Stability and Hypothesis Transfer Learning

Ilja Kuzborskij ILJA.KUZBORSKIJ@IDIAP.CH
Idiap Research Institute, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Francesco Orabona FRANCESCO@ORABONA.COM
Toyota Technological Institute at Chicago, USA

Abstract

We consider the transfer learning scenario, where the learner does not have access to the source domain directly, but rather operates on the basis of *hypotheses* induced from it – the Hypothesis Transfer Learning (HTL) problem. Particularly, we conduct a theoretical analysis of HTL by considering the *algorithmic stability* of a class of HTL algorithms based on Regularized Least Squares with biased regularization. We show that the relatedness of source and target domains accelerates the convergence of the Leave-One-Out error to the generalization error, thus enabling the use of the Leave-One-Out error to find the optimal transfer parameters, even in the presence of a *small* training set. In case of unrelated domains we also suggest a theoretically principled way to prevent negative transfer, so that in the limit we recover the performance of the algorithm not using any knowledge from the source domain.

1. Introduction

The standard assumption in supervised machine learning algorithms is to have models trained and tested on samples drawn from the same probability distribution. However, this assumption is often violated in practical applications.

A more general setting is the one in which the marginal distributions over training and testing domains are different but related. This is the problem of Domain Adaptation (DA), where a successful scheme typically utilizes large unlabeled samples from both domains to adapt a source hypothesis to the target domain. Previous work has addressed in detail the theory of DA and proposed algorithms that critically depend on optimal weighting parameters given by the theoretical analysis (Ben-David et al., 2010a;b; Mansour et al., 2009b; Cortes et al., 2008). However, in practice, the learner needs access to sufficient unlabeled samples from both domains to estimate these parameters. Even if unlabeled data are abundant, the estimation of these parameters can be computationally prohibitive in some scenarios. A hypothetical example is a large number of domains involved or, for instance, when one acquires new domains incrementally. Here, keeping unlabeled data from all the domains and reestimating parameters is a necessity.

To overcome this practical limitation, a new framework has been analyzed by a number of works (Fei-Fei et al., 2006; Yang et al., 2007; Orabona et al., 2009; Mansour et al., 2009a; Tommasi et al., 2010; Kuzborskij et al., 2013). In this framework, that we will call Hypothesis Transfer Learning (HTL), unlike DA, only *source hypotheses* trained on a source domain are utilized. The attractive quality of HTL is the fact, that it assumes no explicit access to the source domain, nor any knowledge about the relatedness of the source and target distributions. Although, this setting has been explored empirically with success, a formal theory of HTL is mostly missing. Hence it is unclear how to recover optimal transfer parameters and what properties of the source hypothesis affect generalization.

In this paper, we take a step towards a theory of HTL. In particular, we analyze the generalization ability of a class of HTL algorithms stemming from Regularized Least Squares (RLS) with biased regularization. We assume access to a given number of source hypotheses and a *small* set of training samples from tar-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

get domain. Rather than relying on oracle inequalities for tuning the optimal parameters, we use the Leave-One-Out (LOO) risk. The LOO risk is known to have low bias compared to empirical risk or cross-validation (Elisseeff & Pontil, 2003), thus making it preferrable in a small sample regime.

In the following, we will show that the variance of the LOO estimator for the considered algorithms decreases with the increasing quality of the source hypothesis over the target domain. We do so by employing the notion of *hypothesis stability* (Bousquet & Elisseeff, 2002), and upper bounding the second-order moment of the difference between the expected risk and the LOO risk. In addition, we propose how a hypothetical algorithm could avoid negative transfer in the case of unrelated domains, while in worst case scenario recovering the generalization guarantees of RLS. As a side effect, we improve polynomial generalization bounds of Bousquet & Elisseeff (2002) for RLS. Finally, from the stability theory point of view, this work also tries to address an open question by Elisseeff & Pontil (2003): *"Is there a way to incorporate prior knowledge via stability?"*, thus exposing a connection between stability and the Hypothesis Transfer Learning.

The rest of the paper is organized as follows. We formally state the HTL problem in Section 2 and introduce analyzed algorithms in Section 4. The main result comes in Section 5, particularly in Theorem 2, with implications discussed in Section 5.1. The proof of the main result can be found in Section 5.2, while related work on DA and HTL is covered in Section 3. Finally we draw some conclusions and discuss future work in Section 6.

2. Definitions and Problem Setting

In the following, we denote with small and capital bold letters respectively column vectors and matrices, e.g. $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_d]^T \in \mathbb{R}^d$ and $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$. The expected value of a random variable distributed according to p is denoted by $\mathbb{E}_{z \sim p}[\cdot]$, and multiple random variables as $\mathbb{E}_{z \sim p, z' \sim p}[\cdot]$.

Denoting by $\mathcal{X}$ and $\mathcal{Y}$ respectively the input and output space of the learning problem, the training set S is defined as $\{(\boldsymbol{x}_i, y_i) : 1 \leq i \leq m\}$, drawn i.i.d. from $\mathcal{X} \times \mathcal{Y}$ according to the probability distribution μ. We also define a supervised learning algorithm as follows.

Definition 1 (Supervised learning algorithm)**.** *A supervised learning algorithm is a map*

$$A : (\mathcal{X} \times \mathcal{Y})^m \mapsto \mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$$

that maps the training set S onto a hypothesis $f_S \in \mathcal{F}$.

The LOO training set is defined as $S^{\setminus i} = \{(\boldsymbol{x}_j, y_j) : 1 \leq j \leq i-1 \text{ and } i+1 \leq j \leq m\}$ and hypothesis $f_{S^{\setminus i}}$ is produced by an algorithm A given training set $S^{\setminus i}$.

To measure the accuracy of a learning algorithm, we introduce the non-negative *loss* function $\ell(f, (\boldsymbol{x}, y))$, which measures the cost incurred predicting $f(\boldsymbol{x})$ instead of y. In the following we will focus on the square loss, $\ell(f, (\boldsymbol{x}, y)) = (f(\boldsymbol{x}) - y)^2$, for its appealing computational properties.

The *expected risk* of a hypothesis f_S, with respect to the probability distribution p, is then defined as

$$R_p(A, S) = R_p(f_S) := \mathbb{E}_{(\boldsymbol{x},y) \sim p}[\ell(f_S, (\boldsymbol{x}, y))],$$

while the *empirical risk* is

$$\hat{R}(A, S) := \frac{1}{m} \sum_{i=1}^{m} \ell(f_S, (\boldsymbol{x}_i, y_i)).$$

We also define the *LOO risk* as

$$\hat{R}^{\,\text{loo}}(A, S) := \frac{1}{m} \sum_{i=1}^{m} \ell(f_{S^{\setminus i}}, (\boldsymbol{x}_i, y_i)).$$

2.1. Hypothesis Transfer Learning Problem

In addition to the training set S for the *target* domain, drawn according to μ, we now introduce a *source* domain. Let $\mathcal{X}'$ and $\mathcal{Y}'$ be respectively the input and output space of a *source* domain. The *source* training set $S' = \{(\boldsymbol{x}'_i, y'_i) : 1 \leq i \leq m'\}$ is drawn i.i.d. from $\mathcal{X}' \times \mathcal{Y}'$ according to the probability distribution μ'. Denote by f' the hypothesis generated by any learning algorithm over the source training set S'. The aim of an HTL algorithm is to use the source hypothesis f' to improve the performance of a supervised learning algorithm over S. More formally, we define HTL algorithm as follows

Definition 2 (HTL algorithm)**.** *An HTL algorithm is a map*

$$A^{htl} : (\mathcal{X} \times \mathcal{Y})^m \times \mathcal{F}' \mapsto \mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}} \tag{1}$$

that maps S and the source hypothesis $f' \in \mathcal{F}'$ onto a target hypothesis $f_S^{htl} \in \mathcal{F}$.

The aim of the HTL algorithm is to satisfy the Improvement Condition (IC):

$$R_\mu(A^{htl}, (S, f')) \leq R_\mu(A^{htl}, (S, \mathbf{0})) \ . \tag{2}$$

We define a failure to satisfy IC as a *negative transfer*. Note that we do not assume any relationship between μ and μ'. We are only interested in the observable improvement of the generalization error on the target domain.

3. Related Work

We start by introducing related work from DA field, closely related to the HTL problem. Most DA algorithms can de described by the general map

$$A^{\mathrm{DA}} : (\mathcal{X} \times \mathcal{Y})^m \times \mathcal{X}^{m_u} \times (\mathcal{X}' \times \mathcal{Y}')^{m'} \mapsto \mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}, \quad (3)$$

where in addition to previous notation, m_u is the size of unlabeled set. Typically it is assumed that $m \ll m'$, $m \ll m_u$ and μ, μ' are different.

A theoretical analysis of (3) has been proposed by Ben-David et al. (2010a), considering case $m = 0$, and alternatively $m > 0$, but $m \ll m'$. The work proves a VC-bound on the expected risk of a target hypothesis, when the target hypothesis minimizes a convex combination of the empirical source and target risks. Hence the optimal DA algorithm critically depends on weight parameters that control the importance of source and target distributions. However, these optimal parameters depend on a divergence term between unlabeled training samples drawn from source and target marginal distributions, that appears additively in their bound. The nature of this divergence term was later explained by Ben-David et al. (2010b): the additive divergence term is inevitable unless an algorithm has access to labeled training data from the target domain. A similar setting was also examined by Mansour et al. (2009b), who instead derived Rademacher complexity bounds, but in practice proposed a very similar sample reweighting scheme. Once again, an additive divergence term appears in the bounds. Although related, problem (1) is not directly reducible to (3), since it depends on the properties of the learning algorithm that generates the source hypothesis. Apart from that, the source domain is inaccessible. These facts render the mentioned DA bounds inapplicable for analysis of the Hypothesis Transfer Learning.

As pointed out by Mansour et al. (2009a), the source training set (or multiple sets) might not be available due to the prohibitive size. With this motivation for HTL they attack a multiple source type of problem (1). They pose an assumption that the target distribution μ_{mix} is a mixture of n source distributions. With n source hypotheses, it was shown, that if $R_{\mu'_i}(f'_i) \leq \epsilon \ \forall i \in \{1, \ldots, n\}$, then $R_{\mu_{\text{mix}}}(\sum_{i=1}^n \beta_i f'_i(x)) \leq \epsilon + \delta, \ \delta \geq 0$. The result is insightful, since it relates the risk of target and source hypotheses. However, again, the optimal weights depend on unknown quantities that are estimated from unlabeled samples, partially defeating the original purpose of the algorithm.

HTL have been also considered from Bayesian perspective. Li & Bilmes (2007) have analyzed a Bayesian approch to solving (1) via PAC-Bayes bounds and arrived at an additive KL-divergence term. It was shown that for logistic regression, the divergence term is upper bounded by $\|f - f'\|$, leading to biased regularization in the learning algorithms. Indeed, Bayesian linear regression with f'-mean Gaussian prior over f leads to exact recovery of $\|f - f'\|$ in optimization problem (Bishop, 2006). Another example of the Bayesian HTL approach was proposed by Fei-Fei et al. (2006) for visual object detection task. Results of Li & Bilmes (2007) hint that generative methods like the one in Fei-Fei et al. (2006) could also be related to biased regularization.

A number of empirical attempts have tried to justify HTL. An SVM-like algorithm with regularizer $\|f - f'\|$ was proposed by Yang et al. (2007) for video concept detection task. Orabona et al. (2009) suggested a parametrized variant $\|f - \beta f'\|$ for Least-Squares SVM, then extended to multiple sources by Tommasi et al. (2010). Leveraging on this idea, a recent HTL multiclass formulation explored a class-incremental transfer setting (Kuzborskij et al., 2013). While some of these methods demonstrated impressive practical potential, their theoretical nature remains unclear.

We also briefly mention that variance bounds on the LOO risk were derived by Zhang (2003). In our work we are more interested in the generalization ability of the algorithm, that is estimated through the concept of algorithmic stability (Bousquet & Elisseeff, 2002) and the LOO risk.

4. Hypothesis Transfer Learning through Regularized Least Squares

Without loss of generality, in the following we will assume that $\mathcal{Y}, \mathcal{Y}' = [-B; B]$, where $B \in \mathbb{R}$ and $\|\boldsymbol{x}\| \leq 1, \ \boldsymbol{x} \in \mathcal{X} = \mathbb{R}^d$.

We will consider linear algorithms, extended to non-linear ones through the use of kernels. Hence $f(\boldsymbol{x})$ will be expressed as the inner product of a vector $\boldsymbol{w}$, learned from the training data, and the sample $\boldsymbol{x}$.

We assume, that only the target training set S and source hypothesis f' are given, so that the source training set S' is not required. The main objective of this analysis is to identify the effect of f' on the generalization properties of A^{htl}. For this reason, we would like to bound the expected risk of A^{htl} with terms depending on the characteristics of f'. In particular, we expect that a smaller risk $R_\mu(f')$ should improve the generalization of A^{htl}, compared to the case when $f' \equiv 0$.

As said above, we proceed by specializing A^{htl} to a particular class of algorithms, the RLS with biased regularization. This will allow us to arrive at a generalization bound where all the relevant quantities are computable in a closed form.

4.1. Biased Regularized Least Squares

The RLS algorithm consists in solving the following optimization problem

$$\min_{\boldsymbol{u}} \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{u}^\top \boldsymbol{x}_i - y_i)^2 + \lambda\|\boldsymbol{u}\|^2 \ . \qquad (4)$$

The interest of RLS lies in its strong theoretical guarantees and in the fact that solution can be expressed in a closed form (Rifkin et al., 2003). As a useful consequence, its LOO prediction function is expressed in closed form as well, allowing a very efficient model selection (Cawley & Talbot, 2007). It is also possible to arrive at (4) from a Bayesian perspective by putting a **0**-mean Gaussian prior over the parameters of a linear regression model (Bishop, 2006). Note that the same formulation can be used for both classification and regression problems (Rifkin et al., 2003).

Assuming that the source hypothesis $f'(\boldsymbol{x})$ is expressed as $\boldsymbol{x}^\top \boldsymbol{w}'$, and $\boldsymbol{w}'$ belongs to the same space of $\boldsymbol{w}$, Orabona et al. (2009) proposed the use of a biased regularization to solve hypothesis transfer learning problems efficiently. More formally they defined the following algorithm.

Algorithm 1. *The Hypothesis Transfer Learning Algorithm based on Regularized Least Squares produces a hypothesis* $f_S^{htl}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}_S$*, where*

$$\boldsymbol{w}_S = \operatorname*{argmin}_{\boldsymbol{u}} \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{u}^\top \boldsymbol{x}_i - y_i)^2 + \lambda\|\boldsymbol{u} - \boldsymbol{w}'\|^2. \quad (5)$$

Analogously, one can see the formulation of an Algorithm 1 as a Bayesian linear regression with $\boldsymbol{w}'$-mean Gaussian prior distribution. The solution of an Algorithm 1 can be expressed in closed form, in fact from the first order optimality condition we get

$$\begin{aligned}
\boldsymbol{w}_S = \ & \operatorname*{argmin}_{\boldsymbol{u}} \frac{1}{m}\|\boldsymbol{X}^\top \boldsymbol{u} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{u} - \boldsymbol{w}'\|^2 \\
& \Rightarrow \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{w}_S - \boldsymbol{y}) + m\lambda(\boldsymbol{w}_S - \boldsymbol{w}') = 0 \qquad (6) \\
& \Rightarrow \boldsymbol{X}(\boldsymbol{X}^\top \hat{\boldsymbol{w}}_S + \boldsymbol{X}^\top \boldsymbol{w}' - \boldsymbol{y}) + m\lambda\hat{\boldsymbol{w}}_S = 0 \\
& \Rightarrow (\boldsymbol{X}\boldsymbol{X}^\top + m\lambda\boldsymbol{I})\hat{\boldsymbol{w}}_S = \boldsymbol{X}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w}' \\
& \Rightarrow \hat{\boldsymbol{w}}_S = (\boldsymbol{X}\boldsymbol{X}^\top + m\lambda\boldsymbol{I})^{-1}\boldsymbol{X}(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{w}') \\
& \Rightarrow \hat{\boldsymbol{w}}_S = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + m\lambda\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{w}')
\end{aligned}$$

where in (6) we used $\hat{\boldsymbol{w}}_S := \boldsymbol{w}_S - \boldsymbol{w}'$ and in the last step we used the identity $(\boldsymbol{X}\boldsymbol{X}^\top + m\lambda\boldsymbol{I})^{-1}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + m\lambda\boldsymbol{I})^{-1}$ to express the solution in dual variables. So, the solution to the problem is given by $\boldsymbol{w}_S = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + m\lambda\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{w}') + \boldsymbol{w}'$, due the definition of $\hat{\boldsymbol{w}}_S$.

Using the fact that the LOO risk of Algorithm 1 can be written in closed form, Orabona et al. (2009) proposed to weight the source hypothesis $\boldsymbol{w}'$ by a scalar β, optimized in order to minimize the LOO risk.

In the following we show how to generalize this approach to the generic source hypotheses f' and how to obtain a generalization guarantee for it.

5. Analysis by Hypothesis Stability

We now propose a more general version of Algorithm 1.

Algorithm 2. *RLS transfer algorithm by altering training set as* $\{(\boldsymbol{x}_i, y_i - f'(\boldsymbol{x}_i)) : 1 \le i \le m\}$ *produces a hypothesis*

$$f_S^{htl'}(\boldsymbol{x}) = T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S) + f'(\boldsymbol{x}),$$

where

$$\hat{\boldsymbol{w}}_S := \operatorname*{argmin}_{\boldsymbol{u}} \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{u}^\top \boldsymbol{x}_i - y_i + f'(\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{u}\|^2,$$

and the truncation function $T_C(\hat{y})$ *is defined as* $T_C(\hat{y}) = \min(\max(\hat{y}, -C), C)$.

If $f'(x)$ is equal $\boldsymbol{x}^\top \boldsymbol{w}'$, where $\boldsymbol{w}'$ belongs to the same space of $\boldsymbol{w}_S$, and $C = \infty$, Algorithms 1 and 2 are completely equivalent, because they have exactly the same solution. However, Algorithm 2 is more general because it allows $\boldsymbol{w}'$ to belong to a different space. Hence it captures the notion of biased regularization, and generalizes it to any type of source hypothesis f'. This algorithm also captures and generalizes many of the ideas present in the previous works on HTL (Fei-Fei et al., 2006; Yang et al., 2007; Orabona et al., 2009; Mansour et al., 2009a; Tommasi et al., 2010). Still the use of a specific loss, the square loss, will allow us to have an efficient computation as well. Also, this formulation allows to truncate the prediction within the range $[-C; C]$, that helps to greatly improve the theoretical guarantees and the practical performance too. In fact it is easy to see that if $C \ge B + \|f'\|_\infty$, then $(T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S) + f'(\boldsymbol{x}) - y)^2 \le (\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S + f'(\boldsymbol{x}) - y)^2$.

Our goal is to upper bound the expected risk of Algorithm 2, keeping in mind the effect of f'. To this end, we propose to employ the stability framework of Bousquet & Elisseeff (2002). Our choice is motivated by the fact that bounds arising from the stability analysis are free from complexity measures. Hence, the generalization bound of interest will be composed mostly from computable quantities, thus making it more practical, e.g. for finding the optimal transfer parameters.

In particular, we can upper bound the moments of the random variable $R_\mu(A,S) - \hat{R}^{\,\mathrm{loo}}(A,S)$ with a quantity that captures the stability of the learning algorithm. The second order moment can then be used to obtain polynomial bounds, through the Chebyshev's inequality (Bousquet & Elisseeff, 2002).

There exist various definitions of stability (Bousquet & Elisseeff, 2002), but the one we will use is the hypothesis stability.

Definition 3 (Hypothesis Stability (Bousquet & Elisseeff, 2002))**.** *An algorithm A has a hypothesis stability γ with respect to the loss function ℓ if $\forall i \in \{1, \dots, m\}$ the following holds*

$$\mathbb{E}_{S,(\boldsymbol{x},y)}\left[|\ell(f_S,(\boldsymbol{x},y)) - \ell(f_{S^{\setminus i}},(\boldsymbol{x},y))|\right] \le \gamma \ .$$

We will use a slight variation of the polynomial bound of Bousquet & Elisseeff (2002). The reason is, that the Theorem 11 of Bousquet & Elisseeff (2002) has the term $\frac{M^2}{2}$, that is not affected by $R_\mu(f')$. Instead, we exchange $\frac{M^2}{2}$ for the term $\mathbb{E}_S[\ell(f_{S^{\setminus i}}, z_i)]$.

Theorem 1. *For a supervised learning algorithm A with hypothesis stability γ, and M such that $\ell(f_{S^{\setminus i}},(\boldsymbol{x},y)) \le M$, for any $i \in \{1,\dots,m\}$, we have*

$$\begin{aligned}&\mathbb{E}_S[(R_\mu(A,S) - \hat{R}^{\,\mathrm{loo}}(A,S))^2] \\ &\qquad \le \frac{M\,\mathbb{E}_S[\ell(f_{S^{\setminus i}},(\boldsymbol{x}_i,y_i))]}{m} + 3M\gamma \ .\end{aligned}$$

Proof (Sketch). We trace the occurrence of $\frac{M^2}{2m}$ to the proof of Lemma 9 (Bousquet & Elisseeff, 2002). At the beginning of the proof they suggest the following inequality

$$\begin{aligned}&\mathbb{E}_S[(R_\mu(f) - \hat{R}^{\,\mathrm{loo}}(f))^2] \\ &\quad \le \frac{1}{m}\mathbb{E}_S[\ell(f_{S^{\setminus i}},z_i)(M - \ell(f_{S^{\setminus i}},z_j))] \\ &\quad + \mathbb{E}_{S,z,z'}[\ell(f_{S^{\setminus i}},z)\ell(f_{S^{\setminus i}},z') - \ell(f_{S^{\setminus i}},z)\ell(f_{S^{\setminus i}},z_i)] \\ &\quad + \mathbb{E}_{S,z,z'}[\ell(f_{S^{\setminus i}},z_i)\ell(f_{S^{\setminus i}},z_j) - \ell(f_{S^{\setminus i}},z)\ell(f_{S^{\setminus i}},z_i)] \ .\end{aligned}$$

Here we are only interested in first term, since it is the origin of the term $\frac{M^2}{2m}$. Using the fact that $\ell(f_{S^{\setminus i}},z_j) \ge 0$, we have

$$\begin{aligned}&\frac{1}{m}\mathbb{E}_S[\ell(f_{S^{\setminus i}},z_i)(M - \ell(f_{S^{\setminus i}},z_j))] \\ &\qquad = \frac{1}{m}\mathbb{E}_S[\ell(f_{S^{\setminus i}},z_i)(M - \ell(f_{S^{\setminus i}},z_j))] \\ &\qquad \le \frac{M}{m}\mathbb{E}_S[\ell(f_{S^{\setminus i}},z_i)] \ . \qquad \square\end{aligned}$$

Note that our bound in the worst case loses only a constant multiplicative factor with respect to the one of Bousquet & Elisseeff (2002).

With this theorem we can prove the following result.

Theorem 2. *Set $\lambda \ge \frac{1}{m}$. If $C \ge B + \|f'\|_\infty$, then for Algorithm 2 we have*

$$\begin{aligned}&\mathbb{E}_S[(R_\mu(f_S^{htl'}) - \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}))^2] \\ &\qquad = \mathcal{O}\left(C^2 \frac{T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right) + R_\mu(f')}{m\lambda}\right) \ .\end{aligned}$$

If $C = \infty$, then for Algorithm 2 we have

$$\begin{aligned}&\mathbb{E}_S[(R_\mu(f_S^{htl'}) - \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}))^2] \\ &\qquad = \mathcal{O}\left(\frac{R_\mu(f')(\|f'\|_\infty + B)^2}{m\lambda^3}\right) \ .\end{aligned}$$

The proof can be found in Section 5.2, while in the next Section we discuss the implications of this theorem.

5.1. Implications

First consider the case of $f' \equiv 0$. This case corresponds to learning without any source hypothesis, without transfer learning. If we set $C = \infty$, we have that the second-order moment is bounded by $\mathcal{O}\left(\frac{B^4}{m\lambda^3}\right)$, which is exactly the bound that can be obtained using the results in Bousquet & Elisseeff (2002) for RLS. We see this by combining Theorem 11 with Lemma 23 and obtaining a polynomial generalization bound[1] $\mathcal{O}\left(\frac{B^2}{\lambda^{1.5}\sqrt{m}}\right)$. Considering the second moment, the bound matches ours.

However, if we know the range $[-B;B]$, we can set C accordingly and obtain that the second moment is bounded by $\mathcal{O}\left(\frac{B^2}{m\lambda}\right)$. Thanks to the truncation, the bound is greatly improved over the polynomial bound with square loss in Bousquet & Elisseeff (2002).

We now turn our attention to the case where $f' \not\equiv 0$. In this case, the key quantity is $R_\mu(f')$, an indirect measure of how the source and target domains are related. This term takes the role of the divergence between source and target distribution (Ben-David et al., 2010a;b; Mansour et al., 2009b), however, this is a more intuitive measure which is directly linked to the loss: how the source hypothesis is going to perform on the target domain, the new task? In addition, it is multiplicative to all bound terms, while mentioned divergence terms are additive, even if the bounds are generally incomparable. Based on its value, we have various

[1] The same bound also appears in (De Vito et al., 2005) (p17, footnote 2).

regimes of interest. If $\frac{R_\mu(f')}{\lambda} \to 0$, we have the surprising result that $\mathbb{E}_S[(R_\mu(f_S^{htl'}) - \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}))^2] \to 0$. This implies that the expected risk approaches the LOO risk, with probability 1. In other words, the transfer learning decreases the variance of the LOO in case when the source and target domains are related. This also implies that we can expect the tuning of any parameter of the algorithm (e.g. the type of kernel) through the minimization of the LOO risk, to have optimal performance, even with a small training set. This is the first theoretical explanation of why the algorithms of Orabona et al. (2009); Tommasi et al. (2010) showed reliable performance despite a small training set. Note that $R_\mu(f')$ has to be small with respect to λ. In other words, the better the source hypothesis on the target domain, the more stable an HTL algorithm must be according to Theorem 3. Looking at Algorithm 1, this makes sense, since a very stable algorithm will generate a hypothesis that does not deviate much from the source $\boldsymbol{w}'$.

So far we have outlined the benefits of $R_\mu(f')$, but it is reasonable to ask what happens when this quantity is high, that is when the two domains are unrelated. From the bound, we see that Algorithm 2 is also robust against a mispecified source hypothesis f'. In fact, due to truncation, the rate is exactly the same as obtained in the non-transfer case. If we supply the algorithm with a "bad" source hypothesis, in the limit it will have the performance of an algorithm that learns just using the training set. Again, this robustness is achieved also thanks to the truncation, which avoids excessive growth of the loss. In other words, Algorithm 2 is resistant to negative trasfer. We actually suspect that the truncation is necessary only for the proof, and in fact, Tommasi et al. (2010) already noticed this robust behaviour for Algorithm 1.

We now consider the case when the source hypothesis f' is a weighted combination of n source hypotheses f'_i, that is $f' = \sum_{i=1}^n \beta_i f'_i$. This weighting strategy is equivalent to the ones used in the works on DA, but with the important difference that now these weights can be efficiently estimated from the target training set. In particular, one interpretation of the Theorem 2 yields

$$\min_{\boldsymbol{\beta}} R_\mu(f_S^{htl'}) \le \min_{\boldsymbol{\beta}} \hat{R}^{\,\mathrm{loo}}(f_S^{htl'}) + \mathcal{O}\left(\frac{\|\boldsymbol{\beta}\|^2}{\sqrt{\lambda m}}\right) .$$

Hence the bound suggests an efficient and principled way to find $\boldsymbol{\beta} = [\beta_1, \dots, \beta_n]^\top$. In other words, it is enough to minimize the LOO risk with respect to $\boldsymbol{\beta}$, taking into account the regularization term, thus turning $\boldsymbol{\beta}$ into a parameter of an optimization problem. Note that Orabona et al. (2009); Tommasi et al. (2010) already realized the empirical need to constrain $\boldsymbol{\beta}$, but here we demonstrate a principled form of the regularization. Note that the $\mathcal{O}(\cdot)$ notation used in the bound above hides the confidence variable δ, which should be tuned. Yet, here we are mainly interested in the correct form of the objective function for finding the transfer parameters, as a way of using theory to guide practice. Moreover, regardless of the specific procedure used to estimate the optimal value of $\boldsymbol{\beta}$, as noted above we expect the algorithm to be robust to negative transfer, at least in the asymptotic limit.

5.2. Proof of Theorem 2

To prove the bound of Theorem 2 we need to upper bound the quantities M, γ, and $\mathbb{E}_S[\ell(f_{S^{\setminus i}}, z_i)]$ of Theorem 1. To do so, we proceed by stating and proving additional lemmas. Particularly, M and $\mathbb{E}_S[\ell(f_{S^{\setminus i}}, z_i)]$ are considered in Lemma 4, while γ is bounded in Theorem 3.

We first present two technical lemmas. The first one is needed to bound the effect of the truncation, the second one is a closed-form formula for calculating the change in truncated predictions of RLS when a new sample point is added. This result is related to the well-known closed-form formula for LOO risk for RLS, e.g. see Cawley & Talbot (2007).

Lemma 1. *Let $\alpha \ge 1$, and $C \ge |y|$, then*

$$\begin{aligned}(T_C(\Delta) - y)^2 &\le (T_C(y + \alpha(\Delta - y)) - y)^2 \\ &\le \alpha^2 (T_C(\Delta) - y)^2 \ .\end{aligned}$$

Proof. We only prove the upper bound, noting that the proof of the lower bound is similar. The proof follows from an analysis of all the possible cases. The lemma trivially holds when $|y + \alpha(\Delta - y)| \le C$. For $\Delta > C$, the bound holds because $y + \alpha(\Delta - y) > C$; the same reasoning applies for $\Delta < -C$. The last case is when $\frac{C-y}{\alpha} + y < \Delta < C$. We have $(T_C(y + \alpha(\Delta - y)) - y)^2 = (C - y)^2$. Note that $C \ge y$ implies that $\frac{C-y}{\alpha} + y > y$, so $\Delta > y$ and this implies the stated bound. The case is analogous $-C < \Delta < -\frac{C+y}{\alpha} + y$: we have that $T_C(y + \alpha(\Delta - y)) = -C$ and $-\frac{C+y}{\alpha} + y \le y$ because $C + y \ge 0$, hence $\Delta < y$. □

Lemma 2. *Let $\boldsymbol{w}_S$ be the hypothesis produced by the RLS algorithm given training set S. For any i-th sample $(\hat{\boldsymbol{x}}, \hat{y}) \in S$, we have that the hypothesis $\boldsymbol{w}_{S^{\setminus i}}$ produced by the same RLS algorithm on a training set $S^{\setminus i}$ satisfies*

$$\begin{aligned}(T_C(\hat{\boldsymbol{x}}^\top \boldsymbol{w}_S) - \hat{y})^2 &\le (T_C(\hat{\boldsymbol{x}}^\top \boldsymbol{w}_{S^{\setminus i}}) - \hat{y})^2 \\ &\le \left(1 + \frac{1}{m\lambda}\right)^2 (T_C(\hat{\boldsymbol{x}}^\top \boldsymbol{w}_S) - \hat{y})^2 \ .\end{aligned}$$

Proof. The $\boldsymbol{w}_{S^{\setminus i}}$ is given by

$$\boldsymbol{w}_{S^{\setminus i}} = \operatorname*{argmin}_{\boldsymbol{u}} \frac{1}{m}\|\boldsymbol{X}^\top \boldsymbol{u} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{u}\|^2,$$

where $\boldsymbol{X}$ is a matrix $d \times (m-1)$ and $\boldsymbol{y}$ an $m-1$ dimensional vector, respectively the matrix of the training samples and vector of the training labels without the sample i. Let $\boldsymbol{M} := \boldsymbol{X}^\top \boldsymbol{X} + m\lambda \boldsymbol{I}$, then

$$\hat{\boldsymbol{x}}^\top \boldsymbol{w}_{S^{\setminus i}} = \hat{\boldsymbol{x}}^\top \boldsymbol{X}\boldsymbol{M}^{-1}\boldsymbol{y}.$$

It is straightforward to see that $\hat{\boldsymbol{x}}^\top \boldsymbol{w}_S$ is equal to

$$\begin{bmatrix} \hat{\boldsymbol{x}}^\top \boldsymbol{X} & \|\hat{\boldsymbol{x}}\|^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{M} & \boldsymbol{X}^\top \hat{\boldsymbol{x}} \\ \hat{\boldsymbol{x}}^\top \boldsymbol{X} & \|\hat{\boldsymbol{x}}\|^2 + m\lambda \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y} \\ \hat{y} \end{bmatrix}. \tag{7}$$

Expanding the middle term and using the block-wise matrix inversion property (Petersen & Pedersen, 2008) we get

$$\begin{aligned} \begin{bmatrix} \boldsymbol{M} & \boldsymbol{X}^\top \hat{\boldsymbol{x}} \\ \hat{\boldsymbol{x}}^\top \boldsymbol{X} & \|\hat{\boldsymbol{x}}\|^2 + m\lambda \end{bmatrix}^{-1} &= \begin{bmatrix} \boldsymbol{M}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{bmatrix} \\ &+ \frac{1}{a}\begin{bmatrix} \boldsymbol{M}^{-1}\boldsymbol{X}^\top \hat{\boldsymbol{x}} \\ -1 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{x}}^\top \boldsymbol{X}\boldsymbol{M}^{-1} & -1 \end{bmatrix}, \end{aligned}$$

where $a := \|\hat{\boldsymbol{x}}\|^2 + m\lambda - \hat{\boldsymbol{x}}^\top \boldsymbol{X}\boldsymbol{M}^{-1}\boldsymbol{X}^\top \hat{\boldsymbol{x}}$. Plugging this result into (7) yields

$$\hat{\boldsymbol{x}}^\top \boldsymbol{w}_S = \hat{\boldsymbol{x}}^\top \boldsymbol{w}_{S^{\setminus i}} - \frac{a - m\lambda}{a}(\hat{\boldsymbol{x}}^\top \boldsymbol{w}_{S^{\setminus i}} - \hat{y}) \ .$$

So we have

$$\begin{aligned} &(T_C(\hat{\boldsymbol{x}}^\top \boldsymbol{w}_{S^{\setminus i}}) - \hat{y})^2 \\ &\quad = \left(T_C\left(\frac{a}{m\lambda}(\hat{\boldsymbol{x}}^\top \boldsymbol{w}_S - \hat{y}) + \hat{y}\right) - \hat{y}\right)^2 \ . \end{aligned} \tag{8}$$

Observing that $0 \le \hat{\boldsymbol{x}}^\top \boldsymbol{X}\boldsymbol{M}^{-1}\boldsymbol{X}^\top \hat{\boldsymbol{x}} \le \|\hat{\boldsymbol{x}}\|^2$, we have that $1 \le \frac{a}{m\lambda} \le 1 + \frac{\|\hat{\boldsymbol{x}}\|^2}{m\lambda}$, hence we use the upper bound in Lemma 1 to derive the stated upper bound. Analogously, the lower bound follows from (8) and the lower bound in Lemma 1. □

We proceed by bounding $\mathbb{E}_S[\ell(\boldsymbol{w}_{S^{\setminus i}}, (\boldsymbol{x}_i, y_i))]$, and M in Lemma 4.

Lemma 3. *The following bounds hold for the hypothesis $\hat{\boldsymbol{w}}_S$ produced by the Algorithm 2*

$$\mathbb{E}_S\|\hat{\boldsymbol{w}}_S\|^2 \le \frac{1}{\lambda}R_\mu(f') \ ,$$

and

$$\|\hat{\boldsymbol{w}}_S\|^2 \le \frac{1}{\lambda}(B + \|f'\|_\infty)^2 \ .$$

Proof. We define

$$Q(\boldsymbol{u}) := \frac{1}{m}\sum_{i=1}^m (\boldsymbol{x}_i^\top \boldsymbol{u} - y_i + f'(\boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{u}\|^2 \ .$$

Using the definition of $\hat{\boldsymbol{w}}_S$ in Algorithm 2, we have that

$$Q(\hat{\boldsymbol{w}}_S) \le Q(\boldsymbol{0}) = \hat{R}(f') \ . \tag{9}$$

Hence we get $\|\hat{\boldsymbol{w}}_S\|^2 \le \frac{\hat{R}(f')}{\lambda}$. Now

$$\mathbb{E}_S\|\hat{\boldsymbol{w}}_S\|^2 \le \frac{1}{\lambda}\mathbb{E}_S\hat{R}(f') = \frac{1}{\lambda}R_\mu(f') \ .$$

For the second upper bound, from (9) it also follows

$$\|\hat{\boldsymbol{w}}_S\|^2 \le \frac{1}{m\lambda}\sum_{i=1}^m (f'(\boldsymbol{x}_i) - y_i)^2 \le \frac{1}{\lambda}(B + \|f'\|_\infty)^2 \ . \ \square$$

Lemma 4. *Assume $(\boldsymbol{x}, y)$ drawn according to μ. For Algorithm 2 the following bounds hold $\forall i \in \{1, \dots, m\}$*

$$\begin{aligned} &\sup_{\boldsymbol{x},y} \ (T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_{S^{\setminus i}}) - y + f'(\boldsymbol{x}))^2 \\ &\le \left(1 + \frac{1}{m\lambda}\right)^2 \left(T_C\left(\frac{B + \|f'\|_\infty}{\sqrt{\lambda}}\right) + B + \|f'\|_\infty\right)^2, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}_S[(T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S) - y + f'(\boldsymbol{x}_i))^2] \\ &\le 2\left(T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right) + R_\mu(f')\right) \ . \end{aligned}$$

Proof. We use Lemma 2, and the Cauchy-Schwarz inequality to derive

$$\begin{aligned} &\sup_{\boldsymbol{x},y} \ (T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_{S^{\setminus i}}) - y + f'(\boldsymbol{x}))^2 \\ &\le \left(1 + \frac{1}{m\lambda}\right)^2 \sup_{\boldsymbol{x},y} \ (T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S) - y + f'(\boldsymbol{x}))^2 \\ &\le \left(1 + \frac{1}{m\lambda}\right)^2 \sup_{\boldsymbol{x},y} \ \left(|T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S)| + B + \|f'\|_\infty\right)^2 \ . \end{aligned}$$

The term $|T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S)|$ can be simultaneosly upper bounded using C and, using Cauchy-Schwarz inequality, by $\|\hat{\boldsymbol{w}}_S\|$. Hence using the second result of Lemma 3 we obtain the first result.

For the second upper bound, using the elementary inequality $(a+b)^2 \le 2(a^2 + b^2)$, in an analogous way we have

$$\begin{aligned} &\mathbb{E}_S[(T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S) - y + f'(\boldsymbol{x}))^2] \\ &\quad \le 2\mathbb{E}_S[(T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S))^2 + (f'(\boldsymbol{x}) - y)^2] \\ &\quad = 2\left(\mathbb{E}_S(T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S))^2 + R_\mu(f')\right) \ . \end{aligned}$$

Again, the first term in the left hand side of the last inequality can be simultaneosly upper bounded using C^2 and, using Cauchy-Schwarz inequality, by $\|\hat{\boldsymbol{w}}\|^2$. Hence the first result of Lemma 3 concludes the proof. □

Now, we are ready to upper-bound the hypothesis stability for Algorithm 2.

Theorem 3. *The hypothesis stability of Algorithm 2 is upper bounded as*

$$\gamma \leq \frac{4}{m\lambda}\left(2 + \frac{1}{m\lambda}\right)\left(T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right) + R_\mu(f')\right) \ .$$

Proof. We start by upper bounding the $|\cdot|$ term in the hypothesis stability from Definition 3. Using Lemma 2, with $\hat{y} = y - f'(\boldsymbol{x})$ and $\hat{\boldsymbol{x}} = \boldsymbol{x}$, we have

$$\begin{aligned}
&\left|(f_S^{htl'}(\boldsymbol{x}) - y)^2 - (f_{S^{\setminus i}}^{htl'}(\boldsymbol{x}) - y)^2\right| \\
&= (f_{S^{\setminus i}}^{htl'}(\boldsymbol{x}) - y)^2 - (f_S^{htl'}(\boldsymbol{x}) - y)^2 \\
&= (T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_{S^{\setminus i}}) - y + f'(\boldsymbol{x}))^2 - \\
&\qquad (T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S) - y + f'(\boldsymbol{x}))^2 \\
&\leq \left(\left(1 + \frac{1}{m\lambda}\right)^2 - 1\right)(T_C(\boldsymbol{x}^\top \hat{\boldsymbol{w}}_S) - y + f'(\boldsymbol{x}))^2 \ .
\end{aligned}$$

Taking now the expectation with respect to S and using the second result of Lemma 4 we have the stated bound. □

With the results above is now easy to prove Theorem 2.

Proof of Theorem 2. Using the upper bound in Lemma 2 and the second result in Lemma 4, we have

$$\begin{aligned}
&\mathbb{E}_S[\ell(f_{S^{\setminus i}}, (\boldsymbol{x}_i, y_i))] \\
&\quad \leq 2\left(1 + \frac{1}{m\lambda}\right)^2 \left(T_{C^2}\left(\frac{R_\mu(f')}{\lambda}\right) + R_\mu(f')\right) \ .
\end{aligned}$$

It is now enough to introduce the upper bounds on M (first result of Lemma 4), γ, and $\mathbb{E}_S[\ell(f_{S^{\setminus i}}, z_i)]$ into Theorem 1. □

6. Conclusions and Future Work

In this paper we have formally introduced the HTL problem and analyzed a class of RLS algorithms with biased regularization that can be used to solve this problem. Our main result is a generalization bound in terms of the Leave-One-Out (LOO) risk, obtained through the notion of hypothesis stability. We point out the key quantity $R_\mu(f')$ and expose its theoretical and practical advantages over analogues in the theory of DA. In particular, we showed that if source and target domains are related, hence $R_\mu(f')$ is small, the LOO risk converges faster to the expected risk and the HTL decreases the variance of the LOO. In the case of unrelated domains, we still match the theoretical guarantees of Regularized Least Squares trained solely on the target domain. As a side effect of our analysis, thanks to the truncation we have improved the polynomial generalization bounds of Bousquet & Elisseeff (2002) for RLS[2].

In future work, we will focus on the theoretical extension to a more general class of algorithms. Since it is possible to express the LOO predictions in a closed form for Algorithm 2, we intend to analyze its stability with respect to any positive loss function, decoupling the loss used in the algorithm from the one used in the analysis, in a similar way as in, e.g., Orabona et al. (2012). Accomplishing that suggests the possiblity for analysis of intriguing HTL scenarios through different loss functions, such as multiclass and hierarchical classification.

Another direction lies in improving the results by obtaining high probability bounds. We note that Bousquet & Elisseeff (2002) have also proved high probability bounds for another notion of stability, the uniform stability, but the assumption is too strong in our case, since it relates stability to the infinity norm of the losses which cannot be linked to the definition of the source risk $R_\mu(f')$. The rigid assumption on uniform stability was also noted by Kutin & Niyogi (2002), who proved exponential PAC-style bounds for weaker notions of stability. However we forecast two problems with applying their framework. First, the bound on the loss is additive in the generalization bound and cannot be linked to the source risk. In our case we managed to avoid this by modifying the proof of Bousquet & Elisseeff (2002) in Theorem 1 and an analogous and non-trivial change of Kutin & Niyogi (2002) results would be needed. Second, the alternative framework only bounds the discrepancy of expected and empirical risks. Instead we prefer to study the use the LOO risk, since it is empirically more robust with small training set.

Acknowledgments

I.K. is supported by the Swiss National Science Foundation Sinergia project Ninapro.

[2] The suboptimality of bounds in Bousquet & Elisseeff (2002) for RLS is also discussed by Zhang (2003).

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J.W. A theory of learning from different domains. *Machine Learning*, 79(1): 151–175, 2010a.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility Theorems for Domain Adaptation. *JMLR W&CP*, 9:129–136, 2010b.

Bishop, C.M. *Pattern Recognition and Machine Learning*. Springer-Verlang New York, 2006.

Bousquet, O. and Elisseeff, A. Stability and Generalization. *Journal of Machine Learning Research*, 2: 499–526, 2002.

Cawley, G. C. and Talbot, N. L. C. Preventing Over-Fitting during Model Selection via Bayesian Regularisation of the Hyper-Parameters. *Journal of Machine Learning Research*, 8:841–861, 2007.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. Sample Selection Bias Correction Theory. In *Algorithmic Learning Theory*, pp. 38–53. Springer, 2008.

De Vito, E., Caponnetto, A., and Rosasco, L. Model Selection for Regularized Least-Squares Algorithm in Learning Theory. *Found. Comput. Math.*, 5(1): 59–85, February 2005.

Elisseeff, A. and Pontil, M. Leave-one-out Error and Stability of Learning Algorithms with Applications. In *Advances in Learning Theory: Methods, Models and Applications*, pp. 111–125. VIOS Press, 2003.

Fei-Fei, L., Fergus, R., and Perona, P. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4): 594–611, 2006.

Kutin, S. and Niyogi, P. Almost-everywhere algorithmic stability and generalization error. In *Eighteenth conference on Uncertainty in artificial intelligence*, pp. 275–282, 2002.

Kuzborskij, I., Orabona, F., and Caputo, B. From N to N+1: Multiclass Transfer Incremental Learning. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. IEEE, 2013.

Li, X. and Bilmes, J. A bayesian divergence prior for classifier adaptation. In *Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain Adaptation with Multiple Sources. In *Advances in neural information processing systems*, volume 21, pp. 1041–1048, 2009a.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *Computational Learning Theory*, 2009b.

Orabona, F., Castellini, C., Caputo, B., Fiorilla, A.E., and Sandini, G. Model Adaptation with Least-Squares SVM for Adaptive Hand Prosthetics. In *Robotics and Automation, IEEE International Conference on*, pp. 2897–2903. IEEE, 2009.

Orabona, F., Cesa-Bianchi, N., and Gentile, C. Beyond logarithmic bounds in online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Petersen, K.B. and Pedersen, M.S. The matrix cookbook. *Technical University of Denmark*, 2008.

Rifkin, R., Yeo, G., and Poggio, T. Regularized Least-Squares Classification. In *Advances in Learning Theory: Methods, Models and Applications*, pp. 131–154. VIOS Press, 2003.

Tommasi, T., Orabona, F., and Caputo, B. Safety in Numbers: Learning Categories from Few Examples with Multi Model Knowledge Transfer. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pp. 3081–3088. IEEE, 2010.

Yang, J., Yan, R., and Hauptmann, A.G. Cross-Domain Video Concept Detection Using Adaptive SVMs. In *Proceedings of the 15th international conference on Multimedia*, pp. 188–197. ACM, 2007.

Zhang, T. Leave-one-out Bounds for Kernel Methods. *Neural Computation*, 15(6):1397–1437, 2003.