous Definition 37

convex program.

$$\begin{aligned}
\max_{p_d:\, d\in\mathcal{D}} \quad & \sum_{d\in\mathcal{D}} p_d \ln \frac{1}{p_d} \\
\text{s.t.} \quad & \sum_{d\in\mathcal{D}} p_d \sum_{u\in\mathcal{A}} d(u,v) \geq c_v = n - \pi^{\mathrm{Ag}}_{\mathcal{H}}(n) - \|v - \mathcal{H}\|_0 && \forall v \in \mathcal{B} \\
& \sum_{d\in\mathcal{D}} p_d = 1 \\
& p_d \geq 0 && \forall d \in \mathcal{D}
\end{aligned}$$

Feasibility of this program holds due to the fact that $\epsilon^{\mathrm{Ag}}_{\mathcal{H}}(n)$ is the optimal error rate, and that by Lemma 63 it therefore corresponds to a $(n - \pi^{\mathrm{Ag}}_{\mathcal{H}}(n))$-agnostic-orientation. Uniqueness follows from strict concavity of the objective. These suffice to prove (1.) and (2.).

To address (3.), we need to again take the dual of the program. Strong duality still holds from the weak version of Slater's condition, as the constraints are still affine. We omit the full dual derivation as it is straightforward and similar to that of `MaxEnt` from the proof of Theorem 21 in Appendix D.3, replacing c with c_v everywhere. Since c_v falls away when we take the partial derivative w.r.t. p_d or z, the rest of the dual derivation is identical. Therefore, (3.) follows from the proof of Theorem 21, which makes no use of c.

Note that both the number of edges and nodes will be much larger in the agnostic Hamming graph $G_{\mathrm{Ag}}(\mathcal{H}|_S)$ than the realizable OIG $G(\mathcal{H}|_S)$. The number of assignments in their bipartite counterparts will reflect this. Although the dual of the agnostic and realizable max entropy programs may appear to take the same value for any fixed assignment d, since the dual does not have a dependence on c, the agnostic max entropy convex program primal actually has a much larger number of primal and dual variables which will impact their associated optimal values. ■

Proposition 70 demonstrates that the randomized learner from the agnostic case can be viewed as Bayesian, just as the randomized learner from the realizable case. Furthermore, this indicates the existence of an SRM for randomized agnostic learners, as discussed in the closing discussion of Section 4.3 for the realizable setting. In particular, the associated regularizer tracks KL divergence from the prior ρ defined in (3.) of Proposition 70.

Appendix F. Related Work

Learnability and optimal learning rates. The importance of learners beyond the classical paradigm of empirical risk minimization (ERM) was discussed by Shalev-Shwartz et al. (2010), who exhibit a learnable class that is not learnable by any ERM learner. The authors propose a notion of *stable* learning rules, which they demonstrate characterizes learnability in this setting. Notably, however, they work in Vapnik's general learning setting, which generalizes Valiant's PAC framework in some respects but restricts learners to be proper. Consequently, it would deem as unlearnable such problems as (Daniely and Shalev-Shwartz, 2014, Theorem 1), a learnable multiclass problem whose learners are all improper. These problems are of central interest to us, and thus the proper learning techniques for achieving stability from Shalev-Shwartz et al. (2010) do not resolve our central inquiry.

There is a rich history of related work for one-inclusion graphs, commencing with the seminal work of Haussler et al. (1994), who used the OIG algorithm of Alon et al. (1987) to obtain error guarantees for VC classes. Subsequently, many follow-up works have studied OIGs and their error. Rubinstein et al. (2009) propose a combinatorial technique called *shifting* to obtain better (sub)graph density bounds for the OIG, and by extension, better mistake bounds. Within their work, they make use of Hall's theorem, but fall short of characterizing the transductive error exactly with any notion similar to the Hall complexity which we propose in Section 3. We note, however, that they were indeed the first to observe the connection between the OIG, the bipartite OIG variant, and Hall's theorem, and as such do not make the claim that our perspective is completely novel.

Our generalized one-inclusion graph proposed in Appendix E is designed to capture learning in the agnostic setting, by extending the vertex set to include all vertices in the *Hamming graph*, not only those labelings of sample points realizable by an $h \in \mathcal{H}$. This modification closely resembles that proposed by Long (1998) for the case of binary classification, though they study realizable learning with distribution shift (rather than learning in the agnostic case). Daniely and Shalev-Shwartz (2014) made several notable advances in the understanding of one-inclusion graphs for multiclass learning, including by exhibiting a learnable problem without any proper learners, improving the analysis of errors incurred by optimal OIG learners, and introducing the DS dimension for measuring the complexity of a hypothesis class. Recently, the breakthrough result of Brukhim et al. (2022) used OIGs to prove that the DS dimension indeed characterizes PAC learnability for multiclass classification over arbitrary label sets. They also demonstrate that the related Natarajan dimension, in contrast, does not characterize learnability.

OIGs are also a key ingredient in the proof of learnability for *partial* concept classes, as studied in Alon et al. (2022); Kalavasis et al. (2022). More recently, Aden-Ali et al. (2023b) demonstrated that optimal orientations of OIGs, despite attaining optimal transductive error, do not necessarily attain optimal PAC error, resolving in the negative a conjecture of Warmuth (2004). The same authors demonstrate in Aden-Ali et al. (2023a) that although OIGs are not sample optimal in the PAC model, a simple aggregation of OIGs is optimal for multiclass classification over finite label sets. In realizable regression, the recent work of Attias et al. (2023) employed OIGs to define the scaled γ-OIG dimension and demonstrate that it characterizes learnability (unlike the fat shattering dimension). In robust learning, Montasser et al. (2022) designed an optimal learner using their *global one-inclusion graph*.

Regularization. Trading off empirical risk with a notion of model complexity harks back to at least the work of Tikhonov (1943). Structural risk minimization, the formalization of this notion within the statistical learning theory community, is usually credited to the celebrated work of Vapnik and Chervonenkis (1974). There is a large body of work examining how regularizers can impact the speed and stability across learning and optimization (see Zhou et al. (2024); Rosset et al. (2007); Lee et al. (2010); Sridharan et al. (2008) and references therein). More recently, there is good reason to believe that popular algorithms such as gradient descent, when applied on neural networks, act as implicit (and perhaps *data-dependant*) regularizers (Smith et al., 2021; Neyshabur et al., 2017).

Our local unsupervised regularizer, however, is unusual in that it can be thought of as an *unsupervised pre-training* algorithm in the transductive setting, which first examines only the unlabeled datapoints in the training set (including the test point), and then uses this to construct a regularizer with which to perform SRM. The connection between regularization and unsupervised pre-training was proposed at least as far back as Erhan et al. (2010). There, the authors demonstrate empirically

that in the context of deep learning (as it existed in 2010), unsupervised pre-training can be thought of as an implicit form of regularization through initialization.

Unsupervised pre-training has also seen a reasonable amount of practical success in domains such as computer vision (Carreira et al., 2016; Chen et al., 2017) and natural language processing (Radford et al., 2018). On the theoretical side, Ge et al. (2023) perform a study of unsupervised pre-training in which they assume the existence of an underlying latent variable model, and perform maximum likelihood parameter estimation as the unsupervised step. They then perform empirical risk minimization with the pre-trained model. This setup is somewhat similar in flavor to our algorithms, where the local unsupervised regularizers can be viewed as a form of unsupervised pre-training, and where we perform ERM on the training data plus regularizer. However, their setup generally differs from ours and they do not focus on characterizing the learnability of multiclass problems. While there is a modest amount of attention from the community in understanding theoretical properties of unsupervised pre-training as viewed through the lens of self-supervised learning (Lee et al., 2021; HaoChen et al., 2021) — especially as it relates to language models (Saunshi et al., 2021) — this work usually does not take place in the fundamental setting of supervised multiclass learning. Furthermore, unsupervised pre-training usually employs separate datasets for the supervised and unsupervised training phases, whereas our unsupervised regularizer employs the same dataset for both phases of training.

Perhaps most related to our formalization of regularizers from the perspective of the theory community is the work of Hopkins et al. (2022), who consider the task of extending arbitrary realizable learners into learners for the agnostic case. In the context of our framework, the extension they provide can be seen as a type of regularization (though not described as so in their work). In particular, their recipe for transforming realizable learners into agnostic learners can be seen as using an unsupervised regularizer in order to restrict focus to a collection of certain hypotheses, on which it then performs empirical risk minimization. Note that restricting focus to certain hypotheses can be implemented as a "hard" regularizer assigning value ∞ to the omitted hypotheses and value zero to the others. This deviates, however, from our setting in several important respects. First, the predictors to which this procedure restricts focus are only elements of $\mathcal{H}$ if the realizable learner used as input in the reduction is a proper learner. (And, as we have seen, there exist learnable multiclass problems without any proper learners.) Secondly, the technique uses distinct datasets for regularization and minimization of empirical risk, in contrast to our transductive setting. Lastly, and most notably, the result relies on one's being supplied a realizable learner to begin with, whereas we are primarily concerned with the design of learners "from scratch."