

A Notion of Task Relatedness Yielding Provable Multiple-Task Learning Guarantees

Shai Ben-David¹ and Reba Schuller Borbely²

¹ David, R., Cheriton School of Computer Science,
University of Waterloo,
Waterloo, ON, N2L 1G3 Canada

shai@cs.uwaterloo.ca,

² reba_schuller@hotmail.com

Summary. The approach of learning of multiple “related” tasks simultaneously has proven quite successful in practice; however, theoretical justification for this success has remained elusive. The starting point for previous work on multiple task learning has been that the tasks to be learned jointly are somehow “algorithmically related”, in the sense that the *results* of applying a specific learning algorithm to these tasks are assumed to be similar. We offer an alternative approach, defining relatedness of tasks on the basis of similarity between the example generating distributions that underlie these tasks.

We provide a formal framework for this notion of task relatedness, which captures a sub-domain of the wide scope of issues in which one may apply a multiple task learning approach. Our notion of task similarity is relevant to a variety of real life multitask learning scenarios and allows the formal derivation of generalization bounds that are strictly stronger than the previously known bounds for both the learning-to-learn and the multitask learning scenarios. We give precise conditions under which our bounds guarantee generalization on the basis of smaller sample sizes than the standard single-task approach³.

1.1 Introduction

Most of the work in machine learning focuses on learning tasks that are encountered separately, one task at a time. While great success has been achieved in this type of framework, it is clear that it neglects certain fundamental aspects of human and animal learning. Human beings face each new learning task equipped with knowledge gained from previous similar learning tasks.

³ A preliminary version of this paper appears in the proceedings of COLT’03, [BDS03]

Furthermore, human learning frequently involves approaching several learning tasks simultaneously; in particular, humans take advantage of the opportunity to compare and contrast similar categories in learning to classify entities into those categories.

It is natural to attempt to apply these observations to machine learning—what kind of advantage is there in setting a learner to work on several tasks sequentially or simultaneously? Intuitively, there should certainly be some advantage, especially if the tasks are closely related in some way. And, indeed, much experimental work [Bax95, IE96, Thr96, Hes98, Car97] has validated this intuition. However, thus far, there has been relatively little progress on theoretical justification for these results.

Relatedness of tasks is key to the multitask learning (MTL) approach. Obviously, one cannot expect that information gathered through the learning of a set of tasks will be relevant to the learning of another task that has nothing in common with the already learned tasks.

Previous work on MTL, or *Learning to Learn*, treated the notion of relatedness using a 'functional' approach. For example, consider one of the more systematic theoretical analysis of a simultaneous learning model to date, Baxter's Learning To Learn work, e.g., [Bax00]. In Baxter's work the similarity between jointly learned tasks is manifested solely through a model selection criterion. Namely, the advantage of learning tasks together relies on the assumption that the tasks share a common optimal inductive bias, reflected by a common optimal (or near-optimal) hypothesis class.

We try to determine under what circumstances one can expect different tasks to be related in a 'learning useful' way. We focus on the sample generating distributions underlying the learning tasks, and define task relatedness as an explicit relationship between these distributions. Our notion seems to capture a sub-domain of the realm of applications to which multi-task learning may be relevant.

Not surprisingly, by limiting the discussion to problems that can be modeled by our data generating mechanism we leave many potential MTL scenarios outside the scope of our discussion. However, there are several interesting problems that can be treated within our framework. For these problems we can reap the benefits of having a mathematical notion of relatedness and prove sample size upper bounds for MTL learning that are better than any previous proven bounds.

The rest of the paper is organized as follows: Section 1.2 formally introduces multiple task learning and describes our notion of task relatedness. We state and prove our generalization error bound for this framework in section 1.3. In section 1.6, we analyze the generalized VC-dimension parameter on which this bound depends, and we compare this bound for multiple task learning to the analogous bounds for the single task approach. That is, we examine when can the error bounds for learning a given task improve by allowing the learner to access samples generated by different but related tasks.

1.1.1 Previous Work

The only theoretical analysis of multitask learning that we are aware of is the work of Baxter [Bax00], and the recent work of Ben-David et. al. [BDGS02].

The main question that we are interested in is when does multitask learning provide an advantage over the single task approach. In order to achieve this, we introduce a concrete notion of what it means for tasks to be "related," and evaluate multi- versus single-task learning for tasks related in this manner. Our notion of relatedness between tasks is inspired by [BDGS02] which deals with the problem of integrating disparate databases. We extend the main generalization error result from [BDGS02] to the multitask learning setting, strengthen it, and analyze the extent to which it offers an improvement over single task learning.

The main technical tool that we use is the generalized VC-dimension of Baxter [Bax00]. Baxter applies his version of VC-dimension to bound the *average* error of a set of predictors over a class of tasks in terms of the average empirical error of these predictors. In contrast with Baxter's analysis, we view multitask learning as having one 'focus of interest' task that one wishes to learn and view the extra related tasks as just an aid towards learning the main task. In this context, bounds on the average error over all tasks are not good enough. We show that when one is dealing with tasks that are related in the sense that we define, the Baxter generalization bound can be strengthened to hold for the error of each single task.

We should point out the distinction between the problem considered herein and the co-training approach of [BM98]. Co-training makes use of extra "tasks" to compensate for having only a small amount of labeled data. However, in co-training, the extra tasks are assumed to be different "views" of the *same* sample, whereas our extra tasks are independent samples from different distributions. Thus, despite its relevance to multitask learning, previous work on co-training cannot be directly applied to the problem at hand.

1.2 A Data Generation Model for Related Tasks

Formally, the typical (single-task) classification learning problem is modeled as follows: Given a domain \mathcal{X} and a random sample S drawn from some unknown distribution P on $\mathcal{X} \times \{0, 1\}$, find a hypothesis $h : \mathcal{X} \rightarrow \{0, 1\}$ such that for randomly drawn (x, b) , with high probability $h(x) = b$. This problem is some times referred to as "statistical regression".

The multiple task learning problem is the analogous problem for multiple distributions. However, the focus is on the potential advantage to each learning task from the data available for the other tasks. Given domain \mathcal{X} and unknown distributions P_0, \dots, P_n over $\mathcal{X} \times \{0, 1\}$, a learner is presented with a sequence of random samples S_0, \dots, S_n drawn from these P_i 's respectively, and has to come up with a hypothesis $h : \mathcal{X} \rightarrow \{0, 1\}$ such that, for (x, b) drawn randomly

from P_0 , $h(x) = b$ with high probability. What we focus on is the extent to which the samples S_i , for $i \neq 0$ be utilized to help find a good hypothesis for predicting the labels of P_0 .

As we have mentioned previously, it is intuitive that the benefit of having access to samples from multiple tasks depends on the "relatedness" between the different tasks. While there has been empirical success with sets of tasks related in various ways, thus far, no formal definition of "relatedness" has yielded any theoretical results to that effect.

1.2.1 Our Notion of Relatedness Between Learning Tasks

We define a data generation mechanism which serves to determine our notion of related tasks.

The basic ingredient in our definition is a set \mathcal{F} of transformations $f : \mathcal{X} \rightarrow \mathcal{X}$. We say that tasks are \mathcal{F} -related if, for some fixed probability distribution over $\mathcal{X} \times \{0, 1\}$, the data in each of these tasks is generated by applying some $f \in \mathcal{F}$ to that distribution. The next definition formalizes this notion.

Definition 1. For a measure space $(\mathcal{X}, \mathcal{A})$, where \mathcal{X} denotes a domain set, and \mathcal{A} is a σ -algebra of its subsets, we discuss probability distributions, P over $\mathcal{X} \times \{0, 1\}$, for which the P -measurable sets are the σ -algebra generated by the sets of the form $A \times B$, for $A \in \mathcal{A}$ and $B \subseteq \{0, 1\}$.

- For a function, $f : \mathcal{X} \rightarrow \mathcal{X}$, let $f[\mathcal{A}]$ be $\{A \subseteq \mathcal{X} : f^{-1}(A) \in \mathcal{A}\}$, and let $f[P]$ be the probability distribution over $\mathcal{X} \times \{0, 1\}$ defined by having the probability distribution $f[P]$ assign to a set $T \subseteq \mathcal{X} \times \{0, 1\}$, the probability $f[P](T) = P(\{(f(x), b) \mid (x, b) \in T\})$.

Let \mathcal{F} be a set of transformations $f : \mathcal{X} \rightarrow \mathcal{X}$, and let P_1, P_2 be probability distributions over $\mathcal{X} \times \{0, 1\}$.

- We say that P_1, P_2 are \mathcal{F} -related if there exists some $f \in \mathcal{F}$ such that $P_1 = f[P_2]$ or $P_2 = f[P_1]$.
- We say that two samples are \mathcal{F} -related if they are samples from \mathcal{F} -related distributions.

In our learning scenario, we assume the data (both the training samples and the test examples) are generated by some probability distributions, $\{P_i : i \leq n\}$ that are (pairwise) \mathcal{F} -related. We assume that the learner knows the set of indices of the distributions, $\{1, \dots, n\}$ and the family of functions \mathcal{F} but does not know the data-generating distributions nor which specific function f relates any given pair of distributions. As input, the learner gets samples, $\{S_i : i \leq n\}$, each S_i drawn i.i.d. from P_i . Consequently, the advantage that a learner can derive for a specific task from access to a sample drawn from some other \mathcal{F} -related task depends of the richness of the family of transformations \mathcal{F} . The larger this set gets, the looser the notion of \mathcal{F} -relatedness is.

Clearly there are many examples of potential applications of simultaneous learning that do not fit into this model of relatedness. However, there are various interesting examples where this notion seems to provide a satisfactory mathematical model of the similarity between the tasks in a set of related learning problems.

Our framework applies to scenarios in which the learner’s prior knowledge includes knowledge of some family \mathcal{F} of transformations, such that all the tasks for which this MTL learning approach will be applied are mutually \mathcal{F} -related. One domain in which such \mathcal{F} -relatedness prior knowledge may be applied is in situations where many different sensors collect data for the same classification problem. For example, consider a set of cameras located in the lobby of some high security building. Assume that they are all used to automatically detect unauthorized visitors, based on the images they record. Clearly, each of these cameras has its own bias, due to a different height, light conditions, angle, etc. While it may be difficult to determine the exact bias of each camera, it may be feasible to define mathematically a set of image transformations \mathcal{F} such that the data distributions of images collected by of all these recorders are \mathcal{F} -related. Another area in which such a notion of similarity is applicable is that of database integration. Suppose there are several databases available, each of which obtains its information from the same data pool, yet represents its information with a different database schema. For the purpose of classification prediction, our results in the next section eliminate the need for the difficult undertaking of database integration, treating each database as one task in a multiple task learning problem.

1.3 Learning \mathcal{F} -Related Tasks

In this section, we analyze multiple task learning for \mathcal{F} -related tasks. Our main idea is to separate the information contained in the training data into information that is *invariant* under transformations from \mathcal{F} and data that is \mathcal{F} -sensitive. We utilize the training samples from the extra tasks to learn the \mathcal{F} -invariant aspects of the predictor. For example, if our domain is the two dimensional Euclidean space, and \mathcal{F} is a family of isometries of the plane, then geometric shapes are \mathcal{F} -invariant and can be deduced from translated images of the original data distributions, while their location in the plane is \mathcal{F} -sensitive. We formalize this basic idea in a way that allows precise quantification of the potential benefits to be drawn from such multi-task training data. We derive sample complexity bounds that demonstrate the merits of this algorithmic approach.

We formalize our notion of \mathcal{F} -relevant information through an appropriate partitioning of the learner’s set of potential label predictors. Given hypothesis space, \mathbb{H} , we create a family, \mathcal{H} , of hypothesis spaces consisting of sets of hypotheses in \mathbb{H} which are equivalent up to transformations in \mathcal{F} . We assume that \mathcal{F} forms a group under function composition and that \mathbb{H} is closed under

the action of \mathcal{F} . As is standard, we will write $[h]_{\sim_{\mathcal{F}}}$, or simply $[h]$, to denote the equivalence class of h under $\sim_{\mathcal{F}}$.

Definition 2. *Let \mathcal{F} be a set of transformations over a domain set \mathcal{X} , and let \mathbb{H} be a hypothesis space over that domain.*

- *We say that \mathcal{F} acts as a group over \mathbb{H} , if*
 1. *\mathbb{H} is closed under transformations from \mathcal{F} . Namely, for every $f \in \mathcal{F}$ and every $h \in \mathbb{H}$, $h \circ f \in \mathbb{H}$, and*
 2. *\mathcal{F} forms a group under function composition. Namely, \mathcal{F} is closed under transformation composition and inverses (for every $f, g \in \mathcal{F}$, the inverse transformation, f^{-1} , and the composition, $f \circ g$ are also members of \mathcal{F}).*
- *When \mathcal{F} acts as a group over \mathbb{H} , we define equivalence relation $\sim_{\mathcal{F}}$ on \mathbb{H} by:*

$$h_1 \sim_{\mathcal{F}} h_2 \text{ iff there exists } f \in \mathcal{F} \text{ such that } h_2 = h_1 \circ f.$$

We shall consider the family of hypothesis spaces, $\mathcal{H} = \{[h] : h \in \mathbb{H}\}$ - the family of all equivalence classes of \mathbb{H} under $\sim_{\mathcal{F}}$, (equivalently, $\mathcal{H} = \mathbb{H} / \sim_{\mathcal{F}}$).

Our learning paradigm consists of two stages. In the first stage, the learner considers all of the sample sets and uses them to learn the aspects of the task that are invariant under \mathcal{F} . In our setting, this means finding a $\sim_{\mathcal{F}}$ equivalence class, $[h]$, that is best suited for our prediction. In the second stage, the learner considers only the training sample that comes from the distribution of the target task (say, P_1), to figure out which specific predictor $h' \in [h]$ to choose as its final hypothesis. The benefit from the extra tasks examples is therefore realized through the reduction of the hypotheses search space, from the original \mathbb{H} to the subset $[h]$. The smaller \mathcal{F} is, the smaller each $[h]$ will be (since $[h] = \{h \circ f : f \in \mathcal{F}\}$), and so the larger is the benefit of multitasking.

To make this outline more concrete, let us consider again the two dimensional Euclidean space as our domain, \mathcal{X} , and let \mathcal{F} be a family of isometries of the plane. Let \mathbb{H} be the class of all axis-aligned rectangles. In this case, by viewing examples generated by distributions P_i that are \mathcal{F} -related to some target P_1 , one can learn about the length and width of the best rectangle predictor, but not about its location in the plane. More formally, in this case, for every rectangle, h , its $\sim_{\mathcal{F}}$ -equivalence class is the set of all rectangles that are isomorphic to h . It is not hard to see that the VC-dimension of such a class is 3, which is lower than the VC-dimension of the class of all axis aligned rectangles in the plain (which is 4). In Section 1.5 analyze the VC-dimension of such classes in arbitrary Euclidean dimensions and observe a similar reduction (from $2d$ to $\lfloor \frac{3d}{2} \rfloor$). This reduction in the complexity of the hypothesis class is where we gain from having samples from extra tasks (i.e., extra \mathcal{F} -related distributions).

1.4 Generalization error bounds

Following standard notation, we denote the *true error*, and *empirical error*, respectively, of a hypothesis as follows. For distribution P ,

$$Er^P(h) = P(\{(x, b) \in \mathcal{X} \times \{0, 1\} : h(x) \neq b\}).$$

And for a sample, S , of points in $\mathcal{X} \times \{0, 1\}$,

$$\hat{Er}^S(h) = \frac{|\{(x, b) \in S : h(x) \neq b\}|}{|S|}.$$

Lemma 1. *Let $f : \mathcal{X} \rightarrow \mathcal{X}$ and let P_1 and P_2 be probability distributions over \mathcal{X} .*

For any hypothesis, $h : \mathcal{X} \rightarrow \{0, 1\}$, a probability distribution, P , over $\mathcal{X} \times \{0, 1\}$ and $f : \mathcal{X} \rightarrow \mathcal{X}$,

$$Er^{f[P]}(h \circ f) = Er^P(h). \quad (1.1)$$

This is an immediate consequence of the the definition of the image, $f[P]$ of a distribution P and of the error, $Er^P(h)$.

Using this fact, we can deduce that the equivalence classes of \mathbb{H} perform equally well on the different tasks in the following sense.

Definition 3. *For any hypothesis space, H , define*

$$Er^P(H) = \inf_{h \in H} Er^P(h).$$

Thus, we judge the performance of a hypothesis space on a given task by the performance of the best hypothesis in the space on that task.

Lemma 2. *Let P_1, P_2 be \mathcal{F} -related distributions and \mathcal{F} be a group under function composition. If H is closed under the action of \mathcal{F} then $Er^{P_1}(H) = Er^{P_2}(H)$.*

Proof. We need to show that

$$\inf_{h \in H} Er^{P_1}(h) = \inf_{h \in H} Er^{P_2}(h).$$

It suffices to show that for every $h \in H$ there exist $h', h'' \in H$ such that $Er^{P_2}(h') \leq Er^{P_1}(h)$ and $Er^{P_1}(h'') \leq Er^{P_2}(h)$.

Since P_1, P_2 be \mathcal{F} -related, and \mathcal{F} is a group (so each $f \in \mathcal{F}$ has its inverse there) there exist $f, f' \in \mathcal{F}$ such that $P_1 = f[P_2]$ and $P_2 = f'[P_1]$. Since H is closed under the action of \mathcal{F} , both $h' = h \circ f$, and $h'' = h' \circ f'$ are members of H . Applying Lemma 1, we get $Er^{P_2}(h') = Er^{P_1}(h)$ and $Er^{P_1}(h'') = Er^{P_2}(h)$, so we are done. \square

Before we continue, we require some background.

1.4.1 Background from Baxter [Bax00]

Baxter [Bax00] discusses the following problem. Given a set of tasks and a set of hypothesis spaces, choose the hypothesis space which performs best on the set of the tasks. He provides a bound for the generalization error for this problem in terms of a generalized VC-dimension parameter. In particular, he bounds the rate of convergence of the *average* (over all the tasks) of the empirical errors to the true error.

Baxter's generalization error bound depends on the following notion of generalized VC-dimension for families of hypothesis spaces.

Notation:

For function $g : Y \rightarrow Z$ and $\bar{y} = (y_1, \dots, y_n) \in Y^n$, $\bar{g}(\bar{y})$ will denote $(g(y_1), \dots, g(y_n)) \in Z^n$.

Definition 4. 1. Given a matrix of $m \times n$ domain points, for every hypothesis class $H \in \mathcal{H}$ we consider the collection of all the $\{0, 1\}$ matrices that can be generated by applying n hypotheses from H to the n rows of the matrix (respectively). Formally, denoting for each $i \leq n$, $\bar{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$,

$$H_{n,m}(\bar{x}_1, \dots, \bar{x}_n) = \left\{ \begin{bmatrix} h_1(x_{1,1}), h_1(x_{1,2}), \dots, h_1(x_{1,m}) \\ \vdots \\ h_n(x_{n,1}), h_n(x_{n,2}), \dots, h_n(x_{n,m}) \end{bmatrix} : h_1, \dots, h_n \in H \right\}.$$

2. For family \mathcal{H} of hypothesis spaces, we take the union, over all classes $H \in \mathcal{H}$, of these sets of $\{0, 1\}$ matrices, and count how many matrices are in that union. Finally, we take the maximum of that number over all possible choices of the underlying matrix of $m \times n$ domain points. Namely,

$$\Pi_{\mathcal{H}}(n, m) = \max_{\bar{x}_1, \dots, \bar{x}_n \in X^m} \left| \bigcup_{H \in \mathcal{H}} H_{n,m}(\bar{x}_1, \dots, \bar{x}_n) \right|.$$

Definition 5. $d_{\mathcal{H}}(n) = \max\{m : \Pi_{\mathcal{H}}(n, m) = 2^{nm}\}$.

The following statements follow directly from the above definitions:

Proposition 1. For every family of classes, \mathcal{H} , and for every n ,

1. $\sup\{\text{VC-dim}(H) : H \in \mathcal{H}\} \leq d_{\mathcal{H}}(n) \leq \text{VC-dim}(\bigcup\{H : H \in \mathcal{H}\})$.
2. In particular, if \mathcal{H} consists of just one class, $\mathcal{H} = \{H\}$, then, for every n , $d_{\mathcal{H}}(n) = \text{VC-dim}(H)$.
3. $d_{\mathcal{H}}(n+1) \leq d_{\mathcal{H}}(n)$.

We can now state the relevant result from [Bax00] on multitask learning, which appears as corollary 13 in [Bax00].⁴

⁴ Note that although [Bax00] only states that $\frac{1}{n} \sum_{i=1}^n \text{Er}^{P_i}(h_i) \leq \frac{1}{n} \sum_{i=1}^n \hat{\text{Er}}^{S_i}(h_i) + \epsilon$, it is clear from the proofs in [Bax00] that this stronger form holds.

Theorem 1. Let \mathcal{H} be any permissible boolean hypothesis space family⁵, and let S_1, \dots, S_n be a sequence of random samples from distributions P_1, \dots, P_n (respectively) on $\mathcal{X} \times \{0, 1\}$. If the number of examples, in each sample S_i satisfies

$$|S_i| \geq \frac{88}{\epsilon^2} \left[2d_{\mathcal{H}}(n) \log \frac{22}{\epsilon} + \frac{1}{n} \log \frac{4}{\delta} \right],$$

then with probability at least $1 - \delta$ (over the choice of S_1, \dots, S_n), for any $H \in \mathcal{H}$, and $h_1, \dots, h_n \in H$,

$$\left| \frac{1}{n} \sum_{i=1}^n Er^{P_i}(h_i) - \frac{1}{n} \sum_{i=1}^n \hat{Er}^{S_i}(h_i) \right| \leq \epsilon.$$

Note that this theorem only bounds the *average* generalization error over the different tasks.

1.4.2 Bounding the Generalization Error for *Each* Task

We are now ready to state and prove one of our main results, which gives an upper bound on the sample complexity of finding a $\sim_{\mathcal{F}}$ -equivalence class which is near-optimal for *each* of the tasks. This is significant, since the goal of multitask learning is to use extra tasks to improve performance on one particular task.

Theorem 2. Let \mathcal{F} be a family of domain transformations of some domain set \mathcal{X} and let \mathbb{H} be a family of binary valued function on that domain so that \mathcal{F} acts as a group over \mathbb{H} . For any $h \in \mathbb{H}$, let $[h]$ denote the equivalence class of h under the relation $\sim_{\mathcal{F}}$ (or, equivalently, the trajectory of h under the transformations of \mathcal{F}) and let $\mathcal{H} = \{[h] : h \in \mathbb{H}\}$. Let $P_1 \dots P_n$ be a set of \mathcal{F} -related probability distributions over $\mathcal{X} \times \{0, 1\}$ and let S_1, \dots, S_n be a sequence of random samples, each generated *i.i.d.* from the corresponding distribution P_i .

Then, if the number of examples, in each sample S_i satisfies

$$|S_i| \geq \frac{88}{\epsilon^2} \left[2d_{\mathcal{H}}(n) \log \frac{22}{\epsilon} + \frac{1}{n} \log \frac{4}{\delta} \right], \quad (1.2)$$

then with probability at least $1 - \delta$, for every $h \in \mathbb{H}$,

$$\left| Er^{P_1}([h]) - \inf_{h_1, \dots, h_n \in [h]} \frac{1}{n} \sum_{i=1}^n \hat{Er}^{S_i}(h_i) \right| \leq \epsilon.$$

⁵ Permissibility, introduced by Ben-David in [BEHW89] is a “weak measure-theoretic condition satisfied by almost all ‘real-world’ hypothesis space families” that is required for the VC type uniform convergence bounds to hold. Throughout this paper we shall assume that all our classes are permissible.

Proof. Observe that lemma 2 implies that for any $h \in \mathbb{H}$ and any $1 \leq j \leq n$,

$$Er^{P_j}([h]_{\sim \mathcal{F}}) = \inf_{h_1, \dots, h_n \in [h]_{\sim \mathcal{F}}} \frac{1}{n} \sum_{i=1}^n Er^{P_i}(h_i).$$

The result now follows from theorem 1. \square

By combining standard generalization error result for single task learning with theorem 2, we now have an sample complexity bound for our full, two-stage, learning paradigm.

Definition 6 (MT-ERM Paradigm). *Given classes \mathcal{F} and \mathbb{H} as above, and a sequence of labeled sample sets, S_1, \dots, S_n , the MT-ERM (Multi-Task Empirical Risk Minimization) paradigm for \mathcal{F} and \mathbb{H} works in two steps as follows:*

1. Pick $h^* \in \mathbb{H}$ that minimizes $\inf_{h_1, \dots, h_n \in [h]} \sum_{i=1}^n \hat{E}r^{S_i}(h_i)$ over all $[h] \in \mathbb{H}/\sim_{\mathcal{F}}$.
2. Pick $h^\diamond \in [h^*]$ that minimizes $Er^{S_1}(h')$ over all $h' \in [h^*]$, and output h^\diamond as the learner's hypothesis.

Theorem 3. *Let (P_1, \dots, P_n) , (S_1, \dots, S_n) , \mathcal{F} and \mathbb{H} be as in the previous theorem. Let $d_{max} = \max_{h \in \mathbb{H}} VC\text{-dim}([h]_{\sim \mathcal{F}})$. Let h^\diamond be the output of an $(\mathcal{F}, \mathbb{H})$ -MT-ERM algorithm. Then, for every $\epsilon_1, \epsilon_2, \delta > 0$, if*

$$|S_1| \geq (64/\epsilon_1^2)[2d_{max} \log(12/\epsilon_1) + \log(8/\delta)]$$

and, for all $i > 1$

$$|S_i| \geq \frac{88}{\epsilon_2^2} \left[2d_{\mathcal{H}}(n) \log \frac{22}{\epsilon_2} + \frac{1}{n} \log \frac{8}{\delta} \right]$$

then, with probability greater than $(1 - \delta)$

$$Er^{P_1}(h^\diamond) \leq \inf_{h \in \mathbb{H}} Er^{P_1}(h) + 2(\epsilon_1 + \epsilon_2)$$

Proof. Let h^\sharp be the best P_1 label predictor in \mathbb{H} . That is, $h^\sharp = \arg \min_{h \in \mathbb{H}} Er^{P_1}(h)$. Let $[h^*]$ be the equivalence class picked in the first stage of the MT-ERM paradigm. I.e., $[h^*]$ is a minimizer of $\inf_{h_1, \dots, h_n \in [h]} \sum_{i=1}^n \hat{E}r^{S_i}(h_i)$ over all $[h] \in \mathbb{H}/\sim_{\mathcal{F}}$.

By the choice of h^* ,

$$\inf_{h_1, \dots, h_n \in [h^*]} \sum_{i=1}^n \hat{E}r^{S_i}(h_i) \leq \inf_{h_1, \dots, h_n \in [h^\sharp]} \sum_{i=1}^n \hat{E}r^{S_i}(h_i).$$

By Theorem 2, with probability greater than $(1 - \delta/2)$,

$$\inf_{h_1, \dots, h_n \in [h^\#]} \sum_{i=1}^n \hat{E}r^{S_i}(h_i) \leq Er^{P_1}([h^\#]) + \epsilon_1,$$

and also,

$$Er^{P_1}([h^*]) \leq \inf_{h_1, \dots, h_n \in [h^*]} \sum_{i=1}^n \hat{E}r^{S_i}(h_i) + \epsilon_1.$$

Combining these three inequalities, we get that with probability greater than $(1 - \delta/2)$,

$$Er^{P_1}([h^*]) \leq Er^{P_1}([h^\#]) + 2\epsilon_1.$$

Finally, the second stage of the MT-ERM algorithm is just a standard ERM algorithm yielding, with probability greater than $(1 - \delta/2)$, a hypothesis $h^\circ \in [h^*]$, whose P_1 error is within $2\epsilon_2$ of the best hypothesis there, namely of $Er^{P_1}([h^*])$.

Recall, that the common (single task) ERM paradigm requires a sample of size

$$|S_1| \geq (64/\epsilon^2)[\text{VC-dim}(\mathbb{H} \log(12/\epsilon) + \log(4/\delta)] \quad (1.3)$$

to find a hypothesis h° that with probability greater than $(1 - \delta)$ has

$$Er^{P_1}(h^\circ) \leq \inf_{h \in \mathbb{H}} Er^{P_1}(h) + 2\epsilon$$

It follows that the extent by which the samples from the extra P_i 's help depends on the gap between the parameters d_{max} and $d_{\mathcal{H}}$ and $\text{VC-dim}(\mathbb{H})$.

Next we examine what the values of these parameters for the specific case of learning axis-aligned rectangles in \mathfrak{R}^d . Finally, in section 1.6 we analyze these parameters for general classes \mathbb{H} and \mathcal{F} .

1.5 Analysis of Axis-Aligned Rectangles under Euclidean Shifts

Let $\mathcal{X} = \mathbb{R}^d$, and \mathbb{H} be the set of characteristic functions of axis-aligned rectangles, i.e., functions that map to 1 all points within some fixed rectangle $[a_1, a_1 + b_1] \times \dots \times [a_d, a_d + b_d]$ and map to 0 to all other points. Let \mathcal{F} be the set of Euclidean shifts, i.e., functions of the form $f(x_1, \dots, x_d) = (x_1 + v_1, \dots, x_d + v_d)$, where $v_1, \dots, v_d \in \mathbb{R}$. As above, we let \mathcal{H} denote $\mathbb{H}/\sim_{\mathcal{F}}$. Note that indeed in this case \mathcal{F} acts as a group over \mathbb{H} .

Claim. For $d > 1$ and $n > d$, $d_{\mathcal{H}}(n) \leq d + \lfloor \frac{d}{2} \rfloor$.

Proof. We will see in theorem 5 below, that for H as above and $n > d$,

$$d_{\mathcal{H}}(n) = \max_{[h] \in \mathcal{H}} \text{VC-dim}([h]).$$

So, it suffices to show that for any $[h] \in \mathcal{H}$, $\text{VC-dim}([h]) \leq d + \frac{d}{2}$. We prove this as the following lemma.

Lemma 3. *Let \mathbf{r} be an axis-aligned rectangle in \mathbb{R}^d , and let $F(\mathbf{r})$ be the class of all Euclidean shifts of \mathbf{r} . Then $\text{VC-dim}(F(\mathbf{r})) \leq d + \frac{d}{2}$.*

Proof. Suppose $[h]$ shatters set U . (I.e., for any $V \subseteq U$, there exists $h' \in [h]$ such that for all $x \in U$, $h'(x) = 1 \iff x \in V$. We say that such an h' obtains subset V of U .)

Then, in order to obtain the complements of each of the singleton subsets of U , each point $x \in U$ must have some coordinate k_x in which its value is either the greatest or the least among the k_x th coordinate of all points in U .

For a given point, $p \in \mathbb{R}^m$, let $p(k)$ denote its k th coordinate.

Assume $|U| > d + \frac{d}{2}$. Then, there must exist at least $d + 1$ points $p \in U$ for which k_p is unique, i.e., for every other coordinate k , there exist points $y, z \in U$ such that $y(k) > p(k)$ and $z(k) < p(k)$. And since we are in \mathbb{R}^d , there exist two such points, p and q such that $k_p = k_q$ and both k_p and k_q are unique. Call this coordinate k .

Now, what we have is points $p, q \in U$ such that $p(k) > x(k)$, and $q(k) > x(k)$ for all $x \in U - \{p, q\}$, and for every $k' \neq k$ there exist points $y, z \in S$ such that $y(k) > p(k), q(k)$ and $z(k) < p(k), q(k)$.

We proceed to show that no $h' \in [h]$ obtains the subset $U - \{p, q\}$.

Since $[h]$ must obtain the subset of S consisting of U itself, the length of the side in coordinate j for any $h' \in [h]$ must be at least $\max_{x, y \in U} |x(j) - y(j)|$. Without loss of generality, let us say that h obtains U . Then, any subset of U obtained by any $h' \in [h]$ consists of those points in U that remain after removing axis-parallel slices of h on up to d of its faces, with no two opposing faces sliced.

However, the only slices that can remove p and q without removing any other points from S are the two opposing slices in coordinate k , so, indeed, the subset $U - \{p, q\}$ cannot be obtained by any $h' \in [h]$. \square

Note that the VC-dimension of the class of axis-aligned rectangles in \mathbb{R}^d is $2d$. Comparing equation 1.2 to the corresponding standard VC-dimension generalization error bound [VC71] (shown in equation 1.3), we have the following.

Claim. For the purpose of learning the rectangle side lengths, VC-dimension considerations provide better accuracy guarantees for n shifted samples each of size m than for a single sample of size $n(\frac{8}{11}m - c)$, where c is a constant depending on the desired accuracy and the Euclidean dimension. Furthermore, c is small enough so that each sample may be smaller than that needed to obtain the same guarantees for a single data set of size m .

Previously, [BDGS02] considered the PAC setting, that is, the setting in which the learner is guaranteed that there exists a hypothesis $h \in \mathbb{H}$ that achieves zero error under the data generating distribution (in our case, P_1). For that setting, they showed that for the \mathcal{H} and \mathcal{F} of the example above, n shifted samples each of size m provide better accuracy guarantees than a

single sample of size $n(m - c')$, where c' is a constant depending on the desired accuracy and the Euclidean dimension. Our analysis here provides nearly as strong a result for the more realistic 'agnostic' setting, where the assumption of the existence of a zero error h is waived.

1.6 Analysis of $d_{\mathcal{H}}(n)$

In this section we investigate the parameters $d_{\mathcal{H}}$ that, along with d_{max} , determines the sample complexity (or, equivalently, the generalization error bounds) of multi-task learning in our setting of \mathcal{F} -related learning tasks derived in Theorem 3.

By Proposition 1, $d_{\mathcal{H}}(n + 1) \leq d_{\mathcal{H}}(n)$ for any n . Thus, we see from eq. 1.2 that once we have committed ourselves to the multitask approach, extra tasks can only be beneficial.

Note that since our collection, \mathcal{H} , is made of the equivalence classes $[h]_{\mathcal{F}}$ (formed by the functions of \mathcal{F}) over an initial hypotheses space, \mathbb{H} , the union of all these classes, $\bigcup\{H : H \in \mathcal{H}\}$, equals \mathbb{H} . Therefore, by Proposition 1, $d_{max} \leq d_{\mathcal{H}}(n) \leq VC\text{-dim}(\mathbb{H})$ (where, as above, $d_{max} = \max_{h \in \mathbb{H}} VC\text{-dim}([h])$). Thus, the best we can hope for is $d_{\mathcal{H}}(n) = d_{max}$. We conjecture that for any \mathbb{H} of finite VC-dimension and any \mathcal{F} , this lower bound is attained for all sufficiently large n . The following two theorems support this conjecture.

Notation:

Let $|h|$ denote the cardinality of the support of h , i.e., $|h|$ denotes $|\{x \in \mathcal{X} : h(x) = 1\}|$. Also, for a function h and a vector $\bar{x} = (x_1, \dots, x_n)$, let $\bar{h}(\bar{x}_1) = (h(x_1), \dots, h(x_n))$.

Theorem 4. *If there exists M such that $|h| \leq M$ for all $h \in \mathbb{H}$, then there exists n_0 such that for all $n \geq n_0$,*

$$d_{\mathcal{H}}(n) = \max_{h \in \mathbb{H}} VC\text{-dim}([h]_{\sim_{\mathcal{F}}}).$$

(Recall that \mathcal{H} denotes $\mathbb{H}/\sim_{\mathcal{F}}$.)

Proof. Assume $d_{\mathcal{H}}(n) \geq m$, and let $\bar{x}_1, \dots, \bar{x}_n$ be such that

$$\left| \left\{ \begin{bmatrix} \overline{h \circ f_1(\bar{x}_1)} \\ \vdots \\ \overline{h \circ f_n(\bar{x}_n)} \end{bmatrix} : f_1 \dots f_n \in \mathcal{F}, h \in \mathbb{H} \right\} \right| = 2^{nm}$$

Consider $h_0 \in \mathbb{H}$ and f_1, \dots, f_n such that

$$\begin{bmatrix} \overline{h_0 \circ f_1(\bar{x}_1)} \\ \vdots \\ \overline{h_0 \circ f_n(\bar{x}_n)} \end{bmatrix} = \begin{bmatrix} 1 \dots 1 \\ \vdots \ddots \vdots \\ 1 \dots 1 \end{bmatrix}$$

Note that for each i , there exists $S_i \subseteq h_0$ such that \bar{x}_i is some permutation of $\{f_i^{-1}(z) : z \in S_i\}$.

Say $|h_0| = K$. Then if $n > \binom{K}{m} 2^m$, then there exists $S \subseteq h_0$ and i_1, \dots, i_{2^m} such that $S_{i_j} = S$ for $1 \leq j \leq 2^m$. Let $\sigma_1, \dots, \sigma_{2^m}$ be the corresponding permutations.

Finally, letting v_1, \dots, v_{2^m} be an enumeration of all vectors of length m over $\{0, 1\}$, letting N be any $m \times n$ matrix over $\{0, 1\}$ whose i_j^{th} row is $\sigma_j(v_{i_j})$, and letting h_* and f'_1, \dots, f'_n be such that

$$\begin{bmatrix} \overline{h_* \circ f'_1(\bar{x}_1)} \\ \vdots \\ \overline{h_* \circ f'_n(\bar{x}_n)} \end{bmatrix} = N,$$

we see that $[h_*]_{\sim_{\mathcal{F}}}$ shatters S , so $m \leq \text{VC-dim}([h_*]_{\sim_{\mathcal{F}}})$.

To eliminate the dependence on $|h_0| = K$, we set $n_0 = \left(\binom{M}{M/2} - 1\right) 2^M$, noting that $n_0 \geq \left(\binom{K}{m} - 1\right) 2^m$ for all $K, m \leq M$. \square

So, we see that for any class of hypotheses bounded in size (i.e., the size of their support), for sufficiently large n , $d_{\mathcal{H}}(n)$ obtains its minimum possible value of d_{\max} . However, many natural hypothesis spaces consist of hypotheses that are not only unbounded, but infinite in size. In the following theorem, we show that this conjecture also holds for a natural hypothesis space consisting of infinite hypotheses.

Theorem 5. *Let \mathcal{X}, \mathbb{H} , and \mathcal{F} be the rectangles with shifts as in section 1.5, and let \mathcal{H} denote $\mathbb{H}/\sim_{\mathcal{F}}$ as usual. For $n > d$,*

$$d_{\mathcal{H}}(n) = \max_{h \in \mathbb{H}} \text{VC-dim}([h]_{\sim_{\mathcal{F}}}).$$

Proof. let $\bar{x}_1, \dots, \bar{x}_n \in (\mathbb{R}^d)^m$ be such that

$$\left| \left\{ \begin{bmatrix} \overline{h \circ f_1(\bar{x}_1)} \\ \vdots \\ \overline{h \circ f_n(\bar{x}_n)} \end{bmatrix} : f_1 \dots f_n \in \mathcal{F}, h \in \mathbb{H} \right\} \right| = 2^{nm}.$$

For $y \in \mathbb{R}^d$, and $1 \leq k \leq d$, we will denote by $y(k)$ the k th coordinate of y .

For $k = 1, \dots, d$, let $w_k = \max\{|y(k) - z(k)| : y, z \in \bar{x}_i \text{ for some } 1 \leq i \leq n\}$. w_1, \dots, w_d is the sequence of minimal possible side lengths for any rectangle h such that

$$\begin{bmatrix} \overline{h \circ f_1(\bar{x}_1)} \\ \vdots \\ \overline{h \circ f_n(\bar{x}_n)} \end{bmatrix} = \begin{bmatrix} 1 \dots 1 \\ \vdots \\ 1 \dots 1 \end{bmatrix}$$

Without loss of generality, let us assume that $\bar{x}_1, \dots, \bar{x}_n$ are ordered such that these maxima are attained within $\bar{x}_1, \dots, \bar{x}_d$.

Now, for any binary sequence $b = (b_1, \dots, b_m)$, there exists some $h_b \in \mathbb{H}$ such that

$$\begin{bmatrix} \overline{h_b \circ f_1(\bar{x}_1)} \\ \vdots \\ \overline{h_b \circ f_n(\bar{x}_n)} \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \\ b_1 & \dots & b_m \end{bmatrix}$$

Clearly, no such h_b can have its k th side length less than w_k . Furthermore, there is no advantage in having any k th side length greater than w_k . Thus, we see that if $h = [0, w_1] \times \dots \times [0, w_k]$, then $[h]_{\sim_{\mathcal{F}}}$ shatters \bar{x}_n , so $\text{VC-dim}([h]_{\sim_{\mathcal{F}}}) \geq m$. \square

So, we see that it is not uncommon for $d_{\mathcal{H}}(n)$ to attain its minimum possible value, d_{max} . As that value can be significantly less than $\text{VC-dim}(\mathbb{H})$, it is reasonable to expect that, in many cases, the MT-ERM bound of Theorem 3 is significantly less than the standard ERM bound (Equation 1.3). Thus our bounds can guarantee generalization on the basis of smaller size than the standard VC-dimension considerations for the single-task approach.

Ben-David, et. al. [BDGS02] provide the following further results on $d_{\mathcal{H}}(n)$.

Theorem 6. *If \mathcal{F} is finite and $\frac{n}{\log(n)} \geq \text{VC-dim}(\mathbb{H})$, then*

$$d_{\mathcal{H}}(n) \leq 2 \log(|\mathcal{F}|).$$

Note that this result leads us to scenarios under which $d_{\mathcal{H}}(n)$ is arbitrarily smaller than $\text{VC-dim}(\mathbb{H})$. Indeed, as long as \mathcal{F} is finite, no matter how complex \mathbb{H} is, $d_{\mathcal{H}}(n)$ remains bounded by $2 \log(|\mathcal{F}|)$. Furthermore, in practice, the requirement that \mathcal{F} be finite is not an unreasonable one, since real world problems come with bounded domains and real world computations have limited numerical accuracy.

Furthermore, [BDGS02] provides the following generalization of this result.

Theorem 7. *If $\sim_{\mathcal{F}}$ is of finite index⁶, k , and $n \geq \frac{\log k}{4b \log b}$, then*

$$d_{\mathcal{H}}(n) \leq \frac{\log k}{n} + 4b \log b,$$

where

$$b = \max \left(\max_{H \in \mathbb{H}/\sim_{\mathcal{F}}} \text{VC-dim}(H), 3 \right).$$

This shows that even if \mathcal{F} is infinite, $d_{\mathcal{H}}(n)$ cannot grow arbitrarily with increasing complexity of \mathcal{H} .

⁶ The *index* of an equivalence relation is the number of equivalence classes into which it partitions its domain.

1.7 Conclusions and Future Work

We have presented a useful notion of relatedness between tasks for multiple task learning. This notion of relatedness provides a natural model for a variety of real world learning scenarios. We have derived generalization error bounds for learning of multiple tasks related in this manner. These bounds depend on a generalized VC-dimension parameter, which can be significantly less than the ordinary VC-dimension, thus improving on the usual bounds for the single task approach. We have provided analysis of this parameter and its relationship to the usual VC-dimension, and we have given precise conditions under which our multitask approach provides generalization guarantees based on smaller sample size than the single task approach.

This work is a significant step towards the goal of a full theory of multiple task learning. With the restriction to a special type of relatedness of tasks, we have been able to obtain sample size bounds which are significantly better than previously proven bounds for the learning to learn scenario.

Hopefully, this work will stimulate future work in several directions. There is room for a more thorough understanding of the conditions under which multi-task learning is advantageous over the single task approach in our scenario; in particular, a greater understanding of the generalized VC-dimension parameter would provide such insight. It would also be fruitful to relax the requirements on the set of transformations through which the tasks are related, allowing these transformations to be arbitrary rather than bijections, and perhaps even allowing the actual transformations between the tasks to be merely approximated by the set of known transformations. Finally, the quest for further applicable notions of relatedness between tasks remains the key to a thorough understanding of multiple task learning.

We believe that this work provides convincing evidence that a theoretical understanding of multiple task learning and its advantage over the single task approach is a promising research endeavor worth pursuing.

References

- [Bax95] Jonathan Baxter. Learning Internal Representations. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers, 1995.
- [Bax00] Jonathan Baxter. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.
- [BDGS02] Shai Ben-David, Johannes Gehrke, and Reba Schuller. A Theoretical Framework for Learning from a Pool of Disparate Data Sources. In *Proceedings of the The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.
- [BDS03] Shai Ben-David, and Reba Schuller. Exploiting Task Relatedness for Multiple Task Learning. In *Proceedings of the The Sixteenth Annual Conference on Learning Theory (COLT)*, 2003.
- [BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.
- [BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers, 1998.
- [Car97] Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.
- [Hes98] Tom Heskes. Solving a Huge Number of Similar Tasks: A Combination of Multi-Task Learning and a Hierarchical Bayesian Approach. In *International Conference on Machine Learning*, pages 233–241, 1998.
- [IE96] N. Intrator and S. Edelman. How to Make a Low-Dimensional Representation Suitable for Diverse Tasks. *Connection Science*, 8, 1996.

- [KV97] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts, 1997.
- [Thr96] S. Thrun. Is learning the n-th thing any easier than learning the first? In D. Touretzky and M Mozer, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 640–646, 1996.
- [VC71] V. Vapnik and A. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theoret. Probl. And Its Appl*, 16(2):264–280, 1971.