Multi-task Learning with Labeled and Unlabeled Tasks

Anastasia Pentina [1] **Christoph H. Lampert** [1]

Abstract

In multi-task learning, a learner is given a collection of prediction tasks and needs to solve all of them. In contrast to previous work, which required that annotated training data is available for all tasks, we consider a new setting, in which for some tasks, potentially most of them, only unlabeled training data is provided. Consequently, to solve all tasks, information must be transferred between tasks with labels and tasks without labels. Focusing on an instance-based transfer method we analyze two variants of this setting: when the set of labeled tasks is fixed, and when it can be *actively* selected by the learner. We state and prove a generalization bound that covers both scenarios and derive from it an algorithm for making the choice of labeled tasks (in the active case) and for transferring information between the tasks in a principled way. We also illustrate the effectiveness of the algorithm by experiments on synthetic and real data.

1. Introduction

In the *multi-task learning* setting (Caruana, 1997) a learner is given a collection of prediction tasks that all need to be solved. The hope is that the overall prediction quality can be improved by processing the tasks jointly and sharing information between them. Indeed, theoretical as well as experimental studies have shown that information transfer can reduce the amount of annotated examples per task needed to achieve good performance under various assumptions on how the learning tasks are related.

All existing multi-task learning approaches have in common, however, that they need at least some labeled training data for every task of interest. In this paper, we study a new and more challenging setting, in which for a subset of the tasks (typically the large majority) only unlabeled data is available. In practice, it is highly desirable to be able to handle this situation for problems with a very large number of tasks, such as sentiment analysis for market studies: for different products different attributes matter and, thus, each product should be have its own predictor and forms its own learning task. At the same time annotating data for each such task is prohibitive, especially when new products are constantly added to the market. Another example are prediction problems, for which the *fixed cost* of obtaining any labels for a task can be high, even when the *variable cost* per label are reasonable. This is a well-known issue when using crowd sourcing for data annotation: recruiting and training annotators first imposes a large overhead, and only afterwards many labels can be obtained within a short time and at a low cost.

A distinctive feature of the setting we study is that it requires two types of information transfer: between the labeled tasks and from labeled to unlabeled ones. While the first type is common in multi-task learning, none of the existing multi-task methods is able to handle the second type. In contrast, information transfer from labeled to unlabeled tasks is commonly studied in domain adaptation research, where, however, transfer of the first type is typically not considered. Thus, the setting of *multi-task learning with labeled and unlabeled tasks* can be seen as a blend of traditional multi-task learning and domain adaptation.

In this work we focus on a transfer method that learns a predictor for every task of interest by minimizing a task-specific convex combination of training errors on the labeled tasks (Ben-David et al., 2007; Mansour et al., 2009). We choose this method because it allows us to capture both types of information transfer – between the labeled tasks and from labeled to unlabeled ones – in a unified fashion. Clearly, the success of this approach depends on the choice of the weights in the convex combinations. Moreover, one can expect it also to depend on the subset of labeled tasks as well, because some subsets of tasks might be more informative and representative than the others. This suggests that it will be beneficial if the labeled subset is not arbitrary but if it can be chosen in a data-dependent way. We refer to this learning scenario, where initially every task is represented only by a set of unlabeled examples and the learner can choose for which tasks to request some labels, as *active task selection*.

[1]IST Austria. Correspondence to: Anastasia Pentina <apentina@ist.ac.at>.

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

Our main result is a generalization bound that quantifies both of the aforementioned effects: it relates the total multi-task error to quantities that depend on the subset of labeled tasks and on the task-specific weights used for information transfer. Using the computable quantities in the bound as an objective function and minimizing it numerically, we obtain a principled algorithm for selecting which tasks to have labeled (in the active task selection scenario) and for choosing task-specific weights and predictors for all tasks, labeled as well as unlabeled. We highlight the practical usefulness of the derived method by experiments on synthetic and real data.

The success of any information transfer approach, regardless whether it is applied in the multi-task or the domain adaptation scenario, depends on the relatedness between tasks of interest. Indeed, one cannot expect to benefit from information transfer between the labeled tasks or to be able to obtain solutions of reasonable quality for the unlabeled ones if the given tasks are completely unrelated. An advantage of the method we propose is that from the associated generalization bound we can read off explicitly under which conditions the algorithm can be expected to succeed. In particular, it suggests that the proposed method is likely to succeed if the given set of tasks satisfies the following assumption of *task smoothness*: if two tasks are similar in their marginal distributions, then their optimal prediction functions are also likely to be similar. A more formal definition will be given in Section 3. The task smoothness assumption resembles the classical *smoothness assumption* of semi-supervised learning (Chapelle et al., 2006). It can be expected to hold in many real-world settings with a large number of tasks, for example in the aforementioned case of sentiment analysis: if two products are described using similar words, these words would likely have similar connotation for both products. Note, also, that a similar assumption appears implicitly in (Blanchard et al., 2011).

1.1. Related Work

Most existing multi-task learning methods work in the fully supervised setting and aim at improving the overall prediction quality by sharing information between the tasks. For this they employ different types of transfer: *instance-transfer* methods re-use training samples from different tasks (Crammer & Mansour, 2012), *parameter-transfer* methods assume that the predictors for all tasks are similar to each other in some norm and exploit this fact through specific regularizers (Evgeniou & Pontil, 2004), *representation-transfer* approaches assume that the predictors for all tasks share a common (low-dimensional) representation that can be learned from the data (Argyriou et al., 2007; 2008). Follow-up works extended and generalized these concepts, e.g. by learning the relatedness of tasks (Saha et al., 2011; Kang et al., 2011) or sharing only between subgroups of tasks (Xue et al., 2007; Kumar & Daumé III, 2012; Barzilai & Crammer, 2015). However, all of the above methods require at least some labeled data for each task.

To our knowledge, the only existing multi-task method that can be applied in the considered setting where for some tasks only unlabeled data is available is (Khosla et al., 2012). Motivated by the problem of dataset bias, this method relies on the assumption that different tasks are minor modifications (i.e. biased versions) of the same, true prediction problem. Similarly to (Evgeniou & Pontil, 2004), it uses specific regularizers and trains predictors for all tasks jointly as small perturbations of a common predictor, which corresponds to the hypothetical unbiased task and can potentially be applied to unseen problems. Thus, applied in the considered setting, this method provides one predictor for all unlabeled tasks and treats the labeled ones as slight variations of them.

Information transfer from labeled to unlabeled tasks is the question typically studied in domain adaptation research. In fact, if the set of labeled tasks is fixed, any domain adaptation technique might be used to obtain solutions for unlabeled tasks, in particular those based on source reweighting (Shimodaira, 2000), representation learning (Pan et al., 2011; Glorot et al., 2011), or semi-supervised transfer (Xing et al., 2007). However, by design all domain adaptation methods aim at finding the best predictor on a *single* target task given a *fixed* set of source tasks. Therefore none of them can readily be applied in the active task selection setting, where the learner needs to *select* the labeled tasks that would lead to good performance across *all tasks*.

A second related setting is zero-shot learning (Larochelle et al., 2008; Lampert et al., 2013; Palatucci et al., 2009), where contextual, usually semantic, information is used to solve a learning task for which no training data is available. The situation we are interested in is more specific than this, though, as we assume that unlabeled data of the tasks is available, not context in an arbitrary form. As we will show, this allows us to derive formal performance guarantees that zero-shot learning methods typically lack.

The active task selection scenario is directly related to the question of identifying a representative set of source tasks in domain adaptation, a question that has previously been raised in the context of sentiment analysis (Blitzer et al., 2007). It also shares some features with *active learning*, where the learner is given a set of unlabeled samples and can choose a subset to obtain labels for. A fundamental difference is, however, that in active learning the learner needs to find a single prediction function for all labeled and unlabeled data while in the multi-task setting each task, including unlabeled ones, potentially requires its own predictor.

In the multi-task or zero-shot setting, active learning has so far not found widespread use. Exemplary works in this direction are (Reichart et al., 2008; Saha et al., 2010; Gavves et al., 2015), which, however, use active learning on the level of training examples, not tasks. The idea of choosing tasks was used in active curriculum selection (Ruvolo & Eaton, 2013; Pentina et al., 2015), where the learner can influence the order in which tasks are processed. However these methods nevertheless require annotated examples for all tasks of interest.

2. MTL with Labeled and Unlabeled Tasks

In the multi-task setting the learner observes a collection of prediction tasks and its goal is to learn all of them. Formally, we assume that there is a set of T tasks $\{\langle D_1, f_1\rangle, \dots, \langle D_T, f_T\rangle\}$, where each task t is defined by a marginal distribution D_t over the input space $\mathcal{X}$ and a deterministic labeling function $f_t : \mathcal{X} \to \mathcal{Y}$. The goal of the learner is to find T predictors $h_1, \dots, h_T$ in a hypothesis set $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ that would minimize the average expected risk:

$$\mathrm{er}(h_1, \dots, h_T) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{er}_t(h_t), \qquad (1)$$
$$\text{where} \quad \mathrm{er}_t(h_t) = \mathop{\mathbb{E}}_{x\sim D_t} \ell(h_t(x), f_t(x)).$$

In this work we concentrate on the case of binary classification tasks, $\mathcal{Y} = \{-1, 1\}$, and 0/1-loss, $\ell(y_1, y_2) = 0$ if $y_1 = y_2$, and $\ell(y_1, y_2) = 1$ otherwise.

In the fully-supervised setting the learner is given a training set of annotated examples for every task of interest. In contrast, we consider the scenario where every task t is represented by a set $S_t = \{x_1^t, \dots, x_n^t\}$ of n unlabeled examples sampled i.i.d. according to the marginal distribution D_t. For a subset of k tasks $\{i_1, \dots, i_k\}$, which are either predefined or, in the active scenario, can be selected based on the unlabeled data, the learner is given labels for a random subset $\overline{S}_{i_j} \subset S_{i_j}$ of m points.

To obtain a predictor for any task, labeled or unlabeled, we consider a method that minimizes a convex combination of training errors of the labeled tasks. This choice allows us to capture, in a unified fashion, both types of information transfer – between the labeled tasks and from labeled to unlabeled ones. Formally, for a set of tasks $I = \{i_1, \dots, i_k\} \subset \{1, \dots, T\}$ we define:

$$\Lambda^I = \left\{ \alpha \in [0,1]^T : \sum_{i=1}^{T} \alpha_i = 1;\ \mathrm{supp}\,\alpha \subseteq I \right\} \qquad (2)$$

for $\mathrm{supp}\,\alpha = \{i \in \{1, \dots, T\} : \alpha_i \neq 0\}$. Given a weight vector $\alpha \in \Lambda^I$, the α-weighted empirical error of a hypothesis $h \in \mathcal{H}$ is defined as follows:

$$\widehat{\mathrm{er}}_\alpha(h) = \sum_{i\in I} \alpha_i \widehat{\mathrm{er}}_i(h), \qquad (3)$$
$$\text{where} \quad \widehat{\mathrm{er}}_i(h) = \frac{1}{m} \sum_{(x,y)\in\overline{S}_i} \ell(h(x), y). \qquad (4)$$

In order to obtain a solution for any task t the learner minimizes $\widehat{\mathrm{er}}_{\alpha^t}(h)$ for some $\alpha^t \in \Lambda^I$, where I is the set of labeled tasks, potentially in combination with some regularization.

The success of this approach depends on the subset I of tasks that are labeled and on the weights $\alpha^1, \dots, \alpha^T$. The following theorem quantifies both of these effects and will later be used to chose $\alpha^1, \dots, \alpha^T$ and potentially I in a principled way.

Theorem 1. *Let d be the VC dimension of the hypothesis set $\mathcal{H}$, k be the number of labeled tasks, $S_1, \dots, S_T$ be T sets of size n each, where $S_i \overset{i.i.d.}{\sim} D_i$, and $\overline{S}_1, \dots, \overline{S}_T$ be their random subsets of size m each, for which labels would be provided if the corresponding task is selected as labeled. Then for any $\delta > 0$ with probability at least $1-\delta$ over $S_1, \dots, S_T$ and $\overline{S}_1, \dots, \overline{S}_T$ uniformly for all choices of labeled tasks $I = \{i_1, \dots, i_k\}$ and weights $\alpha^1, \dots, \alpha^T \in \Lambda^I$, provided that they are fully determined by the unlabeled data only, and for all possible choices of $h_1, \dots, h_T \in \mathcal{H}$ the following inequality holds:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{er}_t(h_t) \le \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathrm{er}}_{\alpha^t}(h_t) + \frac{1}{T}\sum_{t=1}^{T}\sum_{i\in I}\alpha_i^t \,\mathrm{disc}(S_t, S_i)$$
$$+ \frac{A}{T}\|\alpha\|_{2,1} + \frac{B}{T}\|\alpha\|_{1,2} + C + D + \frac{1}{T}\sum_{t=1}^{T}\sum_{i\in I}\alpha_i^t\lambda_{ti}, \qquad (5)$$

where

$$\mathrm{disc}(S_t, S_i) = \max_{h,h'\in\mathcal{H}} |\widehat{\mathrm{er}}_{S_t}(h,h') - \widehat{\mathrm{er}}_{S_i}(h,h')|$$

with $\widehat{\mathrm{er}}_{S_i}(h,h') = \frac{1}{n}\sum_{j=1}^{n}\ell(h(x_j^i), h'(x_j^i))$ is the empirical discrepancy between unlabeled samples S_t and S_i, and

$$\lambda_{ij} = \min_{h\in\mathcal{H}}(\mathrm{er}_i(h) + \mathrm{er}_j(h))$$
$$\|\alpha\|_{2,1} = \sum_{t=1}^{T}\sqrt{\sum_{i\in I}(\alpha_i^t)^2}, \quad \|\alpha\|_{1,2} = \sqrt{\sum_{i\in I}\left(\sum_{t=1}^{T}\alpha_i^t\right)^2},$$
$$A = \sqrt{\frac{2d\log(ekm/d)}{m}}, \quad B = \sqrt{\frac{\log(4/\delta)}{2m}}$$
$$C = \sqrt{\frac{8(\log T + d\log(enT/d))}{n}} + \sqrt{\frac{2}{n}\log\frac{4}{\delta}},$$
$$D = 2\sqrt{\frac{2d\log(2n) + 2\log(T) + \log(4/\delta)}{n}}.$$

Proof Sketch (the full proof can be found in the supplemental material). By Theorem 2 in (Ben-David et al., 2010), for any two tasks t and i the following inequality holds for every $h \in \mathcal{H}$:

$$\mathrm{er}_t(h) \leq \mathrm{er}_i(h) + \mathrm{disc}(D_t, D_i) + \lambda_{ti}. \tag{6}$$

Thus, we obtain the following bound on the average expected error over all tasks in terms of the error on the labeled tasks:

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{er}_t(h_t) \leq \frac{1}{T}\sum_{t=1}^{T}\mathrm{er}_{\alpha^t}(h_t) \tag{7}$$

$$+\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in I}\alpha_i^t\,\mathrm{disc}(D_t, D_i) + \frac{1}{T}\sum_{t=1}^{T}\sum_{i\in I}\alpha_i^t\lambda_{ti},$$

$$\text{with}\quad \mathrm{er}_{\alpha^t}(h_t) = \sum_{i\in I}\alpha_i^t \mathop{\mathbb{E}}_{x\sim D_i}\ell(h_t(x), f_i(x)),\quad \text{and} \tag{8}$$

$$\mathrm{er}_{D_i}(h, h') = \mathop{\mathbb{E}}_{x\sim D_i}\ell(h(x), h'(x)),\quad \text{and} \tag{9}$$

$$\mathrm{disc}(D_t, D_i) = \max_{h,h'\in\mathcal{H}}\left|\mathrm{er}_{D_t}(h, h') - \mathrm{er}_{D_i}(h, h')\right| \tag{10}$$

is the discrepancy between two distributions (Kifer et al., 2004; Mansour et al., 2009; Ben-David et al., 2010). In order to prove the statement of the theorem we need to relate the α-weighted expected errors and discrepancies between the marginal distributions in (7) to their empirical estimates.

The proof consists of three steps. First, we show that, conditioned on the unlabeled data, $\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathrm{er}}_{\alpha^t}$ can be upper bounded in terms of $\frac{1}{T}\sum_{t=1}^{T}\widehat{\tilde{\mathrm{er}}}_{\alpha^t}$, where:

$$\tilde{\mathrm{er}}_\alpha(h) = \sum_{i\in I}\alpha_i\tilde{\mathrm{er}}_i(h) = \sum_{i\in I}\frac{\alpha_i}{n}\sum_{j=1}^{n}\ell(h(x_j^i), f_i(x_j^i)).$$

This quantity can be interpreted as a training error if the learner would receive the labels for all the samples for the chosen tasks I. Note that in case of $m = n$ this step is not needed and we can avoid the corresponding complexity terms. In the second step we relate the average α-weighted expected errors to $\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathrm{er}}_{\alpha^t}$. In the third step we conclude the proof by bounding the pairwise discrepancies in terms of their empirical estimates.

Step 1. Fix the unlabeled sets $S_1, \ldots, S_T$. They fully determine the choice of labeled tasks I and the weights $\alpha^1, \ldots, \alpha^T$. Therefore, conditioned on the unlabeled data, these quantities can be considered constant and the bound has to hold uniformly only with respect to $h_1, \ldots, h_T$.

In order to simplify the notation we assume that $I = \{1, \ldots, k\}$ and define:

$$\Phi(\overline{S}_1, \ldots, \overline{S}_k) = \sup_{h_1,\ldots,h_T}\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathrm{er}}_{\alpha^t}(h_t) - \widehat{\tilde{\mathrm{er}}}_{\alpha^t}(h_t). \tag{11}$$

Note that one could analyze this quantity using standard techniques from Rademacher analysis, if the labeled examples were sampled from the unlabeled sets i.i.d., i.e. with replacement. However, since we assume that for every i $\overline{S}_i$ is a subset of S_i, i.e. the labeled examples are sampled randomly without replacement, there are dependencies between the labeled examples. Therefore we utilize techniques from the literature on transductive learning (El-Yaniv & Pechyony, 2007) instead. We first apply Doob's construction to Φ in order to obtain a martingale sequence and then use McDiarmid's inequality for martingales (McDiarmid, 1989). As a result we obtain that with probability at least $1 - \delta/4$ over sampling labeled examples:

$$\Phi \leq \mathop{\mathbb{E}}_{\overline{S}_1,\ldots,\overline{S}_k}\Phi + \frac{1}{T}\sqrt{\sum_{i=1}^{k}\left(\sum_{t=1}^{T}\alpha_i^t\right)^2}\sqrt{\frac{\log(4/\delta)}{2m}}. \tag{12}$$

Now we need to upper bound $\mathbb{E}\,\Phi$. Using results from (Tolstikhin et al., 2014) and (Hoeffding, 1963) we observe that:

$$\mathop{\mathbb{E}}_{\overline{S}_1,\ldots,\overline{S}_k}\Phi(\overline{S}_1, \ldots, \overline{S}_k) \leq \mathop{\mathbb{E}}_{\tilde{S}_1,\ldots,\tilde{S}_k}\Phi(\tilde{S}_1, \ldots, \tilde{S}_k), \tag{13}$$

where $\tilde{S}_i$ is a set of m points sampled from S_i i.i.d. with replacement (in contrast to sampling without replacement corresponding to $\overline{S}_i$). This means that we can upper bound the expectation of Φ over samples with dependencies by the expectation over independent samples. By doing so, applying the symmetrization trick, and introducing Rademacher random variables, we obtain that:

$$\mathop{\mathbb{E}}_{\overline{S}_1,\ldots,\overline{S}_k}\Phi \leq \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{i=1}^{k}(\alpha_i^t)^2}\cdot\sqrt{\frac{2d\log(ekm/d)}{m}}. \tag{14}$$

A combination of (12) and (14) shows that (conditioned on the unlabeled data) with probability at least $1 - \delta/4$ over sampling labeled examples uniformly for all choices of $h_1, \ldots, h_T$ the following holds:

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathrm{er}}_{\alpha^t}(h_t) \leq \frac{1}{T}\sum_{t=1}^{T}\widehat{\tilde{\mathrm{er}}}_{\alpha^t}(h_t) + 1\frac{A}{T}\|\alpha\|_{2,1} + \frac{B}{T}\|\alpha\|_{1,2}. \tag{15}$$

Step 2. Now we relate $\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathrm{er}}_{\alpha^t}$ to $\frac{1}{T}\sum_{t=1}^{T}\mathrm{er}_{\alpha^t}$.

The choice of the tasks to label, I, the corresponding weights, α, and the predictors, h, all depend on the unlabeled data. Therefore, we aim for a bound that is uniform in all three parameters. We define:

$$\Psi(S_1, \ldots, S_T) =$$

$$\sup_{I}\sup_{\alpha^1,\ldots,\alpha^T\in\Lambda^I}\sup_{h_1,\ldots,h_T}\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{T}\alpha_i^t(\mathrm{er}_i(h_t) - \tilde{\mathrm{er}}_i(h_t)).$$

The main instrument that we use here is a refined version of McDiarmid's inequality, which is due to (Maurer, 2006). It allows us to use the standard Rademacher analysis, while taking into account the internal structure of the weights $\alpha^1, \ldots, \alpha^T$. As a result we obtain that with probability at least $1 - \delta/4$ simultaneously for all choices of tasks to be labeled, I, weights $\alpha^1, \ldots, \alpha^T \in \Lambda^I$ and hypotheses $h_1, \ldots, h_T$:

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{er}_{\alpha^t}(h_t) \leq \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathrm{er}}_{\alpha^t}(h_t) + C. \tag{16}$$

Step 3. To conclude the proof we bound the pairwise discrepancies in terms of their finite sample estimates. According to Lemma 1 in (Ben-David et al., 2010) for any pair of tasks i, j and any $\delta > 0$ with probability at least $1 - \delta$:

$$\mathrm{disc}(D_i, D_j) \leq \mathrm{disc}(S_i, S_j) + 2\sqrt{\frac{2d\log(2n) + \log(2/\delta)}{n}}.$$

We apply this result to every pair of tasks and combine the results using the uniform bound argument. This yields the remaining two terms on the right hand side: the weighted average of the sample-based discrepancies and the constant D. By combining the result with (15) and (16) we obtain the statement of the theorem. □

3. Explanation and Interpretation

The left-hand side of inequality (5) is the average expected error over all T tasks, the quantity of interest that the learner would like to minimize but cannot directly compute. It is upper-bounded by the sum of two complexity terms and five task-dependent terms: weighted training errors on the labeled tasks, weighted averages of the distances to the labeled tasks in terms of the empirical discrepancies, two mixed norms of the weights α and a weighted average of λ-s.

The complexity terms C and D behave as $O(\sqrt{d\log(nT)/n})$ and converge to zero when the number of unlabeled examples per task, n, tends to infinity. In contrast, $\frac{A}{T}\|\alpha\|_{2,1} + \frac{B}{T}\|\alpha\|_{1,2}$ in the worst case of $\|\alpha\|_{2,1} = \|\alpha\|_{1,2} = T$ behaves as $O(\sqrt{d\log(km)/m})$ and converges to zero when the number of labeled examples per labeled task, m, tends to infinity. In order for these terms to be balanced, i.e. for the uncertainty coming from the estimation of discrepancy to not dominate the uncertainty from the estimation of the α-weighted risks, the number of unlabeled examples per task n should be significantly (for $k \ll T$) larger than m. However, this is not a strong limitation under the common assumption that obtaining enough unlabeled examples is significantly cheaper than annotated ones.

The remaining terms on the right-hand side of (5) depend on the set of labeled tasks I, the tasks-specific weights α-s and hypotheses h-s. Thus, by minimizing them with respect to these quantities one can expect to obtain values for them that are beneficial for solving all tasks of interest based on the given data. For the theorem to hold, the set of labeled tasks and the weights may not depend on the labels. The part of the bound that can be estimated based on the unlabeled data only, and therefore to select I (in the active scenario) and $\alpha^1, \ldots, \alpha^T$ is:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i\in I} \alpha_i^t \,\mathrm{disc}(S_t, S_i) + \frac{A}{T}\|\alpha\|_{2,1} + \frac{B}{T}\|\alpha\|_{1,2}. \tag{17}$$

The first term in (17) is the average weighted distance from every task to the labeled ones, as measured by the discrepancy between the corresponding unlabeled training samples. This term suggests that for every task t the largest weight, i.e. the highest impact in terms of information transfer, should be put on a labeled task i that has a similar marginal distribution. Note that the employed "similarity", which is captured by the discrepancy, directly depends on the considered hypothesis class and loss function and, thus, is tailored to a particular setting of interest. At the same time, the mixed-norm terms $\|\alpha\|_{1,2}$ and $\|\alpha\|_{2,1}$ prevent the learner from putting all weight on the single closest labeled task and can be seen as some form of regularization. In particular, they encourage information transfer also between the labeled tasks, since minimizing just the first term in (17) for every labeled tasks $i \in I$ would result in all weight to be put on task i itself and nothing on other tasks, because by definition $\mathrm{disc}(S_i, S_i) = 0$.

The first mixed-norm term, $\|\alpha\|_{2,1}$ influences every α^t independently and encourages the learner to use data from multiple labeled tasks for adaptation. Thus, it captures the intuition that sharing from multiple labeled tasks should improve the performance. In contrast, $\|\alpha\|_{1,2}$ connects the weights for all tasks. This term suggests to label tasks that all would be equally useful, thus preventing spending resources on tasks that would be informative for only a few of the remaining ones. Also, it prevents the learner from having super-influential labeled tasks that share with too many others. Such cases would be very unstable in the worst case scenario: mistakes on such tasks would propagate and have a major effect on the overall performance.

The effect of the mixed-norm terms can also be seen through the lens of the convergence rates. Indeed, as already mentioned above, in the case of every α^t having only one non-zero component, $\|\alpha\|_{2,1}$ and $\|\alpha\|_{1,2}$ are equal to T and thus the convergence rate[1] is $\tilde{O}(\sqrt{1/m})$. However, in the opposite extreme, if every α^t weights all the labeled tasks equally, i.e. $\alpha_i^t = 1/k$ for all $t \in \{1, \ldots, T\}$ and

[1] $\tilde{O}(\cdot)$ is an analog of $O(\cdot)$ that hides logarithmic factors

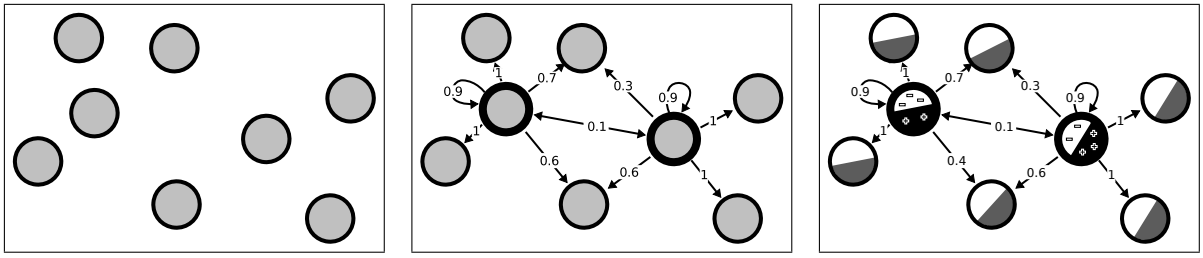


Figure 1. Schematic illustration of *active task selection*. Left: eight unlabeled tasks need to be solved. Center: the subset of tasks to be labeled and between-task weights are determined by minimizing (17). Right: annotated data for labeled tasks is obtained, and prediction functions (black vs. white) for each task are learned using the learned weighted combinations. Sharing can occur between labeled tasks.

$i \in I$, then $\|\alpha\|_{2,1} = \|\alpha\|_{1,2} = \frac{T}{\sqrt{k}}$ and the convergence rate improves to $\tilde{O}(\sqrt{1/km})$, which is the best one can expect from having a total of km labeled examples.

The only term on the right-hand side of (5) that depends on the hypotheses $h_1, \ldots, h_T$ and can be used to make a favorable choice is the weighted training error on the labeled tasks. Thus, the generalization bound of Theorem 1 suggest the following algorithm (Figure 1):

Algorithm 1.
1. estimate pairwise discrepancies between the tasks based on the unlabeled data
2. choose the tasks I to be labeled (in the active case) and the weights $\alpha^1, \ldots, \alpha^T$ by minimizing (17)
3. receive labels for the labeled tasks I
4. for every task t train a classifier by minimizing (3) using the obtained weights α^t.

Note, that this procedure is justified by Theorem 1: all choices are done in agreement with the conditions of the theorem and, because the inequality (5) holds uniformly for all eligible choices of labeled tasks, weights and predictors, the guarantees also hold for the resulting solution.

Algorithm 1 is guaranteed to perform well, if the solution it finds leads to a low value of the right-hand side of (5). By construction, it minimizes all data-dependent terms in (5), except for one quantity that cannot be estimated from the available data:

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{i \in I} \alpha_i^t \lambda_{ti}. \tag{18}$$

While discrepancy captures the similarity between marginal distributions, the λ-terms reflect the similarity between labeling functions: for every pair of task, t, and labeled task, $i \in I$, the corresponding value λ_{ti} is small if there exists a hypothesis that performs well on both tasks. Thus, Algorithm 1 can be expected to perform well, if for any two given tasks t and i that are close to each other in terms of discrepancy (and thus in the minimization of (17) the corresponding α_i^t is large), there exists a hypothesis that performs well on both of them (i.e. the corresponding λ_{ti} is small). We refer to this property of the set of learning tasks as *task smoothness*.

Training predictors for every task of interest using data from all labeled tasks improves the statistical guarantees of the learner. However, it results in empirical risk minimization on up to km samples for T different weighted combinations. Since we are most interested in the situation when T is large, one might be interested in way to reduce the amount of necessary computation. One way to do so is to drop the mixed-norm terms from the objective function (17), in which case it reduces to

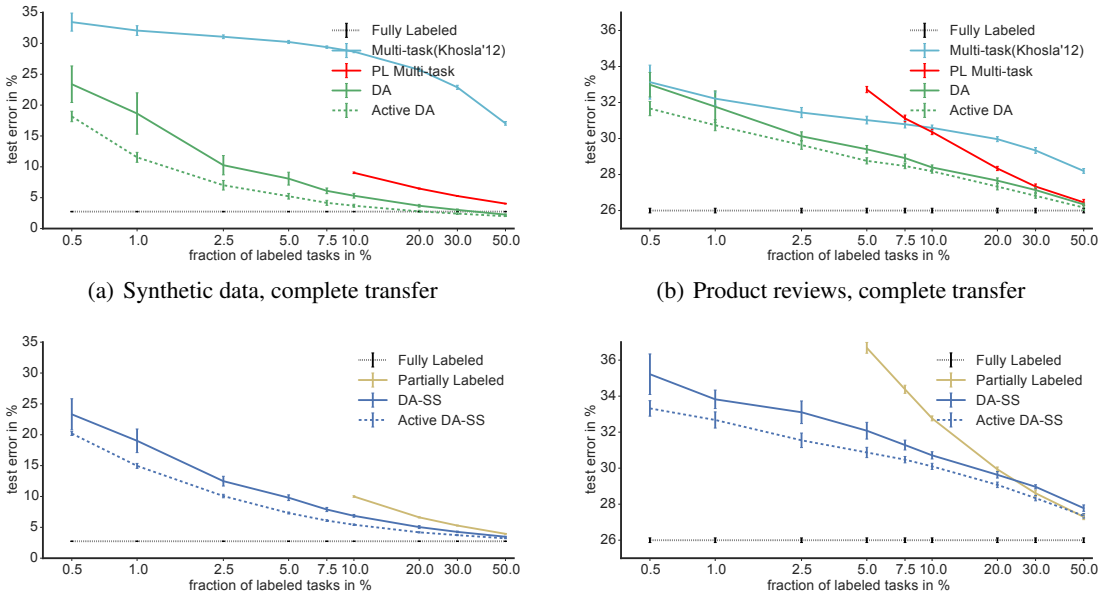
$$\frac{1}{T} \sum_{t=1}^{T} \sum_{i \in I} \alpha_i^t \operatorname{disc}(S_t, S_i). \tag{19}$$

This expression is linear in α and thus minimizing it for a fixed set I will lead to assigning each task to a single labeled task that is closest to it in terms of empirical discrepancy. Each labeled task will be assigned to itself. Consequently, the learner must train only k predictors, one for each labeled task, using only its m samples. The expression (19) can be seen as the k-medoids clustering objective with tasks corresponding to points in the space with (semi-)metric defined by empirical discrepancy and labeled tasks correspond to the centers of the clusters. Thus, this method reduces to k-medoids clustering, resembling the suggestion of Blitzer et al. (2007). Note that, nevertheless, the conditions of Theorem 1 are fulfilled, and thus its guarantees will hold for the obtained solution. The guarantees will be more pessimistic, however, than those from Algorithm 1, as the minimization ignores parts of the bound (5) and will not use the potentially beneficial transfer between labeled tasks.

4. Experiments

To illustrate that the proposed algorithm can also be practically useful, we performed experiments on synthetic and real data. In both cases we choose $\mathcal{H}$ to be the set of all linear predictors with a bias term on $\mathcal{X} = \mathbb{R}^d$.

Synthetic data. We generate $T = 1000$ binary classifica-



(a) Synthetic data, complete transfer

(b) Product reviews, complete transfer

(c) Synthetic data, single-source transfer

(d) Product reviews, single-source transfer

Figure 2. Experimental results on synthetic and real data: average test error and standard deviation over 10 repeats.

tion tasks in $\mathbb{R}^2$. For each task t its marginal distribution D_t is a unit-variance Gaussian with mean μ_t chosen uniformly at random from the set $[-5, 5] \times [-5, 5]$. The label $+1$ is assigned to all points that have angle between 0 and π with μ_t (computed counter-clockwise), the other points are labeled -1. We use $n = 1000$ unlabeled and $m = 100$ labeled examples per task.

Real Data. We curate a *Multitask dataset of product reviews*[2] from the corpus of *Amazon product data*[3] (McAuley et al., 2015a;b). We select the products for which there are at least 300 positive reviews (with scores 4 or 5) and at least 300 negative reviews (with scores 1 or 2). Each of the resulting 957 products we treat as a binary classification task of predicting whether a review is positive or negative. For every review we extract features by first pre-processing (removing all non-alphabetical characters, transforming the rest into lower case and removing stop words) and then applying the sentence embedding procedure of (Arora et al., 2017) using 25-dimensional GloVe word embedding (Pennington et al., 2014). We use $n = 500$ unlabeled samples per task and label a subset of $m = 400$ examples for each of the selected tasks. The remaining data is used for testing.

Methods. We evaluate the proposed method in the case when the set of labeled tasks is predefined (referred to as *DA*) by setting the set I to be a random subset of tasks and minimizing (17) only with respect to α-s and in the active task selection scenario where (17) is minimized with respect to both I and α-s (referred to as *Active DA*). We compare these methods to a multi-task method based on (Khosla et al., 2012), also with random labeled tasks (the same ones as for *DA*). Specifically, we solve:

$$\min_{w,v,b} C\Big(\|w\|^2 + \frac{1}{k}\sum_{t\in I}\|v_t\|^2\Big) + \frac{1-\gamma}{km}\sum_{t\in I,(x,y)\in\overline{S}_t}(w^T x + b - y)^2 + \frac{\gamma}{km}\sum_{t\in I}\sum_{(x,y)\in\overline{S}_t}((w^T + v_t^T)x + (b + b_t) - y)^2 \quad (20)$$

for $\gamma \in \{0, 0.1, \ldots, 1\}$ and use (w, b) for making predictions on all unlabeled tasks and $(w + v_t, b + b_t)$ for each labeled task $t \in I$. For every number of labeled tasks we report the result for γ that has the best test performance averaged over 10 repeats (denoted by *Multi-task*), as an upper performance bound on what could be achieved by model selection.

We also evaluate the discussed simplification of the proposed methods that consists of minimizing (19). We refer to these as *DA-SS* (for random predefined labeled tasks) and as *Active DA-SS* (in the active task selection scenario). The SS stands for *single source*, as in this setting, each task is solved based on information from only one labeled tasks.

To provide further context for the results we also report the results of learning independent ridge regressions with access to labels for all tasks (denoted by *Fully Labeled*). However, this baseline has access to many more annotated examples in total than all other methods. In order to quantify this effect we also consider the setting when the learner

[2] http://cvml.ist.ac.at/productreviews/

[3] http://jmcauley.ucsd.edu/data/amazon/

has access to labels for all tasks, but fewer of them: namely, when the number of labeled tasks is k, the number of labels per task is mk/T, i.e. the total amount of labeled examples is mk, the same as for all other methods. In this case we evaluate two methods. The first one learns ridge regressions for every task independently and thus can be seen as a reference point for the methods that do not involve information transfer between the labeled tasks, i.e. *DA-SS* and *Active DA-SS*. The second reference method is based on (Evgeniou & Pontil, 2004) and consists of minimizing (20) with γ set to 1 and processing all tasks as labeled. This approach transfers information between all the tasks and therefore we refer to it when evaluating the methods that involve information transfer between the labeled tasks, i.e. *DA*, *Active DA* and *Multi-task*.

Implementation. We estimate the empirical discrepancies between pairs of tasks by finding a hypothesis in $\mathcal{H}$ that minimizes the squared loss for the binary classification problem of separating the two sets of instances, as in (Ben-David et al., 2010). To minimize (17) for a given set of labeled tasks we use gradient descent. It is also used as a subroutine when minimizing (17) with respect to both I and α-s, for which we employ the GraSP algorithm (Bahmani et al., 2013). *Active DA-SS* involves the minimization of the k-medoid risk (19), which we perform using a local search as in (Park & Jun, 2009). For both methods for the active task selection scenario we used the heuristic from k-means++ (Arthur & Vassilvitskii, 2007) for initialization. To obtain classifiers for the individual tasks in all scenarios we use least-squares ridge regression. Regularization constants for all methods we selected from the set $\{0\}\cup\{10^{-17}, 10^{-16} \dots 10^{8}\}$ by 5×5-fold cross validation.

Results. The results are shown in Figure 4. First, one can see that the proposed domain adaptation-inspired method *DA* outperforms the multi-task method (20). This could be due to higher flexibility of *DA* compared to *Multi-task* as the latter provides only one predictor for all unlabeled tasks. Indeed, the difference is most apparent in the experiment with synthetic data, where by design there is no single predictor that could perform well on a large fraction of tasks. Results on the product reviews indicate that *DA*'s flexibility of learning a specific predictor for every task can be advantageous in more realistic scenarios as well.

Second, on both datasets both methods for active task selection, i.e. *Active DA* and *Active DA-SS*, outperform the corresponding passive methods, i.e. *DA* and *DA-SS*, systematically across various fractions of the labeled tasks. In particular, both active task selection methods require substantially fewer tasks labeled to achieve the same accuracy as their analogs with randomly chosen tasks. This confirms the intuition that selecting which tasks to label in a data-dependent way is beneficial and demonstrates that Theorem 1 is capable of capturing this effect.

Another interesting observation that can be made from the results in Figure 4 is that both active and passive domain adaptation-inspired methods clearly outperform the corresponding partially labeled baselines, especially for small fractions of labeled tasks. This indicates that having more labels for fewer tasks rather than only few labels for all tasks could be beneficial not only in terms of annotation costs, but also in terms of prediction accuracy.

As the number of labeled tasks gets larger, e.g. half of all tasks, the performance of the active task selection learner becomes almost identical to the performance of the *Fully Labeled* method, even improving over it in the case of multi-source transfer on synthetic data. This confirms the intuition that in the case of many related tasks even a fraction of the tasks can contain enough information for solving all tasks.

5. Conclusion

In this work we introduced and studied a variant of multi-task learning in which annotated data is available only for some of the tasks. This setting combines aspects of traditional multi-task learning, namely the transfer of information between labeled tasks, with aspects typical for domain adaptation problems, namely transferring information from labeled tasks to solve tasks for which only unlabeled data is available. The success of the learner in this setting depends on the effectiveness of information transfer and informativeness of the set of labeled tasks. We analyzed two scenarios: a passive one, in which the set of labeled tasks is predefined, and the *active task selection* scenario, in which the learner decides for which tasks to query labels.

Our main technical contribution is a generalization bound that quantifies the informativeness of the set of labeled tasks and the effectiveness of information transfer. We demonstrated how the bound can be used to make the choice of labeled tasks (in the active scenario) and to transfer information between the tasks in a principled way. We also showed how the terms in the bound have intuitive interpretations and provide guidance under which assumption of tasks relatedness the induced algorithm is expected to work well. Our empirical evaluation demonstrated that the proposed methods work also well in practice.

For future work we plan to further exploit the idea of active learning in the multi-task setting. In particular, we are interested in identifying whether by allowing the learner to make its decision on which tasks to label in an iterative way, rather than forcing it to choose all the tasks at the same time, one could obtain better learning guarantees as well as more effective learning methods.

Acknowledgments

We thank Alexander Zimin and Marius Kloft for useful discussions. This work was in parts funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036.

References

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Multi-task feature learning. In *Conference on Neural Information Processing Systems (NIPS)*, 2007.

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Convex multi-task feature learning. *Machine Learning*, 2008.

Arora, Sanjeev, Liang, Yingyu, and Ma, Tengyu. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*, 2017.

Arthur, David and Vassilvitskii, Sergei. k-means++: the advantages of careful seeding. In *Symposium on Discrete Algorithms (SODA)*, 2007.

Bahmani, Sohail, Raj, Bhiksha, and Boufounos, Petros T. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research (JMLR)*, 14(1), 2013.

Barzilai, Aviad and Crammer, Koby. Convex multi-task learning by clustering. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2015.

Ben-David, Shai, Blitzer, John, Crammer, Koby, and Pereira, Fernando. Analysis of representations for domain adaptation. In *Conference on Neural Information Processing Systems (NIPS)*, 2007.

Ben-David, Shai, Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, and Vaughan, Jennifer Wortman. A theory of learning from different domains. *Machine Learning*, 2010.

Blanchard, Gilles, Lee, Gyemin, and Scott, Clayton. Generalizing from several related classification tasks to a new unlabeled sample. In *Conference on Neural Information Processing Systems (NIPS)*, 2011.

Blitzer, John, Dredze, Mark, and Pereira, Fernando. Biographies, Bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

Caruana, Rich. Multitask learning. *Machine Learning*, 1997.

Chapelle, Olivier, Schölkopf, Bernhard, and Zien, Alexander. *Semi-supervised learning*. The MIT Press, 2006.

Crammer, Koby and Mansour, Yishay. Learning multiple tasks using shared hypotheses. In *Conference on Neural Information Processing Systems (NIPS)*, 2012.

El-Yaniv, Ran and Pechyony, Dmitry. Transductive Rademacher complexity and its applications. In *Workshop on Computational Learning Theory (COLT)*, 2007.

Evgeniou, Theodoros and Pontil, Massimiliano. Regularized multi-task learning. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

Gavves, Efstratios, Mensink, Thomas, Tommasi, Tatiana, Snoek, Cees, and Tuytelaars, Tinne. Active transfer learning with zero-shot priors: Reusing past datasets for future tasks. In *International Conference on Computer Vision (ICCV)*, 2015.

Glorot, Xavier, Bordes, Antoine, and Bengio, Yoshua. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference on Machine Learing (ICML)*, 2011.

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963.

Kang, Zhuoliang, Grauman, Kristen, and Sha, Fei. Learning with whom to share in multi-task feature learning. In *International Conference on Machine Learing (ICML)*, 2011.

Khosla, Aditya, Zhou, Tinghui, Malisiewicz, Tomasz, Efros, Alexei, and Torralba, Antonio. Undoing the damage of dataset bias. In *European Conference on Computer Vision (ECCV)*, 2012.

Kifer, Daniel, Ben-David, Shai, and Gehrke, Johannes. Detecting change in data streams. In *VLDB*, 2004.

Kumar, Abhishek and Daumé III, Hal. Learning task grouping and overlap in multi-task learning. In *International Conference on Machine Learing (ICML)*, 2012.

Lampert, Christoph H., Nickisch, Hannes, and Harmeling, Stefan. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2013.

Larochelle, Hugo, Erhan, Dumitru, and Bengio, Yoshua. Zero-data learning of new tasks. In *Conference on Artificial Intelligence (AAAI)*, 2008.

Mansour, Yishay, Mohri, Mehryar, and Rostamizadeh, Afshin. Domain adaptation: Learning bounds and algorithms. In *Workshop on Computational Learning Theory (COLT)*, 2009.

Maurer, Andreas. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 2006.

McAuley, J. J., Pandey, R., and Leskovec, J. Inferring networks of substitutable and complementary products. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2015a.

McAuley, J. J., Targett, C., Shi, Q., and van den Hengel, A. Image-based recommendations on styles and substitutes. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2015b.

McDiarmid, Colin. On the method of bounded differences. In *Surveys in Combinatorics*, 1989.

Palatucci, Mark, Pomerleau, Dean, Hinton, Geoffrey E, and Mitchell, Tom M. Zero-shot learning with semantic output codes. In *Conference on Neural Information Processing Systems (NIPS)*, 2009.

Pan, Sinno Jialin, Tsang, Ivor W, Kwok, James T, and Yang, Qiang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks (T-NN)*, 2011.

Park, Hae-Sang and Jun, Chi-Hyuck. A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, 36(2), 2009.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2014.

Pentina, Anastasia, Sharmanska, Viktoriia, and Lampert, Christoph H. Curriculum learning of multiple tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Reichart, Roi, Tomanek, Katrin, Hahn, Udo, and Rappoport, Ari. Multi-task active learning for linguistic annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008.

Ruvolo, Paul and Eaton, Eric. Active task selection for lifelong machine learning. In *Conference on Artificial Intelligence (AAAI)*, 2013.

Saha, Avishek, Rai, Piyush, Daumé III, Hal, and Venkatasubramanian, Suresh. Active online multitask learning. In *ICML Workshop on Budget Learning*, 2010.

Saha, Avishek, Rai, Piyush, Daumé III, Hal, and Venkatasubramanian, Suresh. Online learning of multiple tasks and their relationships. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2011.

Shimodaira, Hidetoshi. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000.

Tolstikhin, I., Blanchard, G., and Kloft, M. Localized complexities for transductive learning. In *Workshop on Computational Learning Theory (COLT)*, 2014.

Xing, Dikan, Dai, Wenyuan, Xue, Gui-Rong, and Yu, Yong. Bridged refinement for transfer learning. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2007.

Xue, Ya, Liao, Xuejun, Carin, Lawrence, and Krishnapuram, Balaji. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research (JMLR)*, 8:35–63, 2007.