Risk Bounds for Transferring Representations With and Without Fine-Tuning

Daniel McNamara [1] **Maria-Florina Balcan** [2]

Abstract

A popular machine learning strategy is the transfer of a representation (i.e. a feature extraction function) learned on a source task to a target task. Examples include the re-use of neural network weights or word embeddings. We develop sufficient conditions for the success of this approach. If the representation learned from the source task is fixed, we identify conditions on how the tasks relate to obtain an upper bound on target task risk via a VC dimension-based argument. We then consider using the representation from the source task to construct a prior, which is fine-tuned using target task data. We give a PAC-Bayes target task risk bound in this setting under suitable conditions. We show examples of our bounds using feedforward neural networks. Our results motivate a practical approach to weight transfer, which we validate with experiments.

1. Introduction

A widely used machine learning technique is the transfer of a representation learned from a source task, for which labeled data is abundant, to a target task, for which labeled data is scarce. This may be effective if the tasks approximately share an intermediate representation. For example:

- features learned from an image of a human face to predict age may also be useful for predicting gender
- word embeddings learned to predict word contexts may also be useful for part of speech tagging
- features learned from financial data to predict loan default may also be useful for predicting insurance fraud.

Often a representation is learned by a different organization that may have greater access to data, computational and human resources. Examples are the Google word2vec package (Mikolov et al., 2013), and downloadable pre-trained neural networks.[1] Under this ‘representation-as-a-service’ model, a user may expect to access the representation itself, as well as information about its performance on the source task data on which it was trained. We aim to convert this into a guarantee of the usefulness of the representation on other tasks, which is known *in advance* without the effort or cost of testing the representation on the target task(s). Our analysis also covers the case where the source task is constructed from unlabeled data, as in neural network unsupervised pre-training.

We consider two approaches to transferring a representation learned from a source task to a target task, as shown in Figure 1. We may either treat the representation as fixed, or we may narrow the class of representations considered on the target task, which we refer to as *fine-tuning*. The fixed option may be attractive when very little labeled target task data is available and hence overfitting is a strong concern, while the advantage of fine-tuning is relatively greater hypothesis class expressiveness.

Let X, Y and Z be sets known as the input, output and feature spaces respectively. Let F be a class of *representations*, where $f : X \rightarrow Z$ for $f \in F$. Let G be a class of *specialized classifiers*, where $g : Z \rightarrow Y$ for $g \in G$. Let the hypothesis class $H := \{h : \exists f \in F, g \in G$ such that $h = g \circ f\}$. Let $h_S, h_T : X \rightarrow Y$ be the labeling functions and P_S, P_T be the input distributions for source task S and target task T respectively. We consider the setting $Y = \{-1, 1\}$. Let the risk of a hypothesis h on S and T be $R_S(h) := \mathbb{E}_{x \sim P_S}[h_S(x) \neq h(x)]$ and $R_T(h) := \mathbb{E}_{x \sim P_T}[h_T(x) \neq h(x)]$ respectively. Let $\hat{R}_S(h)$ and $\hat{R}_T(h)$ be the corresponding empirical (i.e. training set) risks. We have m_S labelled points for S and m_T labelled points for T. Let d_H be the VC dimension of H.

The remainder of the paper is structured as follows. In Section 2 we introduce related work. In Sections 3 and 4 we analyze the cases where the transferred representation is fixed and fine-tuned respectively. In Section 5 we apply the results and use them to motivate and test a practical approach to weight transfer in neural networks. We conclude in Section 6 and defer more involved proofs to Section 7.

[1]The Australian National University and Data61, Canberra, ACT, Australia [2]Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Daniel McNamara <daniel.mcnamara@anu.edu.au>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[1] See `http://code.google.com/archive/p/word2vec`, `http://caffe.berkeleyvision.org/model_zoo` and `http://vlfeat.org/matconvnet/pretrained` for examples.

2. Background

Empirical studies have shown the success of transferring representations between tasks (Donahue et al., 2014; Hoffman et al., 2014; Girshick et al., 2014; Socher et al., 2013; Bansal et al., 2014). Word embeddings learned on a source task have been shown (Qu et al., 2015) to perform better than unigram features on target tasks such as part of speech tagging, and comparably or better than embeddings fine-tuned on the target task. Yosinski et al. (2014) learned neural network weights using half of the ImageNet classes, and then learned the other classes with a neural network initialized with these weights, finding a benefit compared to random initialization only with target task fine-tuning. The transfer of representations, both with and without fine-tuning, is widely and successfully used.

Previous work on domain adaptation (Ben-David et al., 2010; Mansour et al., 2009; Germain et al., 2013) has considered learning a hypothesis h on S and re-using it on T, bounding $R_T(h)$ using $R_S(h)$ (measured with labeled source data) and some notion of similarity between P_S and P_T (measured with additional unlabeled target data). Such results motivate a joint optimization using labeled source and unlabeled target data (Ganin et al., 2016; Long et al., 2015) to learn separate mappings $f_S, f_T : X \to Z$ which make the induced distributions in the feature space Z similar, and a hypothesis $g : Z \to Y$ learned from the source labels which can be re-used on T. This approach assumes the tasks become the same if their input distributions can be aligned. We consider a relaxation where the tasks are more weakly related but some representation step can be transferred. We consider learning $f : X \to Z$ on S, re-using it on T, and then learning $g_T : Z \to Y$ from a small amount of labeled target data. Given the widespread use of 'downloadable' representations, where f and g_T are learned separately and there is no joint optimization over source and target data, this is a realistic setting.

Work on lifelong learning relates the past performance of a representation over many tasks to its expected future performance. For a representation $f \in F$ we construct $G \circ f := \{g \circ f : g \in G\}$. Suppose there is a distribution over tasks, known as an environment. Assume several tasks from this environment have been sampled, and that for each task some hypothesis in $G \circ f$ has been selected and its empirical risk evaluated. Previous work has provided bounds on the difference between the average empirical risk and the expected risk of the best hypothesis in $G \circ f$ for a new task drawn from the environment. Such bounds have been given by measuring the complexity of F and G using covering numbers (Baxter, 2000), a variant of the growth function (Galanti et al., 2016), and a distribution-dependent measure known as Gaussian complexity (Maurer et al., 2016). All of these bounds rely on known past performance on a large number of tasks.[2] In practice, however, representations such as neural network weights or word embeddings are often learned using only a single source task, which is the setting we consider.

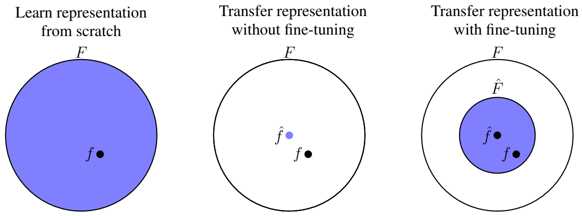
Learn representation from scratch
F
f
Transfer representation without fine-tuning
F
$\hat{f}$
f
Transfer representation with fine-tuning
F
$\hat{F}$
$\hat{f}$
f

Figure 1. A comparison of approaches to learning a representation on a target task, where the search space in each case is the shaded area. Learning from scratch, we search a representation class F for a good representation $f \in F$. Without fine-tuning, we fix a representation $\hat{f}$ learned from the source task. With fine-tuning, we narrow the search to $\hat{F} \subseteq F$ near $\hat{f}$, which still contains f.

3. Representation Fixed by Source Task

Suppose labeled source data is abundant, labeled target data is scarce, and we believe the tasks share a representation. A natural approach to leveraging the source data is to learn $\hat{g}_S \circ \hat{f} \in H$ on S, from which we assume we may extract $\hat{f} \in F$,[3] then conduct empirical risk minimization over $G \circ \hat{f} := \{g \circ \hat{f} : g \in G\}$ on T yielding $\hat{g}_T \circ \hat{f}$. Theorem 1 upper-bounds $R_T(\hat{g}_T \circ \hat{f})$ using four terms: a function ω measuring a transferrability property obtained analytically from the problem setting, the empirical risk $\hat{R}_S(\hat{g}_S \circ \hat{f})$, the generalization error of a hypothesis in H learned from m_S samples, and the generalization error of a hypothesis in G learned from m_T samples. The value of the theorem is that if $\omega(R) = O(R)$, $\hat{R}_S(\hat{g}_S \circ \hat{f})$ is a small constant, $m_S \gg m_T$ and $d_H \gg d_G$,[4] we improve on the VC dimension-based bound for learning T from scratch by avoiding the generalization error of a hypothesis in H learned from m_T samples. Furthermore, we do not settle for bounding $R_T(\hat{g}_T \circ \hat{f})$ in terms of $\hat{R}_T(\hat{g}_T \circ \hat{f})$, which may be large. The theorem can be used to select S given

[2] Pentina & Lampert (2014) extend this analysis to stochastic hypotheses (i.e. distributions over deterministic hypotheses), where for each task we learn a posterior given a prior and training data. The quality of the prior affects the learner's performance. The study proposes using source tasks to learn a 'hyperposterior', a distribution over priors which is sampled to give a prior for each task. Such a hyperposterior may focus the learner on a representation shared across tasks. The study gives a PAC-Bayes bound on the expected risk of using a hyperposterior to learn a new task drawn from the environment, in terms of the average empirical risk obtained using the hyperposterior to learn the source tasks.

[3] This is not possible with knowledge of $\hat{g}_S \circ \hat{f}$ alone, but in the case of feedforward neural networks which we focus on, $\hat{f}$ is known if the weights learned on S are known.

[4] We have $m_S \gg m_T$ if labeled source task data is abundant while labeled target task data is scarce, and $d_H \gg d_G$ if we simplify target task learning by substantially reducing the hypothesis space to be searched.

several options. While we refer to ω in a general form, we give an example in Section 3.1 and expect that others exist.[5]

Theorem 1. *Let $\omega : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function. Suppose P_S, P_T, h_S, h_T, $\hat{f}$, G have the property that $\forall \hat{g}_S \in G$, $\min_{g \in G} R_T(g \circ \hat{f}) \leq \omega(R_S(\hat{g}_S \circ \hat{f}))$. Let $\hat{g}_T := \arg\min_{g \in G} \hat{R}_T(g \circ \hat{f})$. Then with probability at least $1 - \delta$ over pairs of training sets for tasks S and T, $R_T(\hat{g}_T \circ \hat{f}) \leq \omega(\hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2d_H \log(2em_S/d_H) + 2\log(8/\delta)}{m_S}}) + 4\sqrt{\frac{2d_G \log(2em_T/d_G) + 2\log(8/\delta)}{m_T}}$.*

Proof. Let $g_T^* := \arg\min_{g \in G} R_T(g \circ \hat{f})$. With probability at least $1 - \delta$,

$$
\begin{aligned}
& R_T(\hat{g}_T \circ \hat{f}) \\
& \leq \hat{R}_T(\hat{g}_T \circ \hat{f}) + 2\sqrt{\frac{2d_G \log(2em_T/d_G) + 2\log(8/\delta)}{m_T}} \\
& \leq \hat{R}_T(g_T^* \circ \hat{f}) + 2\sqrt{\frac{2d_G \log(2em_T/d_G) + 2\log(8/\delta)}{m_T}} \\
& \leq R_T(g_T^* \circ \hat{f}) + 4\sqrt{\frac{2d_G \log(2em_T/d_G) + 2\log(8/\delta)}{m_T}} \\
& \leq \omega(R_S(\hat{g}_S \circ \hat{f})) + 4\sqrt{\frac{2d_G \log(2em_T/d_G) + 2\log(8/\delta)}{m_T}} \\
& \leq \omega(\hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2d_H \log(2em_S/d_H) + 2\log(8/\delta)}{m_S}}) + 4\sqrt{\frac{2d_G \log(2em_T/d_G) + 2\log(8/\delta)}{m_T}}.
\end{aligned}
$$

Using m training points and a hypothesis class of VC dimension d, with probability at least $1 - \delta$, for all hypotheses h simultaneously, the risk $R(h)$ and empirical risk $\hat{R}(h)$ satisfy $|R(h) - \hat{R}(h)| \leq 2\sqrt{\frac{2d\log(2em/d) + 2\log(4/\delta)}{m}}$ (Mohri et al., 2012). For G this yields the first and third inequalities with probability at least $1 - \frac{\delta}{2}$. For H, because ω is non-decreasing, this yields the fifth inequality with probability at least $1 - \frac{\delta}{2}$. Applying the union bound achieves the desired result. The second inequality is by the definition of $\hat{g}_T$ and the fourth inequality follows from our assumption. □

3.1. Neural Network Example with Fixed Representation

In Theorem 2, we give an example of the property required by Theorem 1 which is specific to a particular problem setting. We consider a neural network with a single hidden layer (see Figure 2). We propose transferring the lower-level weights (corresponding to $\hat{f}$) learned on S, so that only the upper-level weights (corresponding to G) have to be learned on T. We want to show $\hat{f}$ is also useful for T.

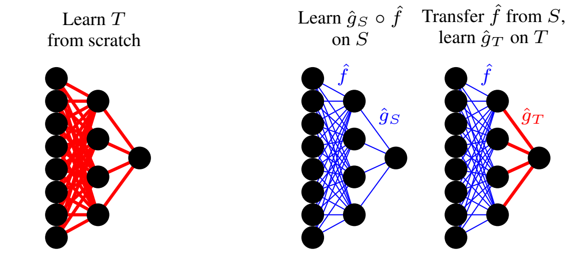
Learn T from scratch
Learn $\hat{g}_S \circ \hat{f}$ on S
Transfer $\hat{f}$ from S, learn $\hat{g}_T$ on T
$\hat{f}$
$\hat{g}_S$
$\hat{f}$
$\hat{g}_T$

Figure 2. Neural network example learning T from scratch (left) and with weights transferred from S (right). Thin blue and thick red lines show weights trained on S and T respectively. Under certain assumptions using weight transfer yields low risk on T.

To do this, we assume that some lower-level weights perform well on both tasks, which is clearly a necessary condition for the *specific* $\hat{f}$ we are transferring to perform well on both tasks. We also assume P_S and P_T have the *relative rotation invariance* property and that the upper-level weights have fixed magnitude. This is so that a point x for which $\hat{f}(x)$ contributes to the risk on T cannot be 'hidden' from the risk of using $\hat{f}$ on S, either through low $P_S(x)$ or low magnitude upper-level weights. Hence $R_S(\hat{g}_S \circ \hat{f})$ reliably indicates the usefulness of $\hat{f}$ on T.

Let $X = \mathbb{R}^n$ and $Z = \mathbb{R}^k$. Let F be the function class such that $f(x) = [a(w_1 \cdot x), \ldots, a(w_k \cdot x)]$, where $w_i \in \mathbb{R}^n$ for $1 \leq i \leq k$, $a : \mathbb{R} \to \mathbb{R}$ is an odd function[6] and $\cdot$ is the dot product. Let G be the function class such that $g(z) = sign(v \cdot z)$, where $v \in \{-1, 1\}^k$. Suppose $\exists f \in F, g_S, g_T \in G$ such that $\max[R_S(g_S \circ f), R_T(g_T \circ f)] \leq \epsilon$. Let $\hat{f}(x) := [a(\hat{w}_1 \cdot x), \ldots, a(\hat{w}_k \cdot x)]$. Given w_i and $\hat{w}_i$, pick nonzero constants α_i and β_i such that $||w_i|| = ||\alpha_i \hat{w}_i - \beta_i w_i||$ and $w_i \cdot (\alpha_i \hat{w}_i - \beta_i w_i) = 0$. Let M be a $2k \times n$ matrix with rows $w_1, \alpha_1 \hat{w}_1 - \beta_1 w_1, \ldots, w_k, \alpha_k \hat{w}_k - \beta_k w_k$. Suppose M is full rank.[7] Suppose $\forall x, x'$ such that $||Mx|| = ||Mx'||$, $P_T(x) \leq cP_S(x')$ for some $c \geq 1$, which we call *relative rotation invariance* and implies P_S and P_T have the same support. If M is an orthogonal matrix then $\forall x, x'$ such that $||x|| = ||x'||$, $P_T(x) \leq cP_S(x')$.[8]

Theorem 2. *Let $\omega(R) := cR + \epsilon(1 + c)$. Then $\forall \hat{g}_S \in G$, $\min_{g \in G} R_T(g \circ \hat{f}) \leq \omega(R_S(\hat{g}_S \circ \hat{f}))$.*

[5] We define ω by relating $R_S(\hat{g}_S \circ \hat{f})$ to $\min_{g \in G} R_T(g \circ \hat{f})$, since we expect this may be feasible analytically as in our example in Section 3.1. However, because we only observe $\hat{R}_S(\hat{g}_S \circ \hat{f})$, in Theorem 1 we use this to bound $R_S(\hat{g}_S \circ \hat{f})$ and then apply ω.

[6] i.e. $a(-x) = -a(x)$. Examples are tanh, sign and identity.

[7] To see that this condition is necessary, consider the following example where M is not full rank. Let $n = 4, k = 2$, $h_S = sign(x_1)$ and $h_T = sign(x_2)$. For $f(x) = [x_1 + x_2, x_1 - x_2]$, $g_S(z) = sign(z_1 + z_2)$ and $g_T(z) = sign(z_1 - z_2)$, we have $R_S(g_S \circ f) = R_T(g_T \circ f) = 0$. On S we learn $\hat{f}(x) = [x_1 + x_3, x_1 - x_3]$ and $\hat{g}_S(z) = sign(z_1 + z_2)$, so that $R_S(\hat{g}_S \circ \hat{f}) = 0$ but in general $\min_{g \in G} R_T(g \circ \hat{f}) > 0$ since $\hat{f}$ ignores x_2.

[8] For example, P_S and P_T are spherical Gaussians. For a zero-mean multivariate Gaussian distribution this is achieved by the whitening transformation $x \to \Lambda^{-1/2} U^T x$, where the columns of U and entries of the diagonal matrix Λ are the eigenvectors and eigenvalues of the distribution's covariance matrix respectively.

4. Representation Fine-Tuned on Target Task

Consider learning $\hat{g}_S \circ \hat{f}$ on S, and then using $\hat{f}$ and $R_S(\hat{g}_S \circ \hat{f})$ to find $\hat{F} \subseteq F$, as in Figure 1. Let $\tilde{h}_{g \circ f}$ be a stochastic hypothesis (i.e. a distribution over H) associated with $g \circ f$ (e.g. $g \circ f$ is the mode of $\tilde{h}_{g \circ f}$). We propose learning T with the hypothesis class $\tilde{H}_{G \circ \hat{F}} := \{\tilde{h}_{g \circ f} : f \in \hat{F}, g \in G\}$ and the prior $\tilde{h}_{\hat{g}_S \circ \hat{f}}$. Learning T from scratch we assume that we would instead use $\tilde{H}_{G \circ F} := \{\tilde{h}_{g \circ f} : f \in F, g \in G\}$ and some fixed prior $\tilde{h}_0 \in \tilde{H}_{G \circ F}$. Let $R_T(\tilde{h}) := \mathbb{E}_{x \sim P_T, h \sim \tilde{h}}[h_T(x) \neq h(x)]$ and compute $\hat{R}_T(\tilde{h})$ on the training set distribution of T.

In Theorem 3 we show that if $\hat{F}$ is 'small enough' so that all $\tilde{h} \in \tilde{H}_{G \circ \hat{F}}$ have a small KL divergence from $\tilde{h}_{\hat{g}_S \circ \hat{f}}$, we may apply a PAC-Bayes bound to the generalization error of hypotheses in $\tilde{H}_{G \circ \hat{F}}$ involving four terms: a function ω measuring a transferrability property, the empirical risk $\hat{R}_S(\hat{g}_S \circ \hat{f})$, the generalization error of a hypothesis in H learned from m_S points, and a weak dependence on m_T. The value of the theorem is that if $\omega(R) = O(R)$, $\hat{R}_S(\hat{g}_S \circ \hat{f})$ is a small constant, and $m_S \gg m_T$, we improve on the PAC-Bayes bound for $\tilde{H}_{G \circ F}$ and $\tilde{h}_0$.[9] $\hat{F}$ is useful if it is also 'large enough' in the sense that $\exists \tilde{h}_{g_T \circ f} \in \tilde{H}_{G \circ \hat{F}}$ such that $R_T(\tilde{h}_{g_T \circ f}) \leq \epsilon$. Here ω quantifies how large the $\hat{F}$ we search on T must be in order to be 'large enough', in terms of $R_S(\hat{g}_S \circ \hat{f})$. While in general such an $\hat{F}$ and ω may not exist, we give an example in Section 4.1.

Theorem 3. *Let $\omega : \mathbb{R} \to \mathbb{R}$ be non-decreasing. Suppose given $\hat{f} \in F$ and $R_S(\hat{g}_S \circ \hat{f})$ estimated from S, it is possible to construct $\hat{F}$ with the property $\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}$, $KL(\tilde{h} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \omega(R_S(\hat{g}_S \circ \hat{f}))$. Then with probability at least $1 - \delta$ over pairs of training sets for tasks S and T, $\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}$, $R_T(\tilde{h}) \leq \hat{R}_T(\tilde{h}) + \sqrt{\frac{\omega(\hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2d_H \log(2em_S/d_H) + 2\log(8/\delta)}{m_S}}) + \log 2m_T/\delta}{2(m_T - 1)}}$.*

Proof. With probability at least $1 - \delta$,

$$
\begin{aligned}
&R_T(\tilde{h}) \\
&\leq \hat{R}_T(\tilde{h}) + \sqrt{\frac{KL(\tilde{h} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) + \log 2m_T/\delta}{2(m_T - 1)}} \\
&\leq \hat{R}_T(\tilde{h}) + \sqrt{\frac{\omega(R_S(\hat{g}_S \circ \hat{f})) + \log 2m_T/\delta}{2(m_T - 1)}}.
\end{aligned}
$$

The first inequality holds with probability at least $1 - \frac{\delta}{2}$ (Shalev-Shwartz & Ben-David, 2014). The second inequality holds by assumption. Furthermore, $R_S(\hat{g}_S \circ \hat{f}) \leq \hat{R}_S(\hat{g}_S \circ \hat{f}) + 2\sqrt{\frac{2d_H \log(2em_S/d_H) + 2\log(8/\delta)}{m_S}}$ with probability at least $1 - \frac{\delta}{2}$ (Mohri et al., 2012) and ω is non-decreasing. The result follows from the union bound. □

4.1. Neural Network Example with Fine-Tuning

We transfer and fine-tune weights in a feedforward neural network with one hidden layer to instantiate the property required by Theorem 3. We learn a deterministic hypothesis of this type on S and obtain k estimated lower-level weight vectors $\hat{w}_i$. Learning T we now consider only lower-level weights near $\hat{w}_i$, corresponding to $\hat{F}$. On T we learn a stochastic hypothesis formed by taking a deterministic network and adding independent sources of spherical Gaussian noise to the lower-level weights and sign-flipping noise to the upper-level weights. The KL divergence between two of the stochastic hypotheses is expressed using the angles between their lower-level weights[10] and a quantity computable from their upper-level weights.

We want to prove that we can construct such an $\hat{F}$ to successfully learn T. To do this, we assume some lower-level weights w_i perform well on both S and T. We make $\hat{F}$ 'small enough' by only including lower-level weights with small angles to $\hat{w}_i$, and 'large enough' by using the risk observed using $\hat{w}_i$ on S to provide an upper bound on the angle between each pair w_i and $\hat{w}_i$. Our assumptions ensure that poor $\hat{w}_i$ cannot be 'hidden' from the risk on S, either through low P_S density in the region of disagreement between w_i and $\hat{w}_i$, or through low magnitude higher-level weights. Hence we know that searching $\hat{F}$ will include w_i.

Let $X = \mathbb{R}^n$ and $Z = \mathbb{R}^k$, where k is odd. Let F be the function class such that $f(x) = [sign(w_1 \cdot x), \ldots, sign(w_k \cdot x)]$, where $w_i \in \mathbb{R}^n$ for $1 \leq i \leq k$. Let G be the function class such that $g(z) = sign(v \cdot z)$, where $v \in \{-1, 1\}^k$. Let B_v be a distribution on $\{-1, 1\}^k$ such that for $v' \sim B_v$, $Pr(v') = \prod_{j=1}^{k} p^{\mathbf{1}(v'_j = v_j)}(1 - p)^{\mathbf{1}(v'_j = -v_j)}$, where $p \in [0.5, 1]$. Let $\tilde{h}_{g \circ f} := g' \circ f'$ such that $v', w'_1, \ldots, w'_k \sim B_v \prod_{i=1}^{k} \mathcal{N}(w_i, \sigma^2 I)$. Suppose $\exists f \in F, g_S, g_T \in G$ such that $\max[R_S(g_S \circ f), R_T(\tilde{h}_{g_T \circ f})] \leq \epsilon$. Let $\hat{f}(x) := [sign(\hat{w}_1 \cdot x), \ldots, sign(\hat{w}_k \cdot x)]$, $\theta(w_i, \hat{w}_i)$ be the angle between w_i and $\hat{w}_i$, and assume $\forall i, ||\hat{w}_i|| = 1$. Define M as in Section 3.1. Let P_S have the *rotation invariance* property $\forall x, x'$ such that $||Mx|| = ||Mx'||$, $P_S(x) \leq cP_S(x')$ for some $c \geq 1$.

Theorem 4. *Given $\hat{f}$ and $R_S(\hat{g}_S \circ \hat{f})$ estimated from S, let $\theta_{\max} := \pi\sqrt{2(k-1)}c(R_S(\hat{g}_S \circ \hat{f}) + \epsilon)$ and $\hat{F} := \{f \in F : \forall i, ||w_i|| = 1 \wedge |\theta(w_i, \hat{w}_i)| \leq \theta_{\max}\}$. Let $\omega(R) := \frac{k}{\sigma^2}[1 - \cos\theta_{\max}] + k[2p - 1 + (1-p)^k]\log_2 \frac{p}{1-p}$. Then $\exists \tilde{h}_{g_T \circ f} \in \tilde{H}_{G \circ \hat{F}}$ such that $R_T(\tilde{h}_{g_T \circ f}) \leq \epsilon$ and $\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}$, $KL(\tilde{h} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \omega(R_S(\hat{g}_S \circ \hat{f}))$.*

[9] Using the restricted deterministic hypothesis class $G \circ \hat{F} := \{h : \exists f \in \hat{F}, g \in G$ such that $h = g \circ f\}$ and a VC dimension-based bound may not improve on H, since possibly $d_{G \circ \hat{F}} = d_H$.

[10] Assuming that the lower-level weight vectors are of fixed magnitude, which is no loss of model expressiveness since we use the sign activation function at the hidden layer.

5. Applications

We show the utility of the risk bounds, and present a novel technique and experiments motivated by our theorems.

5.1. Using the Risk Bounds

The results described yield tighter bounds on risk when transferring representations from S, compared to learning T from scratch. Examples are shown in Figure 3.[11]

We set $\delta = 0.05$. For the top part, we use the example from Section 3.1 and set $n = 10$, $k = 5$. Learning T from scratch with H, we use the bound from Mohri et al. (2012) used previously. The VC dimension of a network of $|E|$ edges using the sign activation is $O(|E| \log |E|)$ (Shalev-Shwartz & Ben-David, 2014), where in our case $|E| = nk+k$. We use $d_H = |E| \log |E|$ in the chart. Transferring a representation from S to T without fine-tuning, we consider the limit $\epsilon \to 0, \hat{R}_S(\hat{g}_S \circ \hat{f}) \to 0, m_S \to \infty$, and hence $\omega(\cdot) \to 0$ by Theorem 2. Furthermore, $d_G \leq k$ since G is finite and hence $d_G \leq \log_2 |G|$ (Shalev-Shwartz & Ben-David, 2014). We use the bound from Theorem 1.

For the bottom part, we use the example from Section 4.1 and set $\sigma^2 = \frac{1}{10}$, $k = 499$, $p = \frac{2}{3}$. Learning T from scratch we use the stochastic hypothesis class $\{\tilde{h}_{g \circ f} : f \in F$ such that $\forall i ||w_i|| = 1, g \in G\}$ and a prior $\tilde{h}_0$ where $\forall i\ w_i = \mathbf{0}$ and $v \in \{-1, 1\}^k$ is arbitrary.[12] Hence we have the bound $KL(\tilde{h}||\tilde{h}_0) \leq 10k + \frac{k}{3}$, which becomes tight for large k. We apply the PAC-Bayes bound (Shalev-Shwartz & Ben-David, 2014) used previously. Transferring a representation from S and fine-tuning on T, we consider the limit $\epsilon \to 0$, $\hat{R}_S(\hat{g}_S \circ \hat{f}) \to 0$, $m_S \to \infty$. We have $KL(\tilde{h}||\tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \frac{k}{3}$ by Theorem 4. We use the bound from Theorem 3.

5.2. Fine-Tuning through Regularization

We relax the hard constraint on $\hat{F}$ from Section 4.1 by using a modified loss function, which we find performs better in practice. Let y_i and $\hat{y}_i$ be the label and prediction respectively for the ith training point. In a fully-connected feedforward neural network with l layers of weights, let $W^{(j)}$ be the jth weight matrix, $\hat{W}^{(j)}$ be its estimate from S (excluding weights for bias units in both cases), and $||\cdot||_2$ be the entry-wise 2 norm. A typical loss function (1) used for training is composed of the sum of training set log loss and L2 regularization on the weights.

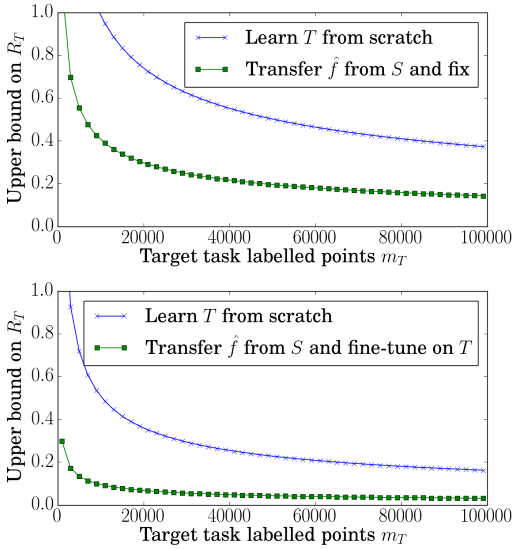

Figure 3. A comparison of risk bounds compared to learning T from scratch, without fine-tuning (top) and with fine-tuning (bottom). The two charts use different parameters (see Section 5.1).

$$\sum_{i=1}^{m}[-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)] + \frac{\lambda}{2} \sum_{j=1}^{l}(||W^{(j)}||_2^2) \tag{1}$$

We replace the regularization penalty with (2).[13]

$$\sum_{j=1}^{l}[\frac{\lambda_1(j)}{2}||W^{(j)} - \hat{W}^{(j)}||_2^2 + \frac{\lambda_2(j)}{2}||W^{(j)}||_2^2] \tag{2}$$

This penalizes estimates of W far from the representation learned on S. Since we expect the tasks to share a low-level representation (e.g. edge detectors for vision, word embeddings for text) but be distinct at higher levels (e.g. image components for vision, topics for text), we set $\lambda_1(\cdot)$ to be a decreasing function, while $\lambda_2(\cdot)$ controls standard L2 regularization. The technique is novel to our knowledge, although other approaches to transferring regularization between tasks exist (Evgeniou & Pontil, 2004; Raina et al., 2006; Argyriou et al., 2008; Ghifary et al., 2014).

5.3. Experiments

We experiment on basic image and text classification tasks.[14] We show that learning algorithms motivated by our theoretical results can help to overcome a scarcity of labeled target task data. Note that we do not replicate the conditions specified in our theorems, nor do we attempt extensive tuning to achieve state-of-the-art performance.

[11] Note that VC dimension risk bounds are known for being rather loose, while PAC-Bayesian bounds are tighter and hence yield non-trivial results in higher dimensions with fewer samples.

[12] This class is as expressive as $\tilde{H}_{G \circ F}$ but by setting $||w_i|| = 1$ the KL divergence of all hypotheses from any prior is bounded, allowing a fair comparison to $\tilde{H}_{G \circ \hat{F}}$. The choice of $\tilde{h}_0$ minimizes worst case KL divergence to a hypothesis in the class.

[13] Basing our approach on (1), we follow the convention that weights connected to bias units are excluded from the regularization penalty. However, the inclusion of these weights in the $||W^{(j)} - \hat{W}^{(j)}||$ term of (2) is a plausible variant.

[14] The MNIST and 20 Newgroups datasets are available at http://yann.lecun.com/exdb/mnist and http://qwone.com/~jason/20Newsgroups respectively.

We randomly partition label classes into sets S_+ and S_-, where $|S_+| = |S_-|$.[15] We construct T_+ by randomly picking from S_+ up to $\gamma := \frac{|S_+ \cap T_+|}{|S_+|}$, then randomly picking from S_- such that $|T_+| = |T_-|$. We let S be the task of distinguishing between S_+ and S_- and T be that of distinguishing T_+ and T_-. Constructing S_+ and T_+ as disjunctions of classes means that the class labels are a perfect representation shared between S and T.

We compare the accuracy on T of four options:

- learn T from scratch (BASE)
- transfer $\hat{f}$ from S, fine-tune f and train g on T using (2) (FINE-TUNE $\hat{f}$)
- transfer $\hat{f}$ from S and fix, train g on T (FIX $\hat{f}$)[16]
- transfer $\hat{g}_S \circ \hat{f}$ from S and fix (FIX $\hat{g}_S \circ \hat{f}$).[17]

We use $\lambda_1(1) = \lambda_2(2) = \lambda := 1$,[18] $\lambda_1(2) = \lambda_2(1) = 0$, $m_T = 500$ and the sigmoid activation function. For MNIST we use raw pixel intensities, a $784 \times 50 \times 1$ network and $m_S = 50000$. For NEWSGROUPS we use TF-IDF weighted counts of most frequent words, a $2000 \times 50 \times 1$ network and $m_S = 15000$. We use conjugate gradient optimization with 200 iterations.

The results are shown in Table 1.[19] When the tasks are non-identical, FINE-TUNE $\hat{f}$ is mostly the strongest but performs better on MNIST. FIX $\hat{f}$ outperforms BASE when $\gamma \geq 0.8$ and hence the tasks are similar. While FIX $\hat{f}$ outperforms FIX $\hat{g}_S \circ \hat{f}$ when the tasks are non-identical on MNIST, on NEWSGROUPS there is no evidence of benefit. When the tasks are identical, FIX $\hat{g}_S \circ \hat{f}$ is the strongest.

It appears that learning an MNIST digit requires a dense weight vector and so $\hat{W}^{(1)}$ tends to encode single digits, which helps transferrability. However, it appears that since we may learn a newsgroup with a sparse weight vector, $\hat{W}^{(1)}$ tends to encode disjunctions of newsgroups which somewhat reduces transferrability. When transferring representations does work, fine-tuning using the regularization penalty proposed in (2) improves performance.

Table 1. Evaluation of transferring representations. Entries are the test set accuracy of the technique (row) for the task (column) averaged over 10 trials, with the best result for each task bolded.

TECHNIQUE	MNIST, $\gamma =$			NEWSGROUPS, $\gamma =$		
	0.6	0.8	1	0.6	0.8	1
BASE	88.4	87.9	87.9	**62.6**	63.2	66.1
FINE-TUNE $\hat{f}$	**91.9**	**93.9**	95.4	62.3	**72.3**	83.3
FIX $\hat{f}$	87.5	92.3	97.3	52.2	69.6	83.3
FIX $\hat{g}_S \circ \hat{f}$	67.4	85.6	**98.1**	55.5	70.7	**83.6**

6. Conclusion

We developed sufficient conditions for the successful transfer of representations both with and without fine-tuning. This is a step towards a principled explanation of the empirical success achieved by such techniques. A promising direction for future work is generalizing the neural network architectures considered (e.g. using multiple hidden layers) and relaxing the distributional assumptions required. Furthermore, in the fine-tuning case it may be possible to upper bound the target task generalization error of hypotheses in $G \circ \hat{F} := \{h : \exists f \in \hat{F}, g \in G \text{ such that } h = g \circ f\}$ using another measure such as the Rademacher complexity of $G \circ \hat{F}$, eliminating the need for stochastic hypotheses.

We proposed a novel form of regularization for neural network training motivated by our theoretical results, which penalizes divergence from source task weights and is stricter for lower-level weights. We validated this technique through applications to image and text classification. Future directions include experiments on more challenging tasks using deeper and more tailored network architectures (e.g. convolutional neural networks).

7. Additional Proofs

We provide complete proofs of Theorems 2 and 4. For brevity, we drop the explicit dependence of f, $\hat{f}$, h_S and h_T on x in our notation where the meaning is clear.

7.1. Proof of Theorem 2

Proof. Let $g_S(z) := sign(v_S \cdot z)$, $g_T(z) := sign(v_T \cdot z)$, $\hat{g}_S(z) := sign(\hat{v}_S \cdot z)$, $\hat{g}_T(z) := sign(d * \hat{v}_S \cdot z)$, where $d := v_S * v_T \in \{-1, 1\}^k$ and $*$ is the elementwise product. It is sufficient to show $R_T(\hat{g}_T \circ \hat{f}) \leq cR_S(\hat{g}_S \circ \hat{f}) + \epsilon(1+c)$.

$$R_T(\hat{g}_T \circ \hat{f})$$

$$= Pr_{x \sim P_T}(h_T d * \hat{v}_S \cdot \hat{f} \leq 0)$$

$$\leq Pr_{x \sim P_T}(h_T d * v_S \cdot f \leq 0, d * v_S \cdot f d * \hat{v}_S \cdot \hat{f} \geq 0) + Pr_{x \sim P_T}(h_T d * v_S \cdot f \geq 0, d * v_S \cdot f d * \hat{v}_S \cdot \hat{f} \leq 0)$$

$$\leq Pr_{x \sim P_T}(h_T d * v_S \cdot f \leq 0) + Pr_{x \sim P_T}(d * v_S \cdot f d * \hat{v}_S \cdot \hat{f} \leq 0)$$

[15] For MNIST there are 10 label classes and for 20 Newgroups there are 20. In both cases the classes are approximately balanced. Note that we ignore the hierarchical structure of the 20 Newsgroups classes, which likely contributes to the lower accuracies reported for all methods for this dataset relative to MNIST.

[16] i.e. logistic regression with L2 regularization and $\hat{f}$ fixed.

[17] Used to isolate the benefit of transferring $\hat{f}$ rather than $\hat{g}_S \circ \hat{f}$.

[18] We explored tuning λ to lift the performance of BASE on MNIST, but found that the results did not materially improve. Potentially $\lambda_1(j)$ and $\lambda_2(j)$ in (2) could be tuned with cross validation on the target task.

[19] For $\gamma = 1$, $h_S = h_T$. We do not consider $\gamma < 0.5$, since that is equivalent to $1 - \gamma$ with the definitions of T_+ and T_- swapped.

$$\leq \epsilon + Pr_{x \sim P_T}(d * v_S \cdot f d * \hat{v}_S \cdot \hat{f} \leq 0)$$

$$\leq \epsilon + cPr_{x \sim P_S}(v_S \cdot f \hat{v}_S \cdot \hat{f} \leq 0)$$

$$\leq \epsilon + c[Pr_{x \sim P_S}(h_S \hat{v}_S \cdot \hat{f} \leq 0, h_S v_S \cdot f \geq 0)+$$
$$Pr_{x \sim P_S}(h_S \hat{v}_S \cdot \hat{f} \geq 0, h_S v_S \cdot f \leq 0)]$$

$$\leq \epsilon + c[Pr_{x \sim P_S}(h_S \hat{v}_S \cdot \hat{f} \leq 0) + Pr_{x \sim P_S}(h_S v_S \cdot f \leq 0)]$$

$$\leq cR_S(\hat{g}_S \circ \hat{f}) + \epsilon(1 + c).$$

The third and final inequalities are due to the shared representation assumption in the problem statement. The fourth inequality holds by Lemma 1. The remaining lines apply simple rules of probability. □

Lemma 1. *Suppose* $\forall x, x'$ *such that* $||Mx|| = ||Mx'||$, $P_T(x) \leq cP_S(x')$. *Let* $f, \hat{f} \in F$, $v, \hat{v}, d \in \{-1, 1\}^k$. *Then* $Pr_{x \sim P_T}(d * v \cdot f d * \hat{v} \cdot \hat{f} \leq 0) \leq cPr_{x \sim P_S}(v \cdot f \hat{v} \cdot \hat{f} \leq 0)$.

Proof. Suppose there is an invertible map $\mathbb{R}^n \to \mathbb{R}^n$ yielding x' on input x, such that $\forall x$, $||Mx|| = ||Mx'||$ and $d * v \cdot f(x) d * \hat{v} \cdot \hat{f}(x) = v \cdot f(x') \hat{v} \cdot \hat{f}(x')$. Then the result follows since $P_T(x) \leq cP_S(x')$ by assumption. Furthermore, if M is an orthogonal matrix, $||x|| = ||x'||$.

Such a map is $x' := (M^T M)^{-1} M^T \tilde{d} * (Mx)$, where $\tilde{d} := [d_1, d_1, \ldots, d_k, d_k]$. We have $\forall i$, $w_i \cdot x' = d_i w_i \cdot x$ and $(\alpha_i \hat{w}_i - \beta_i w_i) \cdot x' = d_i(\alpha_i \hat{w}_i - \beta_i w_i) \cdot x$, and hence $\hat{w}_i \cdot x' = d_i \hat{w}_i \cdot x$ for $\alpha_i, \beta_i \neq 0$. Therefore:

$$d * v \cdot f(x) d * \hat{v} \cdot \hat{f}(x)$$

$$= v \cdot d * f(x) \hat{v} \cdot d * \hat{f}(x)$$

$$= v \cdot f(x') \hat{v} \cdot d * \hat{f}(x)$$

$$= v \cdot f(x') \hat{v} \cdot \hat{f}(x').$$

The first equality is a property of the elementwise and dot products. For the second equality, $a(w_i \cdot x') = a(d_i w_i \cdot x) = d_i a(w_i \cdot x)$ since a is an odd function. Similarly, for the third equality $a(\hat{w}_i \cdot x') = a(d_i \hat{w}_i \cdot x) = d_i a(\hat{w}_i \cdot x)$. □

7.2. Proof of Theorem 4

Proof of $\exists \tilde{h}_{g_T \circ f} \in \tilde{H}_{G \circ \hat{F}}$ *such that* $R_T(\tilde{h}_{g_T \circ f}) \leq \epsilon$.
Recall that w_i are the weight vectors for f and $\hat{w}_i$ are those for $\hat{f}$. Observe that for any w_i such that $w_i \cdot \hat{w}_i < 0$, we have $-w_i \cdot \hat{w}_i > 0$ and $-v_i sign(-w_i \cdot x) = v_i sign(w_i \cdot x)$. Combining this with the assumption $\min\limits_{f \in F, g_S, g_T \in G} \max[R_S(g_S \circ f), R_T(g_T \circ f)] \leq \epsilon$, we conclude $\exists f \in F, g_S, g_T \in G$ such that $\forall i, w_i \cdot \hat{w}_i \geq 0$ and $\max[R_S(g_S \circ f), R_T(\tilde{h}_{g_T \circ f})] \leq \epsilon$.

Let $g_S(z) := sign(v_S \cdot z)$ and $\hat{g}_S(z) := sign(\hat{v}_S \cdot z)$. Let P be a rotation invariant distribution for $c = 1$. To prove $\tilde{h}_{g_T \circ f} \in \tilde{H}_{G \circ \hat{F}}$, by the definition of $\tilde{H}_{G \circ \hat{F}}$ it is sufficient to show $\forall i, |\theta(w_i, \hat{w}_i)| \leq \pi\sqrt{2(k-1)}c(R_S(\hat{g}_S \circ \hat{f}) + \epsilon)$.

$$\frac{\max\limits_i |\theta(w_i, \hat{w}_i)|}{\pi\sqrt{2(k-1)}}$$

$$\leq Pr_{x \sim P}(v_S \cdot f v_S \cdot \hat{f} \leq 0)$$

$$\leq Pr_{x \sim P}(v_S \cdot f \hat{v}_S \cdot \hat{f} \leq 0)$$

$$\leq cPr_{x \sim P_S}(v_S \cdot f \hat{v}_S \cdot \hat{f} \leq 0)$$

$$\leq c[Pr_{x \sim P_S}(h_S v_S \cdot f \leq 0, h_S \hat{v}_S \cdot \hat{f} \geq 0)+$$
$$Pr_{x \sim P_S}(h_S v_S \cdot f \geq 0, h_S \hat{v}_S \cdot \hat{f} \leq 0)]$$

$$\leq c[Pr_{x \sim P_S}(h_S v_S \cdot f \leq 0) + Pr_{x \sim P_S}(h_S \hat{v}_S \cdot \hat{f} \leq 0)]$$

$$\leq c[\epsilon + R_S(\hat{g}_S \circ \hat{f})].$$

The first inequality holds by Lemma 2. The second inequality holds by Lemma 3, using the fact $\forall i, w_i \cdot \hat{w}_i \geq 0$. The third inequality uses the rotation invariance of P_S. The following two lines use basic laws of probability. The final inequality uses the assumption $R_S(g_S \circ f) \leq \epsilon$. □

Proof of $\forall \tilde{h} \in \tilde{H}_{G \circ \hat{F}}, KL(\tilde{h} || \tilde{h}_{\hat{g}_S \circ \hat{f}}) \leq \omega(R_S(\hat{g}_S \circ \hat{f}))$.
For any $\tilde{h}_{g \circ f} \in \tilde{H}_{G \circ \hat{F}}$, $KL(\tilde{h}_{g \circ f} || \tilde{h}_{\hat{g}_S \circ \hat{f}})$

$$= \sum_{i=1}^k [KL(\mathcal{N}(w_i, \sigma^2 I) || \mathcal{N}(\hat{w}_i, \sigma^2 I))] + KL(B_v || B_{\hat{v}_S}).$$

The KL divergence of a product distribution is the sum of the KL divergences of its component distributions. We upper bound both terms and apply the definition of ω.

$$\sum_{i=1}^k KL(\mathcal{N}(w_i, \sigma^2 I) || \mathcal{N}(\hat{w}_i, \sigma^2 I))$$

$$= \tfrac{1}{2\sigma^2} \sum_{i=1}^k ||w_i - \hat{w}_i||^2$$

$$= \tfrac{1}{2\sigma^2} \sum_{i=1}^k (||w_i||^2 + ||\hat{w}_i||^2 - 2||w_i|| ||\hat{w}_i|| \cos|\theta(w_i, \hat{w}_i)|)$$

$$= \tfrac{1}{\sigma^2} \sum_{i=1}^k (1 - \cos|\theta(w_i, \hat{w}_i)|)$$

$$\leq \tfrac{k}{\sigma^2}[1 - \cos(\pi\sqrt{2(k-1)}c(R_S(\hat{g}_S \circ \hat{f}) + \epsilon))].$$

The first equality uses the KL divergence of Gaussian distributions. The second equality uses the law of cosines. The third equality is because $\forall i, ||w_i|| = ||\hat{w}_i|| = 1$ by construction. The inequality follows by the definition of $\hat{F}$ and the fact that $1 - \cos|\theta|$ is non-decreasing for $|\theta| \in [0, \pi]$.

$$KL(B_v || B_{\hat{v}_S})$$

$$\leq \sum_{i=1}^k \tbinom{k}{i} p^i (1-p)^{k-i} \log_2 \tfrac{p^i (1-p)^{k-i}}{(1-p)^i p^{k-i}}$$

$$= k[2p - 1 + (1-p)^k] \log_2 \tfrac{p}{1-p}.$$

The first inequality uses the definition of B_v to express $KL(B_v || B_{\hat{v}_S})$. The equality is a simplification. □

Lemma 2. *Suppose k is odd, $v \in \{-1,1\}^k$, $f, \hat{f} \in F$ such that $\forall i$, $w_i \cdot \hat{w}_i \geq 0$ and P is rotation invariant with $c = 1$. Then* $\frac{\max_i |\theta(w_i, \hat{w}_i)|}{\pi\sqrt{2(k-1)}} \leq Pr_{x\sim P}(v \cdot fv \cdot \hat{f} \leq 0)$.

Proof. Let $v_{-j} := [v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k]$ and define f_{-j} and $\hat{f}_{-j}$ similarly. Let $Pr(\cdot) := Pr_{x\sim P}(\cdot)$.

$$Pr(v \cdot fv \cdot \hat{f} \leq 0)$$

$$\geq Pr(v \cdot fv \cdot \hat{f} < 0)$$

$$\geq Pr(v_{-j} \cdot f_{-j} = 0) Pr(v \cdot fv \cdot \hat{f} < 0 | v_{-j} \cdot f_{-j} = 0)$$

$$= Pr(v_{-j} \cdot f_{-j} = 0) \\ Pr(v_j f_j v_{-j} \cdot \hat{f}_{-j} + f_j \hat{f}_j < 0 | v_{-j} \cdot f_{-j} = 0)$$

$$= Pr(v_{-j} \cdot f_{-j} = 0) \\ [Pr(v_j f_j v_{-j} \cdot \hat{f}_{-j} < -1, f_j \hat{f}_j = 1 | v_{-j} \cdot f_{-j} = 0) + \\ Pr(v_j f_j v_{-j} \cdot \hat{f}_{-j} < 1, f_j \hat{f}_j = -1 | v_{-j} \cdot f_{-j} = 0)]$$

$$\geq Pr(v_{-j} \cdot f_{-j} = 0) \\ [Pr(v_j f_j v_{-j} \cdot \hat{f}_{-j} < -1, f_j \hat{f}_j = -1 | v_{-j} \cdot f_{-j} = 0) + \\ Pr(v_j f_j v_{-j} \cdot \hat{f}_{-j} < 1, f_j \hat{f}_j = -1 | v_{-j} \cdot f_{-j} = 0)]$$

$$= Pr(v_{-j} \cdot f_{-j} = 0) \\ [Pr(v_j f_j v_{-j} \cdot \hat{f}_{-j} < -1, f_j \hat{f}_j = -1 | v_{-j} \cdot f_{-j} = 0) + \\ Pr(v_j f_j v_{-j} \cdot \hat{f}_{-j} > -1, f_j \hat{f}_j = -1 | v_{-j} \cdot f_{-j} = 0)]$$

$$= Pr(v_{-j} \cdot f_{-j} = 0) Pr(f_j \hat{f}_j = -1 | v_{-j} \cdot f_{-j} = 0)$$

$$= Pr(v_{-j} \cdot f_{-j} = 0) Pr(f_j \hat{f}_j = -1)$$

$$= \binom{k-1}{\frac{k-1}{2}} \left(\tfrac{1}{2}\right)^{k-1} \frac{|\theta(w_j, \hat{w}_j)|}{\pi}$$

$$\geq \frac{2^{k-1}}{\sqrt{2(k-1)}} \left(\tfrac{1}{2}\right)^{k-1} \frac{|\theta(w_j, \hat{w}_j)|}{\pi}$$

$$\geq \frac{\max_i |\theta(w_i, \hat{w}_i)|}{\pi\sqrt{2(k-1)}}.$$

The third inequality follows since P is rotation invariant and $w_j \cdot \hat{w}_j \geq 0$. The third and fifth equalities use rotation invariance. The final equality uses rotation invariance and the fact that k is odd. The fourth inequality is a standard lower bound for the central binomial coefficient. The other lines use basic simplifications and laws of probability. □

Lemma 3. *Suppose k is odd, $v, \hat{v} \in \{-1,1\}^k$, $f, \hat{f} \in F$ such that $\forall i$, $w_i \cdot \hat{w}_i \geq 0$ and P is rotation invariant with $c = 1$. Then $Pr_{x\sim P}(v \cdot fv \cdot \hat{f} \leq 0) \leq Pr_{x\sim P}(v \cdot f\hat{v} \cdot \hat{f} \leq 0)$.*

Proof. Let $Pr(\cdot) := Pr_{x\sim P}(\cdot)$ and $\mathbb{E}[\cdot] := \mathbb{E}_{x\sim P}[\cdot]$. Let $Pr(\tilde{f}) := Pr_{x\sim P}([f_1(x)\hat{f}_1(x), \ldots, f_k(x)\hat{f}_k(x)] = \tilde{f})$. Let $d := \hat{v} * v$ and $\Delta(x) := \mathbf{1}(v \cdot f(x)\hat{v} \cdot \hat{f}(x) \leq 0) - \mathbf{1}(v \cdot f(x)v \cdot \hat{f}(x) \leq 0)$. Assume $\hat{v} \neq v$ (if $\hat{v} = v$ then the lemma clearly holds). Let $a(\tilde{f}) := \sum_{i=1}^{k} \mathbf{1}(\tilde{f}_i = 1)$ and let $l := \min_{i:d_i=-1} i$. Let $\tilde{F} := \{\tilde{f} \in \{-1,1\}^k : a(\tilde{f}) > a(d * \tilde{f}) \vee (a(\tilde{f}) = a(d * \tilde{f}) \wedge \tilde{f}_l = 1)\}$.

Let $\Phi(a) := \frac{1}{2^{k-1}} \sum_{b=0}^{\lfloor k/2 \rfloor} \sum_{j=\lceil a/2+b/2-k/4 \rceil}^{b} \binom{a}{j}\binom{k-a}{b-j}$. The term b counts coordinates where $v_i \hat{f}_i = sign(v \cdot f)$, while j counts those where $v_i f_i = sign(v \cdot f)$ and $f_i = \hat{f}_i$.

$$Pr(v \cdot f\hat{v} \cdot \hat{f} \leq 0) - Pr(v \cdot fv \cdot \hat{f} \leq 0)$$

$$= \mathbb{E}[\mathbf{1}(v \cdot f\hat{v} \cdot \hat{f} \leq 0)] - \mathbb{E}[\mathbf{1}(v \cdot fv \cdot \hat{f} \leq 0)]$$

$$= \mathbb{E}[\Delta]$$

$$= \sum_{\tilde{f} \in \tilde{F}} Pr(\tilde{f})\mathbb{E}[\Delta|\tilde{f}] + Pr(d * \tilde{f})\mathbb{E}[\Delta|d * \tilde{f}]$$

$$= \sum_{\tilde{f} \in \tilde{F}} [Pr(\tilde{f}) - Pr(d * \tilde{f})]\mathbb{E}[\Delta|\tilde{f}]$$

$$= \sum_{\tilde{f} \in \tilde{F}} [Pr(\tilde{f}) - Pr(d * \tilde{f})] \\ [Pr(v \cdot fv \cdot \hat{f} \leq 0 | d * \tilde{f}) - Pr(v \cdot fv \cdot \hat{f} \leq 0 | \tilde{f})]$$

$$= \sum_{\tilde{f} \in \tilde{F}} [Pr(\tilde{f}) - Pr(d * \tilde{f})][\Phi(a(d * \tilde{f})) - \Phi(a(\tilde{f}))]$$

$$\geq 0.$$

The second equality uses linearity of expectation. The third equality uses the law of total expectation and the definition of $\tilde{F}$.

The fourth equality holds since $\mathbb{E}[\Delta|d * \tilde{f}]$ $= \sum_{f \in \{-1,1\}^k} Pr(f|d * \tilde{f})\mathbb{E}[\Delta|d * \tilde{f}, f]$ $= -\sum_{f \in \{-1,1\}^k} Pr(f|d * \tilde{f})\mathbb{E}[\Delta|\tilde{f}, f]$ $= -\sum_{f \in \{-1,1\}^k} Pr(f|\tilde{f})\mathbb{E}[\Delta|\tilde{f}, f] = -\mathbb{E}[\Delta|\tilde{f}]$ due to the rotation invariance of P.

The fifth equality holds by expanding Δ, linearity of expectation, and a similar argument to the previous equality to show $Pr(v \cdot f\hat{v} \cdot \hat{f} \leq 0|\tilde{f}) = Pr(v \cdot fv \cdot \hat{f} \leq 0|d * \tilde{f})$.

The sixth equality holds by the rotation invariance of P and the fact that k is odd.

For the final inequality, the right hand term is non-negative since $a(\tilde{f}) \geq a(d * \tilde{f})$ and Φ is non-increasing. The left hand term is also non-negative due to the rotation invariance assumption and the fact that $\forall i, w_i \cdot \hat{w}_i \geq 0$. □

Acknowledgements

Daniel McNamara was a visitor at Carnegie Mellon University during the period of this research, supported by a Fulbright Postgraduate Scholarship.

This work was supported in part by NSF grants CCF-1422910, CCF-1535967, IIS-1618714, and a Microsoft Research Faculty Fellowship.

We thank the anonymous reviewers for their useful comments.

References

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Bansal, Mohit, Gimpel, Kevin, and Livescu, Karen. Tailoring continuous word representations for dependency parsing. In *Association for Computational Linguistics*, pp. 809–815, 2014.

Baxter, Jonathan. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(3):149–198, 2000.

Ben-David, Shai, Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, and Vaughan, Jennifer Wortman. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Donahue, Jeff, Jia, Yangqing, Vinyals, Oriol, Hoffman, Judy, Zhang, Ning, Tzeng, Eric, and Darrell, Trevor. DeCAF: a deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pp. 647–655, 2014.

Evgeniou, Theodoros and Pontil, Massimiliano. Regularized multitask learning. In *International Conference on Knowledge Discovery and Data Mining*, pp. 109–117, 2004.

Galanti, Tomer, Wolf, Lior, and Hazan, Tamir. A theoretical framework for deep transfer learning. *Information and Inference*, 2016.

Ganin, Yaroslav, Ustinova, Evgeniya, Ajakan, Hana, Germain, Pascal, Larochelle, Hugo, Laviolette, François, Marchand, Mario, and Lempitsky, Victor. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Germain, Pascal, Habrard, Amaury, Laviolette, François, and Morvant, Emilie. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *International Conference on Machine Learning*, pp. 738–746, 2013.

Ghifary, Muhammad, Kleijn, W Bastiaan, and Zhang, Mengjie. Domain adaptive neural networks for object recognition. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 898–904. Springer, 2014.

Girshick, Ross, Donahue, Jeff, Darrell, Trevor, and Malik, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

Hoffman, Judy, Guadarrama, Sergio, Tzeng, Eric S, Hu, Ronghang, Donahue, Jeff, Girshick, Ross, Darrell, Trevor, and Saenko, Kate. LSDA: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pp. 3536–3544, 2014.

Long, Mingsheng, Cao, Yue, Wang, Jianmin, and Jordan, Michael I. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105, 2015.

Mansour, Yishay, Mohri, Mehryar, and Rostamizadeh, Afshin. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, 2009.

Maurer, Andreas, Pontil, Massimiliano, and Romera-Paredes, Bernardino. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations of Machine Learning*. MIT Press, 2012.

Pentina, Anastasia and Lampert, Christoph H. A PAC-Bayesian bound for Lifelong Learning. In *International Conference on Machine Learning*, pp. 991–999, 2014.

Qu, Lizhen, Ferraro, Gabriela, Zhou, Liyuan, Hou, Weiwei, Schneider, Nathan, and Baldwin, Timothy. Big data small data, in domain out-of domain, known word unknown word: the impact of word representation on sequence labelling tasks. In *Conference on Computational Natural Language Learning*, pp. 89–93, 2015.

Raina, Rajat, Ng, Andrew Y., and Koller, Daphne. Constructing informative priors using transfer learning. In *International Conference on Machine Learning*, pp. 713–720, 2006.

Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Socher, Richard, Ganjoo, Milind, Manning, Christopher D, and Ng, Andrew. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pp. 935–943, 2013.

Yosinski, Jason, Clune, Jeff, Bengio, Yoshua, and Lipson, Hod. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.