# Estimating Relatedness via Data Compression

**Brendan Juba** BJUBA@MIT.EDU
MIT CSAIL, 32 Vassar St., Cambridge MA, 02139 USA

## Abstract

We show that it is possible to use data compression on independently obtained hypotheses from various tasks to algorithmically provide guarantees that the tasks are sufficiently related to benefit from multitask learning. We give uniform bounds in terms of the empirical average error for the true average error of the $n$ hypotheses provided by deterministic learning algorithms drawing independent samples from a set of $n$ unknown computable task distributions over finite sets.

## 1. Introduction

It has been observed that, given a selection of $n$ distinct learning problems, it is possible in some cases to solve the problems more effectively, in the sense of using fewer training samples per task to achieve a given level of generalization performance, by combining the data sets and solving the problems together. This technique is known as *multitask learning.* To see why this is plausible, consider the following example from Caruana (Caruana, 1997):

Suppose we are trying to infer from examples the rules for four functions, $f_1, f_2, f_3, f_4 : \{0,1\}^8 \to \{0,1\}$ where the functions are actually given by the rules $f_1(\overrightarrow{b}) = b_1 \vee \bigoplus_{i=2}^{6} b_i, f_2(\overrightarrow{b}) = \neg b_1 \vee \bigoplus_{i=2}^{6} b_i, f_3(\overrightarrow{b}) = b_1 \wedge \bigoplus_{i=2}^{6} b_i, f_4(\overrightarrow{b}) = \neg b_1 \wedge \bigoplus_{i=2}^{6} b_i$ Caruana observes that the tasks are related in a variety of ways: they are all defined on the same domain, they all ignore $b_7$ and $b_8$, each one uses the same (complicated) subfeature $g(\overrightarrow{b}) = \bigoplus_{i=2}^{6} b_i$, and only one member of each of the pairs $\{(f_1, f_2), (f_3, f_4)\}$ actually depends on the value of $g$. Caruana investigates this problem and many others in the setting of neural nets trained using backpropagation; he shows empirically that if the tasks are "related," then training a single neural net with a common hidden layer and multiple outputs achieves better performance as the number of tasks increases.

Appearing in *Proceedings of the 23$^{rd}$ International Conference on Machine Learning*, Pittsburgh, PA, 2006. 

Perhaps more interestingly, he shows that if tasks are grouped with an unrelated random task, performance degrades, leading us to believe that this phenomenon is tied to the "relatedness" of the tasks. He conducts an empirical investigation into what could be meant by "relatedness," and lists a variety of ways that the multiple tasks can help. Some efficient heuristic measures of relatedness have been developed for neural nets, by Thrun and O'Sullivan (Thrun & O'Sullivan, 1996) and Silver and Mercer (Silver & Mercer, 1998). These measures are based largely on the Euclidean distance between the weight vectors of the trained nets and do not generalize to other learning algorithms.

Baxter (Baxter, 2000) began investigating the problem from a generalized, theoretical perspective, as part of a larger investigation in *bias learning,* where task distributions themselves are randomly sampled from some distribution. Baxter proved PAC-learning-style $\varepsilon - \delta$ sample complexity guarantees for both bias learning and multitask learning in terms of the number of tasks, using measure-theoretic conditions on hypothesis spaces. He also proved theorems characterizing the efficiency of boolean function learning for multiple tasks in terms of the VC-dimension of the hypothesis classes. It is unlikely that these sorts of parameters, used to characterize relatedness, can be efficiently computed in practice, but given that unrelated tasks can hamper performance, it is certainly desirable to find some value that *can* be computed that characterizes relatedness in general.

In a more general setting, Li et al (Li et al., 2003a) have proposed a "universal similarity metric" based on compressibility; they define a Normalized Information Distance over pairs of strings in terms of the Kolmogorov complexities of the two strings concerned. They show that this distance has some nice properties, with the unfortunate exception of computability. To counter this, they propose using ordinary compression to approximate the ideal compression of Kolmogorov complexities. They define the corresponding metric, the Normalized Compression Distance, and demonstrate its suitable behavior on a host of examples.

It is extremely natural to consider applying compressibility as a measure of similarity between tasks, par-

tially because it seems to aptly capture the sorts of similarites that Caruana exhibits – one would expect that the representations of the example functions we discussed earlier should compress substantially for all of the reasons mentioned – but moreover because of the well-established connections between learning and compression. Blumer et al (Blumer et al., 1987) showed that, in order to learn, it suffices to have an "Occam algorithm" that compresses the data. Their result has been sharpened using Kolmogorov complexity by Li et al. (Li et al., 2003b), who, along with Board and Pitt (Board & Pitt, 1990) and Schapire (Schapire, 1990), showed that in many cases, compression is *necessary* for learning. Given this motivating history, and the naturalness of the idea it is surprising that no one has (to our knowledge!) tried to characterize "task-relatedness" in terms of compression.

We aim to recitify this disconnect, and present some preliminary results in this direction. We exhibit a PAC-learning style sample complexity bound in a restricted setting that, for $n$ tasks, independently obtained hypotheses $\overrightarrow{h}$, and fixed $\varepsilon$ and $\delta$, scales as $O(K(\overrightarrow{h})/n)$. Moreover, we argue that one can use any ordinary efficient compression procedure in the style of Normalized Compression Distances to obtain conservative sample complexity estimates—that is, if the sample complexity bound is satisfied using ordinary compression, then PAC-learning guarantees hold, and we can efficiently evaluate whether or not this is so. We also discuss how our result suggests that one may evaluate and select from multiple clusterings of tasks, since it holds for *independently obtained* hypotheses.

## 2. Preliminaries

First, a few words on notation: throughout, we will use lg to denote $\log_2$ and ln to denote $\log_e$. We will denote vectors by arrows, $\overrightarrow{v}$, and matrices by square brackets, $[m]$. Tragically, the nature of this work requires that all constants be carried around explicitly, so that we can verify that they can be found in practice; we apologize in advance for the appearance of all formulas.

### 2.1. Multitask Learning

We first describe Multitask Learning (Caruana, 1997) as viewed from the framework of Baxter (Baxter, 2000). Throughout, we will restrict our attention to computable probability distributions on finite sample spaces, which is significantly more restricted than Baxter's setting, but yields some interesting results (and saves us from some discussions of measure theory). We assume that an input set $X$ and output set $Y$ are given, and that there exist $n$ fixed (but unknown) computable probability distributions, $(P_1, \ldots, P_n) = \overrightarrow{P}$ over the finite set $Z = X \times Y$. We let $\overrightarrow{P}$ denote a joint distribution over $\overrightarrow{z} \in Z^n$ in the natural way, $\overrightarrow{P}(\overrightarrow{z}) = \prod_i P_i(z_i)$. In this way, we assume that we independently sample from the $n$ different task distributions. We will use the standard notions of the expectation and variance of a real-valued random variable, $\mathbb{E}[X] = \sum_z X(z)P(z)$ and $\mathsf{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

We will restrict our attention to deterministic (non-randomized) algorithms. We are free to choose, in the design of our learning algorithm, a hypothesis space $\mathcal{H}$ which is a set of hypotheses, which are functions $h : X \to Y$. $\mathcal{H}^n$ is, naturally, the $n$-fold cartesian product of $\mathcal{H}$ with itself, and describes a selection of $n$ different hypotheses for our $n$ different tasks. We also choose a loss function $l : Y \times Y \to [0, 1]$, and given $\overrightarrow{h} \in \mathcal{H}^n$, define the average loss $\overrightarrow{h}_l : (X \times Y)^n \to [0, 1]$ by $\overrightarrow{h}_l((x_1, y_1), \ldots, (x_n, y_n)) = \frac{1}{n}\sum_{i=1}^n l(h_i(x_i), y_i)$ It is worth noting that our assumption about the loss ranging over $[0, 1]$ is equivalent to assuming that its range is bounded – which is always true over a finite set – and that we have rescaled the loss.

We assume that our learning algorithm is given as input $m$ independent samples from $\overrightarrow{P}$ (which denotes itself a vector of independent samples from each of $P_1, \ldots, P_n$). We refer to this input matrix $[z]$ as a $(n, m)$-sample. The learning algorithm $M$, being deterministic, defines a map $M : (X \times Y)^{n \times m} \to \mathcal{H}^n$ from $(n, m)$-samples to a selection of $n$ hypotheses to our $n$ tasks. Ideally, we would like to choose $n$ hypotheses $\overrightarrow{h} \in \mathcal{H}^n$ to minimize the expected average loss over the $n$ tasks, $\mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h}) = \sum_{\overrightarrow{z} \in Z^n} \overrightarrow{h}_l(\overrightarrow{z})\overrightarrow{P}(\overrightarrow{z})$. This is often approached by minimizing the empirical loss on $[z]$, $\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}) = \frac{1}{m}\sum_{i=1}^m h_l(\overrightarrow{z}_i)$ where $\overrightarrow{z}_1, \ldots, \overrightarrow{z}_m$ are the $m$ column vectors of $[z]$. We will provide a bound on the difference between $\mathrm{er}_{\overrightarrow{P}}$ and $\hat{er}_{[z]}$ that is uniform over all $\overrightarrow{h} \in \mathcal{H}^n$. Thus, the empirical error provides a good estimate for the expected error when the bound holds, and a set of hypotheses that do well on the training sample do well in expectation over the true distributions.

### 2.2. Information Theory

We will require the following standard result from Information Theory due to Shannon, McMillan, and Breiman; we will more or less follow MacKay's (MacKay, 2002) exposition:

**Theorem 1 (Asymptotic Equipartition Prop-**

**erty).** *Let $P$ be a probability distribution over a set $Z$. Denote by $H$ the entropy of $P$, and let $\sigma^2 = \mathsf{Var}\left[\lg \frac{1}{P(z)}\right]$. Now, for a sequence $\overrightarrow{z}$ of $m$ i.i.d. samples from $P$, the "typical set" $T_{m\beta} = \{\overrightarrow{z} \in Z^m : |\frac{1}{m}\lg\frac{1}{P(\overrightarrow{z})} - H| < \beta\}$ satisfies $P(\overrightarrow{z} \in T_{m\beta}) \geq 1 - \frac{\sigma^2}{\beta^2 m}$.*

We will find that it is useful to bound the variance of the self-information of an arbitrary finite distribution, so that we can apply the Asymptotic Equipartition Property when we do not know the distribution $P$.

**Lemma 1.** *For any probability distribution $P$ over a finite set $Z$ with $|Z| \geq 3$, $\mathsf{Var}\left[\lg \frac{1}{P(z)}\right] \leq \lg^2 |Z|$.*

### 2.3. Kolmogorov Complexity

The definitions and results in this section are all covered by Li and Vitányi (Li & Vitányi, 1997).

**Definition 1 (Prefix complexity).** *The* prefix complexity *of a string $x$, denoted $K(x)$, is the length of the shortest string $p$ such that on a fixed universal prefix machine $U$, $U(p) = x$.*

It is known that between any pair of universal prefix machines, the prefix complexities differ by at most a constant amount, though the constant depends on the pair of machines. We will later have cause to specify in more detail the universal prefix machine with respect to which we wish to measure prefix complexity. It is important to note that prefix complexity can be bounded through the use of ordinary data compression algorithms. For an arbitrary compression algorithm which compresses a string $x$ to a string of length $C(x)$, we can add a self-delimiting code prefix to the compressed string, describing its length, so that clearly $K(x) \leq C(x) + 2\lg C(x) + O(1)$.

**Definition 2 (Semimeasure).** *A* semimeasure *for a set $S$ is a function $P : S \to \mathbb{R}^+$ such that $\sum_{x\in S} P(x) \leq 1$. In particular, note that any probability distribution qualifies as a semimeasure.*

**Definition 3 (Universal enumerable discrete semimeasure).** *A* universal enumerable discrete semimeasure *(denoted by $\mathbf{m}$) is an enumerable discrete semimeasure such that for any other discrete enumerable semimeasure $P$, there exists a constant $c_P$ such that for all strings $x$, $P(x) \leq c_P \mathbf{m}(x)$.*

It is important to note that $\mathbf{m}$ is defined in terms of a reference enumeration of the enumerable discrete semimeasures. Much like the reference universal prefix machine, we can specify the enumeration we wish to use in greater detail, and we will have cause to do so. We will make use of the following standard result, due independently to Levin, Gács, and Chaitin:

**Theorem 2 (Coding Theorem).** *There is a constant $c$ such that for all $x$, $-\lg \mathbf{m}(x) = K(x) \pm c$*

We should note that it suffices to use $c = \max\{K(Q), K(T) + 4\}$ where $K(Q)$ is the index in an enumeration of a program computing the "universal a priori probability distribution," $Q(x) = \sum_{U(p)=x} 2^{-|p|}$ for a universal prefix machine $U$, and $K(T)$ is the complexity of a program decoding a Shannon-Fano style coding under $\mathbf{m}$. These programs are fixed in advance.

We will further require the following corollary from Li and Vitányi (Li & Vitányi, 1997):

**Corollary 1.** *If $P$ is an enumerable discrete semimeasure, then for all $x$, $K(x) \leq -\lg P(x) + K(P) + c$, where $K(P)$ is the index of $P$ in the reference enumeration of enumerable semimeasures.*

We will also make use of the following consequence of Markov's inequality:

**Claim 1.** *For any probability distribution $P$, $P[x : -\lg P(x) \leq -\lg \mathbf{m}(x) + \lg k] \geq 1 - \frac{1}{k}$.*

In summary,

**Claim 2.** *With probability at least $1 - 1/k$ over $x$ drawn according to a computable probability distribution $P$, we have a constant $c$ such that*

$$-\lg P(x) - \lg k - c \leq K(x) \leq -\lg P(x) + K(P) + c$$

## 3. Multitask Sample Complexity Bounds

### 3.1. Baxter's Bound

In this section, we review the lemmas leading to the following result of Baxter (Baxter, 2000): for $n$ probability distributions $\overrightarrow{P} = (P_1, \ldots, P_n)$ on $X \times Y$, and a hypothesis space $\mathcal{H}$ with $\mathcal{H}_l^n = \{\overrightarrow{h}_l : \overrightarrow{h} \in \mathcal{H}^n\}$ "permissible," if the number of training samples $m$ satisfies $m \geq \max\{\frac{64}{n\varepsilon^2} \ln \frac{4\mathcal{C}(\frac{\varepsilon}{16}, \mathcal{H}_l^n)}{\delta}, \frac{16}{\varepsilon^2}\}$ then with probability at least $1 - \delta$ over the $(n, m)$-sample $[z]$ used, every $\overrightarrow{h} \in \mathcal{H}_l^n$ satisfies $\mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h}) \leq \hat{\mathrm{er}}_{[z]}(\overrightarrow{h}) + \varepsilon$. We will not discuss the meaning of $\mathcal{C}(\varepsilon, \mathcal{H}_l^n)$ in detail, but informally it is the smallest number $N$ such that for any probability distribution, one can find a set of $N$ functions such that every function in $\mathcal{H}_l^n$ is approximated to within $\varepsilon$ under the measure given by the probability distribution. Although it is quite an appropriate notion, it is unclear how one would compute or estimate this value in practice, which was the motivation for the present work. Likewise, we will not discuss "permissibility," but it seems to be required for the precise reason of the measurability of suprema over permissible sets.

We will make use of several of the lemmas proved by Baxter and combined to yield the aforementioned result. As previously done by Haussler (Haussler, 1992), he uses the parameterized class of metrics $d_\nu$ for $\nu \in \mathbb{R}^+$ given by $d_\nu[x,y] = \frac{|x-y|}{x+y+\nu}$ which has the following easy properties:

**Lemma 2.** *1.* $\forall r,s \geq 0,\ 0 \leq d_\nu[r,s] \leq 1$

*2. For* $0 \leq r \leq s \leq t$, $d_\nu[r,s] \leq d_\nu[r,t]$ *and* $d_\nu[s,t] \leq d_\nu[r,t]$.

*3. For* $0 \leq r,s \leq 1$, $\frac{|r-s|}{\nu+2} \leq d_\nu[r,s] \leq \frac{|r-s|}{\nu}$

Given a $(n,2m)$-sample $[z]$, let $[z(1)]$ denote samples indexed $1,2,\ldots,m$, and let $[z(2)]$ denote samples indexed $m+1,m+2,\ldots,2m$. In this way, $[z(1)]$ and $[z(2)]$ both denote $(n,m)$-samples. Baxter considers the probability of large deviation between empirical estimates of the loss obtained from these samples when elements are randomly permuted between them. This notion is captured by the following permutation group: $\forall m,n \geq 1$, let $\Gamma_{(n,2m)}$ denote the set of all permutations $\sigma$ of the sequence of integer pairs $\{(1,1),(1,2),\ldots,(1,2m),(2,1),\ldots,(n,1),\ldots,(n,2m)\}$ such that for all $1 \leq j \leq m$, either $\sigma(i,j) = (i,m+j)$ and $\sigma(i,m+j) = (i,j)$ or $\sigma(i,j) = (i,j)$ and $\sigma(i,m+j) = (i,m+j)$. We now let $[z_\sigma]$ denote

$$[z_\sigma] = \begin{bmatrix} z_{\sigma(1,1)} & \cdots & z_{(1,2m)} \\ \vdots & \ddots & \vdots \\ z_{\sigma(n,1)} & \cdots & z_{(n,2m)} \end{bmatrix}$$

The lemmas proceed as follows: using this bound on the probability of deviation between two empirical estimates of the loss when elements are randomly permuted between the samples, a bound is obtained on the probability of deviation between the empirical estimates of the loss for two independent samples. This is in turn used to bound the probability of deviation between the true expected average loss and an empirical estimate of the loss, yielding the desired uniform bound. In restating the lemmas, we have tweaked a few parameters, and replaced statements about suprema with statements about fixed hypotheses. These versions are easily verified to follow from essentially the same arguments. For the following lemmas, Baxter acknowledges inspiration from the double symmetrization arguments of Pollard (Pollard, 1984) and Haussler (Haussler, 1992).

**Lemma 3.** *Let* $\overrightarrow{f} : Z^n \to [0,1]$ *be any function that can be written in the form* $\overrightarrow{f}(\overrightarrow{z}) = 1/n \sum_{i=1}^n f_i(z_i)$. *For any* $[z] \in Z^{(n,2m)}$, *if* $\sigma \in \Gamma_{(n,2m)}$ *is chosen uniformly at random, then* $\Pr[\sigma \in \Gamma_{(n,2m)} : d_\nu[\hat{\mathrm{er}}_{[z_\sigma(1)]}(\overrightarrow{f}), \hat{\mathrm{er}}_{[z_\sigma(2)]}(\overrightarrow{f})] > \frac{\alpha}{2}] \leq 2\exp(\frac{-\alpha^2\nu mn}{2})$

**Corollary 2.** *When an* $(n,2m)$*-sample* $[z]$ *is drawn according to any probability distribution* $P$, *we still have* $\Pr[[z] \in Z^{(n,2m)} : d_\nu[\hat{\mathrm{er}}_{[z(1)]}(\overrightarrow{f}), \hat{\mathrm{er}}_{[z(2)]}(\overrightarrow{f})] > \frac{\alpha}{2}] \leq 2\exp(\frac{-\alpha^2\nu mn}{2})$

**Lemma 4.** *Let* $P$ *be a probability measure on* $Z^n$ *and let* $h : Z^n \to [0,1]$. $\forall \nu > 0, 0 < \alpha < 1$, *and* $m \geq \frac{2}{\alpha^2\nu}$, $\Pr[[z] \in Z^{(n,m)} : d_\nu[\hat{\mathrm{er}}_{[z]}(h), \mathrm{er}_P(h)] > \alpha] \leq 2\Pr[[z] \in Z^{(n,2m)} : d_\nu[\hat{\mathrm{er}}_{[z(1)]}(h), \hat{\mathrm{er}}_{[z(2)]}(h)] > \frac{\alpha}{2}]$

Putting these lemmas together, we find

**Claim 3.** *Let* $\overrightarrow{h} : Z^n \to [0,1]$ *be any function that can be written in the form* $\overrightarrow{h}(\overrightarrow{z}) = 1/n \sum_{i=1}^n h_i(z_i)$. *Let* $P$ *be a probability measure on* $Z^n$. *When our samples are drawn according to* $P$, $\forall \nu > 0, 0 < \alpha < 1$, *and* $m \geq \frac{2}{\alpha^2\nu}$, $\Pr[[z] \in Z^{(n,m)} : d_\nu[\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}), \mathrm{er}_P(\overrightarrow{h})] > \alpha] \leq 4\exp(\frac{-\alpha^2\nu mn}{2})$

### 3.2. Kolmogorov Complexity Bounds

It is well-known that under uniform distributions, the Kolmogorov complexity of samples from the distribution are near maximal with high probability. Thus, under the uniform distribution, independent samples have nearly the same Kolmogorov complexity with high probability, so one sample can be used to give a bound on the Kolmogorov complexity of any other sample, and this bound holds with high probability. Our first trick will be to generalize this observation to other distributions:

**Lemma 5.** *Let* $\overrightarrow{P} = (P_1,\ldots,P_n)$ *each be computable probability distributions over a finite set* $Z$. *Let* $H_{\overrightarrow{P}}$ *denote the entropy of the distribution* $\overrightarrow{P}$. *For any* $(n,m)$*-sample* $[z] \in Z^{(n,m)}$ *where each* $z_{ij}$ *is drawn independently according to* $P_i$, *there is a constant* $c$ *such that* $\forall \beta > 0$, $P[[z] \in Z^{(n,m)} : |K([z]) - mH_{\overrightarrow{P}}| > (mn\beta + \max\{\lg k, K(P)\} + c)] < \frac{\lg^2|Z|}{\beta^2 mn} + \frac{1}{k}$

*Proof.* Let $H_{P_i}$ denote the entropy of the distribution $P_i$. First, notice that since each $z_{ij}$ in a $(n,m)$ sample is drawn independently, $\lg \frac{1}{P([z])} = \sum_{i=1}^n \sum_{j=1}^m \lg \frac{1}{P_i(z_{ij})}$, which has expected value $m\sum_{i=1}^n H_{P_i} = mH_{\overrightarrow{P}}$. By the Asymptotic Equipartition Property and lemma 1, we find that $P([z] : |\lg \frac{1}{P([z])} - mH_{\overrightarrow{P}}| > mn\beta) < \frac{\lg^2|Z|}{\beta^2 mn}$ Now, recalling claim 2, there is a constant $c$ such that $|K([z]) - \lg \frac{1}{P([z])}| \leq \max\{K(P), \lg k\} + c$ with probability at least $1 - 1/k$. Furthermore, by the triangle inequality, $|K([x]) - mH_{\overrightarrow{P}}| \leq |K([x]) - \lg \frac{1}{P([x])}| + |\lg \frac{1}{P([x])} - mH_{\overrightarrow{P}}|$ So the claimed statement follows from simple union bounds. □

So now, given a $(n, m)$-sample $[z]$, we can obtain a bound on the Kolmogorov complexity of any $(n, m)$-sample that holds with high probability. Of course, bounding the Kolmogorov complexity implies that we have bounded the number of $(n, m)$-samples that we are likely to see. It is possible to combine this observation with Baxter's lemmas (Baxter, 2000) to obtain a bound on the sample complexity (much as we do in Theorem 3), but such a bound is not particularly desirable for the following reasons.

First, the bound we obtain depends on the Kolmogorov complexity of our $(n, m)$-sample, $K([z])$, where clearly this quantity tends to grow with $m$. In particular, it provides guarantees for all $\delta, \epsilon > 0$ if and only if $K([z]) = o(mn)$; that is, if and only if $[z]$ compresses radically. One interpretation of this statement is that our data set $[z]$ must be *not very random* in the following sense: if some $P_i$ were uniform over some subset of $Z$, then it is known (Li & Vitányi, 1997) that with high probability, each sample from that distribution should not compress below the size of indices in the set by more than a constant amount—we would certainly not expect the sort of behavior necessary for our bound to hold. By constrast, if for example $X = \{1, 2, \ldots, N\}$ and we had samples for some range 1–$m$, then the $x$-values would compress to $O(\lg m)$ bits, since they could be output by a loop with a constant-size body, "output $i$," and a $O(\lg m)$-bit stopping condition, "stop if $i > m$."

It should be clear that the real difficulty with such a bound is that it includes a penalty for the compressed size of randomly sampled $x_{ij} \in X$, so we would really rather have a bound on the size of $y_{ij}$ given $x_{ij}$, or on the size of the hypotheses when compressed. This is much closer in spirit to the approaches discussed by Thrun and O'Sullivan (Thrun & O'Sullivan, 1996) and by Silver and Mercer (Silver & Mercer, 1998), and is likewise much closer to capturing the sense of relatedness described in Caruana's examples (Caruana, 1997). Our much-maligned bound on the compressed size of the data sets yields a bound on the sizes of the compressed size of the hypotheses since the compressed data together with the algorithm used describes the hypotheses, so we are interested in finding a bound that overcomes the above limitations.

We find that it is possible to obtain a bound based on the Kolmogorov complexity of hypotheses as follows: let us return to the setting of ordinary (single-task) learning for a moment. Let $M_1, \ldots, M_n$ be deterministic learning algorithms for our $n$ tasks. We can use these algorithms to obtain an initial $\overrightarrow{h} \in \mathcal{H}^n$ from an $(n, m)$-sample, and depending on how well the set of hypotheses compresses, we can obtain various upper bounds on $\varepsilon$ for various values of $\delta$. This suggests that we may wish to attempt to cluster the tasks (much as done by Thrun and O'Sullivan (Thrun & O'Sullivan, 1996)) in order to find a partition that yields the best $\varepsilon$-$\delta$ tradeoffs; we will see later how this can be done while maintaining the guarantees.

Note that for $M_i : Z^m \to \mathcal{H}$, the range over inputs of length $m$, $\mathcal{H}_{m,i} = \{h \in \mathcal{H} : \exists \overrightarrow{z} \in Z^m M_i(\overrightarrow{z}) = h\}$ is a subset of $\mathcal{H}$ which is finite since $M_i$ is deterministic and $Z$ is finite. Given $n$ computable probability distributions $P_1, \ldots, P_n$, let $\overrightarrow{Q}_m$ be a distribution over $\mathcal{H}_{m,i}$ defined by $Q_{m,i}[h_i] = P_i[\overrightarrow{z} : M_i(\overrightarrow{z}) = h_i]$; observe that $Q_{m,i}$ is computable too. Intuitively, we will argue that Lemma 5 tells us that with high probability, the hypotheses we obtain from $\overrightarrow{Q}_m$ come from a "typical set," in which all hypotheses have roughly the same compressed size. So, we can use the compressed size of our sample hypotheses to estimate the size of the set and apply Baxter's lemma with a union bound.

**Theorem 3.** *Let $\overrightarrow{P} = (P_1, \ldots, P_n)$ each be computable probability distributions over a finite set $Z = (X \times Y)$, let $\mathcal{H}$ be an arbitrary set of hypotheses $h : X \to Y$, and let a loss function $l : Y \times Y \to [0, 1]$ be given. Let $M_1, \ldots, M_n : Z^m \to \mathcal{H}$ be deterministic learning algorithms. When our $(n, m)$-sample $[z]$ is drawn according to $\overrightarrow{P}$, if $m$ satisfies $m \geq \max\{\frac{32 \ln 2}{\varepsilon^2 n}(\lg \frac{2}{\delta} + K(M_1([z]), \ldots, M_n([z])) + 2\max\{\lg \frac{4}{\delta}, K(\overrightarrow{Q}_m)\} + 2c + 3), \frac{16}{\varepsilon^2}\}$ where $n$ satisfies $n \geq \frac{8192 \ln^2 |Z|}{\varepsilon^4 \delta}$ then with probability at least $1 - \delta$, any $\overrightarrow{h} \in \mathcal{H}^n$ with $K(\overrightarrow{h}) \leq K(M_1([z]), \ldots, M_n([z]))$ satisfies $\mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h}) \leq \hat{\mathrm{er}}_{[z]}(\overrightarrow{h}) + \varepsilon$.*

*Proof.* It will help us to note that $|\mathcal{H}_{m,i}| \leq |Z|^m$. Let $\overrightarrow{M}([z])$ denote $(M_1([z]), \ldots, M_n([z])) \in \mathcal{H}^n$. Now define $\mathcal{H}^n([z]) = \{\overrightarrow{h} \in \mathcal{H}^n : K(\overrightarrow{h}) \leq K(\overrightarrow{M}([z]))\}$ and $\mathcal{H}^n_{\mathsf{short}} = \{\overrightarrow{h} \in \mathcal{H}^n : K(\overrightarrow{h}) \leq H_{\overrightarrow{Q}_m} + (n\beta + \max\{\lg k, K(\overrightarrow{Q}_m)\} + c)\}$ so that

$$
\begin{aligned}
P[[z] : \exists \overrightarrow{h} \in \mathcal{H}^n([z]) d_\nu[\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}), \mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h})] > \alpha] \leq \\
P[[z] : |K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| > (n\beta + \max\{\lg k, \\
K(\overrightarrow{Q}_m)\} + c)] + P[[z] : \exists \overrightarrow{h} \in \mathcal{H}^n([z])\ d_\nu[\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}), \\
\mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h})] > \alpha, |K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| \leq (n\beta + \\
\max\{\lg k, K(\overrightarrow{Q}_m)\} + c)] \leq \\
P[[z] : |K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| > (n\beta + \max\{\lg k, K(\overrightarrow{Q}_m) \\
\} + c)] + P[[z] : \exists \overrightarrow{h} \in \mathcal{H}^n_{\mathsf{short}} d_\nu[\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}), \mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h})] > \alpha, \\
|K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| \leq (n\beta + \max\{\lg k, K(\overrightarrow{Q}_m)\} + c)]
\end{aligned}
$$

Since then, $\mathcal{H}^n([z]) \subseteq \mathcal{H}^n_{\mathsf{short}}$

In the last expression, since $|K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| \leq (n\beta + \max\{\lg k, K(\overrightarrow{Q}_m)\} + c)$, $|\mathcal{H}^n_{\mathsf{short}}| \leq 2^{K(\overrightarrow{M}([z]))+2(n\beta+\max\{\lg k, K(\overrightarrow{Q}_m)\}+c)+1}$. Thus, a union bound together with claim 3 yields

$$P[[z] : \exists \overrightarrow{h} \in \mathcal{H}^n_{\mathsf{short}} d_\nu[\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}), \mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h})] > \alpha,$$
$$|K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| \leq (n\beta + \max\{\lg k, K(\overrightarrow{Q}_m)\} + c)]$$
$$\leq \exp(\ln 2(K(\overrightarrow{M}([z])) + 2(n\beta + \max\{\lg k, K(\overrightarrow{Q}_m)\} + c) + 3) - \frac{\alpha^2 \nu m n}{2})$$

whenever $m \geq 2/(\alpha^2\nu)$.

Suppose now we put $k = \frac{4}{\delta}$, $\beta = \sqrt{\frac{4}{n\delta}} m \lg |Z|$, $\nu = 2$, and $\alpha = \frac{\varepsilon}{4}$. Now, if $n \geq \frac{4^4 \cdot 2^2 \cdot 2^2 \cdot 2 \ln^2 |Z|}{\varepsilon^4 \delta}$ and $m \geq \frac{4^2 \cdot 2 \ln 2}{\varepsilon^2 n}(\lg \frac{2}{\delta} + K(\overrightarrow{M}([z])) + 2\max\{\lg \frac{4}{\delta}, K(\overrightarrow{Q}_m)\} + 2c + 3)$ then we find that the latter probability is bounded by $\delta/2$. Therefore, given that additionally $m \geq 4^2/\varepsilon^2$, $P[[z] : \exists \overrightarrow{h} \in \mathcal{H}^n([z]) d_\nu[\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}), \mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h})] > \alpha] \leq \frac{\delta}{2} + P[[z] : |K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| > (n\beta + \max\{\lg 4/\delta, K(\overrightarrow{Q}_m)\} + c)]$ where now, since $|\mathcal{H}_{m,i}| \leq |Z|^m$, we can apply lemma 5 to our sample from $\overrightarrow{Q}_m$ (with its $m = 1$), yielding $P[[z] : |K(\overrightarrow{M}([z])) - H_{\overrightarrow{Q}_m}| > (n\beta + \max\{\lg 4/\delta, K(\overrightarrow{Q}_m)\} + c)] < \frac{\delta}{2}$ so finally $P[[z] : \exists \overrightarrow{h} \in \mathcal{H}^n([z]) d_\nu[\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}), \mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h})] > \alpha] < \delta$ and thus the theorem follows from lemma 2. □

Although we are blithely ignoring the problem of deciding whether or not a hypothesis satisfies the Kolmogorov complexity condition for theorem 3, we observe that we know at least one hypothesis that trivially does—the default hypothesis, $\overrightarrow{h} = (M_1([z]), \ldots, M_n([z]))$. Also, the $1/\varepsilon^2$ and $1/\varepsilon^4$ factors can be improved to $1/\varepsilon$ and $1/\varepsilon^2$, respectively, if one is willing to weaken the error approximation to $\mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h}) \leq \kappa \hat{\mathrm{er}}_{[z]}(\overrightarrow{h}) + \varepsilon$ for some $\kappa > 1$. This is particularly useful if the $\{M_i\}$ in theorem 3 can find a joint hypothesis with $\hat{\mathrm{er}}_{[z]}(\overrightarrow{h}) = 0$. Refer again to Baxter (Baxter, 2000) for more details.

### 3.3. Adapting Kolmogorov Complexity Bounds for Estimation

Upon examining theorem 3, we see that most of the terms are (under reasonable circumstances) known, with the notable exceptions of terms involving Kolmogorov complexities, the unknown distribution function $\overrightarrow{Q}_m$, or both. This work was motivated by the observation that it is possible to provide an upper bound on Kolmogorov complexities in practice. As we noted in section 2.3, we can convert any standard decompression procedure into a decompression procedure which operates on prefix-coded versions of the original compressed strings. We noted that we can use this procedure and the length of the compressed string to bound the Kolmogorov complexity of a string.

Recall that it does not matter too much which universal machine we choose to define Kolmogorov complexity with respect to; the complexities differ by at most a constant amount. We observe that, for any given setting, we can do our accounting with respect to a clever choice of reference machine, one which will permit us to fix the various unknown constants that have crept into our bounds.

**Corollary 3.** *Let $\overrightarrow{P} = (P_1, \ldots, P_n)$, $Z = (X \times Y)$, $\mathcal{H} = \{h : X \to Y\}$, $l : Y \times Y \to [0, 1]$, and $M_1, \ldots, M_n : Z^m \to \mathcal{H}$ be given as in Theorem 3. Let a compression algorithm $M_C$ such that $C(x)$ is the length of $M_C(x)$ and $M_D$, the corresponding decompression algorithm, be given. Let $\overrightarrow{M}(x)$ denote $(M_1(x), \ldots, M_n(x))$. When our $(n, m)$-sample $[z]$ is drawn according to $\overrightarrow{P}$, if $m$ satisfies $m \geq \max\{\frac{32 \ln 2}{\varepsilon^2 n}(\lg \frac{2}{\delta} + C(\overrightarrow{M}([z])) + 2 \lg C(\overrightarrow{M}([z])) + 2\max\{\lg \frac{4}{\delta}, 3\} + 17), \frac{16}{\varepsilon^2}\}$ and $n$ satisfies $n \geq \frac{8192 \ln^2 |Z|}{\varepsilon^4 \delta}$ then with probability at least $1 - \delta$, any $\overrightarrow{h} \in \mathcal{H}^n$ with $K(\overrightarrow{h}) \leq K(\overrightarrow{M}([z]))$ satisfies $\mathrm{er}_{\overrightarrow{P}}(\overrightarrow{h}) \leq \hat{\mathrm{er}}_{[z]}(\overrightarrow{h}) + \varepsilon$.*

*Proof.* Let any universal prefix machine $U$, and any enumeration of the discrete enumerable semimeasures, $\phi_1, \phi_2, \ldots$, be given. First, observe that we can convert $M_D$ into $M'_D$, a prefix-free version: if $M_D(x) = y$, then $M'_D(0^{\lg |x|}|x|x) = y$, and this is uniquely decodable since the first bit of $|x|$ is a '1'. Clearly, the new input has length $|x| + 2 \lg |x|$. We also assume that an algorithm for $P$ exists; an algorithm for $\overrightarrow{Q}_m$ can be easily constructed from the algorithm for $P$. Recall now, that as discussed in section 2.3, $c = \max\{K(Q), K(T) + 4\}$ where $K(Q)$ is the complexity of the index of a program computing the "universal a priori probability distribution" in the reference enumeration used to define $\mathbf{m}$, the universal discrete enumerable semimeasure, and $K(T)$ is the complexity of a program decoding a Shannon-Fano style coding under $\mathbf{m}$.

We observe that now that all of these constants have been fixed, we are free to do the calculation of the Kolmogorov complexities with respect to any machine of our choice. Suppose we now construct a prefix machine $U'$ which first reads a two-symbol code prefix,

which we can interpret as a binary integer. Assume the remaining string is referred to as $x$; depending on the prefix, $U'$ either computes $U(x)$, $M'_D(x)$, or $T(x)$ or checks whether $x = 0$ or 1, outputing 1 if $x = 0$, 2 if $x = 1$, and otherwise looping forever.

Similarly, suppose we make the following modifications to $\phi$: we put $\phi_1 = \overrightarrow{Q}_m$, $\phi_2 = Q'$ where $Q'(x) = \sum_{U'(p)=x} 2^{-|p|}$ is the universal a priori probability distribution for $U'$, and for all $i > 2$, put $\phi'_i = \phi_{i-2}$. Clearly this is still an enumeration of the discrete enumerable semimeasures, where $K(\overrightarrow{Q}_m) = K(Q') = 3$. Note also that since $K(T) + 4 = 6$, the constant $c = 6$. Finally, since on this machine, any string $x$ has a program of length $C(x) + 2\lg C(x) + 2$, $K(x) \leq C(x) + 2\lg C(x) + 2$. The claim now follows directly from theorem 3. □

This bound presently applies only when all $n$ tasks are taken together. It turns out that we can easily extend it to apply the clustering approach suggested earlier.

**Corollary 4.** *Let $T = \{1, \ldots, n\}$ denote our set of tasks, and let $S \subset 2^T$ be any set of clusterings of the tasks of interest. If, for each clustering $s \in S$ of $k$ tasks $s_1, \ldots, s_k$, we have $m \geq \frac{32\ln 2}{\varepsilon_s^2 k}(\lg \frac{2}{\delta} + C(\overrightarrow{M_s}([z])) + 2\lg C(\overrightarrow{M_s}([z])) + 2\lg \frac{4}{\delta} + 17) + \frac{96\ln 2}{\varepsilon_s^2}(1 + \lg(\frac{ne}{k}))$ and $k \geq \frac{8192\ln^2|Z|}{\varepsilon_s^4\delta}$ where, letting $[z|_s]$ denote the restriction of $[z]$ to the tasks contained in $s$, we have put $\overrightarrow{M_s}([z]) = (M_{s_1}([z|_s]), \ldots, M_{s_k}([z|_s]))$ then with probability at least $1 - \delta$, for every cluster $s \in S$, any $\overrightarrow{h} \in \mathcal{H}^k$ with $K(\overrightarrow{h}) \leq K(\overrightarrow{M_s}([z]))$ satisfies $\mathrm{er}_{(P_{s_1},\ldots,P_{s_k})}(\overrightarrow{h}) \leq \hat{\mathrm{er}}_{[z|_s]}(\overrightarrow{h}) + \varepsilon_s$.*

*Proof.* Recalling the proof of theorem 3, notice that the statement we wish to prove is equivalent to (for $\alpha_s = \varepsilon_s/4$) $P[[z] : \exists s \in S, |s| = k \exists \overrightarrow{h} \in \mathcal{H}^k([z|_s]) d_\nu[\hat{\mathrm{er}}_{[z|_s]}(\overrightarrow{h}), \mathrm{er}_{(P_{s_1},\ldots,P_{s_k})}(\overrightarrow{h})] > \alpha_s] < \delta$ where, by corollary 3, it is clear that if the stated conditions are satisfied, then for each set of $k$ tasks, we have $P[[z] : \exists \overrightarrow{h} \in \mathcal{H}^k([z|_s]) d_\nu[\hat{\mathrm{er}}_{[z|_s]}(\overrightarrow{h}), \mathrm{er}_{(P_{s_1},\ldots,P_{s_k})}(\overrightarrow{h})] > \alpha_s] < \frac{\delta}{2^k\binom{ne}{k}^k} \leq \frac{\delta}{2^k\binom{n}{k}}$ and, by a union bound over the sets in $S$, $P[[z] : \exists s \in S, |s| = k \exists \overrightarrow{h} \in \mathcal{H}^k([z|_s]) d_\nu[\hat{\mathrm{er}}_{[z|_s]}(\overrightarrow{h}), \mathrm{er}_{(P_{s_1},\ldots,P_{s_k})}(\overrightarrow{h})] > \alpha_s] < \sum_{k=1}^n \sum_{s\in S,|s|=k} \frac{\delta}{2^k\binom{n}{k}} < \delta$ □

Provided that the clusters are large ($k$ is at least some constant fraction of $n$), it is clear that this bound is not too much worse than that provided by corollary 3. In particular, we still expect the compressed size of the hypotheses to be the dominating term, for each fixed $\varepsilon_s$ and $\delta$.

## 4. Discussion

We have shown that there is a positive relationship between the degree of compression achieved in the hypotheses across various learning tasks and the generalization behavior of these hypotheses. In particular, we have shown that it is possible to algorithmically make $\varepsilon - \delta$ PAC learning style guarantees in a limited setting with sample complexity requirements that feature a desirable "$1/n$ factor," demonstrating that provided that our tasks are similar (in the sense that the hypotheses, for example, exhibit compression) then grouping the tasks together can "share information." Theorem 3 and corollary 3, in particular, show that we can compute an upper bound on generalization behavior based on independently constructed initial hypotheses, making progress on a question posed by Baxter (Baxter, 2000) and allowing us to cluster hypotheses together in hopes of finding groups that compress well algorithmically. The guarantees are sufficiently strong to set the stage for an Occam's Razor style algorithm, since the upper bound it provides applies to any simpler joint hypotheses we find. More generally, these results are interesting because they suggest that it may be possible to intuitively formulate "task relatedness" in the language of Occam's Razor as "tasks are related if their joint description is substantially shorter than the sum of the lengths of their individual descriptions."

And yet, unfortunately, we have not shown much more than the above. Our restrictions to deterministic algorithms and finite sample spaces are severe, and it isn't clear how we might hope to get around these restrictions. Although the behavior of the sample complexity in theorem 3 is tantilizing, the requirements on the number of tasks are prohibitive, and the bounds are almost assuredly quite loose in practice. The Asymptotic Equipartition Property (together with Baxter's lemmas (Baxter, 2000)) has yielded some interesting results, but it seems that to get any sharper results, we will have to find a new trick. The most serious deficiency of the present work is our lack of lower bounds in this setting. At present, we have no idea of how much better we should expect to do, partially due to our lack of nontrivial examples of infinite families of related tasks, but more significantly with respect to task-relatedness, we have established a sufficient condition, but lack necessary conditions.

In other words, we can show that that under certain conditions that compressibility suffices for good gen-

eralization behavior, but we would really like to have conditions that must be satisfied, so that we can algorithmically *reject* groupings of tasks. As Caruana (Caruana, 1997) noticed, unrelated tasks can negatively affect performance, so it is highly desirable to have a method for rejecting groupings of unrelated tasks, or to understand better under what conditions including additional tasks degrades performance. Unfortunately, we have nothing at present to say on this topic. Compressibility is an intuitively appealing measure of relatedness, and seems to capture the sense of relatedness described by Caruana, but until we have shown a firm quantitative relationship, our intuition has accomplished nothing.

Intuition aside, we might expect to be able to use compressibility as a measure of relatedness due to the relationships between compressibility and learning in the single task setting. A potentially fruitful way of framing the relationship between compression and relatedness is to investigate under what conditions a multitask learning algorithm can be used to provide compression, in the style of Schapire (Schapire, 1990), Board and Pitt (Board & Pitt, 1990), Li et al (Li et al., 2003b), and many others. It is entirely possible that existing results for the single task setting can be tweaked in some way to obtain interesting results for the multitask setting. Still, it is not clear that such a result, interesting though it may be, would yield algorithmically testable conditions—after all, it is not possible to algorithmically provide nontrivial lower bounds on the compressibility of strings (Li & Vitányi, 1997).

Another possibility for providing necessary conditions on task-relatedness is investigating the compression-based similarity metrics of Li et al (Li et al., 2003a). Although the Kolmogorov-complexity based metrics suffer from not being approximable from either above or below, their "Normalized Compression Distance" is used as an approximation to an ideal distance in a way that served as partial inspiration for the present work. While it's unclear whether or not anything of substance can be proved directly about the NCD, the fact that it can be evaluated in practice makes it worth seriously investigating. Even in the ideal setting, it would be interesting to see if it is possible to demonstrate a relationship between the compression distances of the hypotheses or data sets of different tasks and the effect of sharing data between those tasks.

## Acknowledgements

Supported by an Akamai Presidential Fellowship. Thanks to Leslie Kaelbling, David Sontag, Madhu Sudan, and the anonymous reviewers for their input.

## References

Baxter, J. (2000). A model of inductive bias learning. *J. Artif. Intell. Res. (JAIR)*, *12*, 149–198.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Inf. Process. Lett.*, *24*, 377–380.

Board, R., & Pitt, L. (1990). On the necessity of occam algorithms. *STOC '90: Proceedings of the twenty-second annual ACM symposium on Theory of computing* (pp. 54–63). New York, NY, USA: ACM Press.

Caruana, R. (1997). Multitask learning. *Mach. Learn.*, *28*, 41–75.

Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Inf. Comput.*, *100*, 78–150.

Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. (2003a). The similarity metric. *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 863–872). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Li, M., Tromp, J., & Vitányi, P. (2003b). Sharpening occam's razor. *Inf. Process. Lett.*, *85*, 267–274.

Li, M., & Vitányi, P. (1997). *An introduction to kolmogorov complexity and its applications (2nd ed.)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

MacKay, D. J. C. (2002). *Information theory, inference & learning algorithms*. New York, NY, USA: Cambridge University Press.

Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*, 197–227.

Silver, D. L., & Mercer, R. E. (1998). The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness. In *Learning to learn*, 213–233. Norwell, MA, USA: Kluwer Academic Publishers.

Thrun, S., & O'Sullivan, J. (1996). Discovering structure in multiple learning tasks: The TC algorithm. *ICML* (pp. 489–497).