Unsupervised Domain Adaptation with a Relaxed Covariate Shift Assumption

Tameem Adel  
University of Manchester, UK  
tameem.hesham@gmail.com

Han Zhao  
Carnegie Mellon University, USA  
han.zhao@cs.cmu.edu

Alexander Wong  
University of Waterloo, Canada  
a28wong@uwaterloo.ca

Abstract

Domain adaptation addresses learning tasks where training is performed on data from one domain whereas testing is performed on data belonging to a different but related domain. Assumptions about the relationship between the source and target domains should lead to tractable solutions on the one hand, and be realistic on the other hand. Here we propose a generative domain adaptation model that allows for modelling different assumptions about this relationship, among which is a newly introduced assumption that replaces covariate shift with a possibly more realistic assumption without losing tractability due to the efficient variational inference procedure developed. In addition to the ability to model less restrictive relationships between source and target, modelling can be performed without any target labeled data (unsupervised domain adaptation). We also provide a Rademacher complexity bound of the proposed algorithm. We evaluate the model on the Amazon reviews and the CVC pedestrian detection datasets.

Introduction

Domain adaptation focuses on modelling problems where the training and test distributions are different but related. The training and test domains are commonly referred to in the domain adaptation literature as the source and target domains, respectively. Domain diversity can emerge as a result of the scarcity of available labeled data from the target domain. It can as well be innate in the problem itself due to, for example, an ongoing change occurring to the source domain like in cases where the original source domain keeps changing over time. Domain adaptation aims at finding solutions for this kind of problem, where the training (source) data are generated from a distribution different from that of the test (target) data, by leveraging the available labeled data from the similar source domain (Ben-David et al. 2007).

More details about seminal works in the domain adaptation literature can be found in Section 1 in the supplementary.

The principal metric of a domain adaptation algorithm is the performance of the target learning hypothesis, which as a matter of fact depends on the labeled source data, (primarily unlabeled) target data and the relationship between the source and target domains. There can be multiple source domains to learn from (Mansour, Mohri, and Rostamizadeh 2009b), but the default setting is one source and one target domain. Regarding target data, several frameworks have access to a target sample containing labeled data instances, and a larger target unlabeled sample, e.g. Blitzer et al.; Oquab et al. (2008; 2014). In other frameworks, like Huang et al. (2006), referred to as unsupervised domain adaptation, the target data is fully unlabeled. The proposed model is an unsupervised domain adaptation framework, although it can incorporate target labeled data, if available, in a forthright manner.

The manner by which source labeled data are utilized in learning the target distribution and consequently the all-important target labels heavily depends on the relationship between the source and target domains. Such relationship is depicted by the learner’s assumptions about the similarities/differences between source and target domains. In this regard, covariate shift represents one of the most widely used domain adaptation assumptions (Sugiyama and Mueller 2005), on which much of the domain adaptation research is based. Covariate shift states that the conditional labeling distributions of both the source and target domains are the same, while their respective marginal data distributions can be different (Storkey and Sugiyama 2006; Ben-David and Urner 2012; 2014). Covariate shift is a valid assumption in some problems, but it can as well be quite unrealistic for many other domain adaptation tasks where the conditional label distributions are not (or, more precisely, not guaranteed to be) identical. The simplification resulting from assuming identical labeling distributions facilitates the quest for a tractable learning algorithm, albeit possibly at the cost of reducing the expressiveness power of the representation, and consequently the accuracy of the resulting hypothesis.

We propose a probabilistic relaxation of the covariate shift assumption where the more uncertain a learner is about the source labeling of an instance, the more uncertain it is that its target conditional labeling probability has a value similar to the corresponding source conditional labeling probability (the more uncertain it is that the label remains the same across domains), and vice versa. We formalize this notion in a generative model that learns a representation from both the target unlabeled data and a probabilistic prospect of the change in their labels across domains, and we use such representation to learn the target labels. The proposed model incorporates a tractable inference procedure for a broad range of problems by exploiting the recent advances in vari-
ational inference (Rezende, Mohamed, and Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013). We focus on promoting the proposed domain adaptation model for the introduced assumption, but the model is generic enough to substantiate other adaptation assumptions, and to be considered a generalization of some other adaptation frameworks. The proposed algorithm can be seen as a Bayesian, more generic and more scalable extension of Adel and Wong (2015).

Our main contributions are as follows:

• We propose a generalized generative modelling framework for unsupervised domain adaptation, which does not require any target labeled data.

• We develop a scalable learning algorithm that broadens the range of solution tractability with less restrictive and possibly more realistic domain adaptation assumptions than covariate shift.

• We propose a relaxed probabilistic version of covariate shift and exploit the proposed model in substantiating this assumption.

• Scalability and tractability of the model are achieved by developing a variational inference procedure tailored to domain adaptation, which is based on recent advances in variational Bayesian procedures, namely approximating the posterior via a recognition model (Rezende, Mohamed, and Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013; Kingma and Welling 2014; Kingma et al. 2014). We provide a lower bound on the marginal likelihood. To the best of our knowledge, this is the first variational inference procedure for domain adaptation.

• Our generalized domain adaptation model yields a rather flexible association between the adaptation assumptions and the learning solution so that conceptual modelling details and their implementations can each be scrutinized more easily and rather separately.

• A Rademacher complexity bound is derived for the proposed algorithm.

• The model is applied to benchmark domain adaptation datasets where its performance is demonstrated by comparison with the achieved state-of-the-art results. There are mainly two experiments: A sentiment analysis task on the Amazon reviews dataset and a pedestrian detection task on the CVC-02 and CVC-04 datasets.

The rest of the paper is organized as follows: The proposed relaxed covariate shift assumption is introduced in Section , followed by an illustration of the proposed unsupervised domain adaptation model. Section describes the developed variational inference procedure. A Rademacher generalization bound is introduced in Section , and finally the two experiments are presented in Section . The related work, the proof of the Rademacher complexity bound, among other details, can be found in the supplementary.

Generative Model for Domain Adaptation

Notation Let $Y$ be a label set. We address classification problems; $\forall y \in Y, y \in Z$, at the source and target domains. Denote the number of labels by $k$, $|Y| = k$. Let $x$ be a data instance. Input to the domain adaptation learner is composed of a source sample, $S$, consisting of $n$ labeled instances, $(x_i, y_i), i \in \{1, \ldots, n\}$, and a target sample $T$ of $m$ unlabeled instances, $x_j, j \in \{1, \ldots, m\}$. Define the loss function of two labeling hypotheses, $h_1$ and $h_2$, by $L(h_1(x), h_2(x)) : y \times y \rightarrow R$, and the expected loss over a distribution, $P$, as $\mathbb{E}_{P} (L(h_1(x), h_2(x)))$. Referring to the Bayes Optimal target labeling function as $h^\text{opt}_T(x)$, the main goal is to develop a learning hypothesis of the target domain, $h_T$, with a maximized classification accuracy, i.e. with a minimized expected target loss, $\mathbb{E}_{P_T}(h_T, h^\text{opt}_T)$. We refer to the proposed model as GenDA.

We begin by describing our relaxed covariate shift-based model for unsupervised domain adaptation. We assume one source and one target domain. We refer to the marginal data distributions in source and target domains as $P_S(x)$ and $P_T(x)$, respectively. The joint data and label distributions of the two domains are denoted by $P_S(x, y)$ and $P_T(x, y)$. Denote by $l_S(x)$ the probability of the assigned source label to $x$, i.e. the probability $P_S(y|x)$, with $y_{\text{max}} = \text{argmax}_{y \in Y} P_S(y|x)$. After learning the source hypothesis, we learn the latent feature space, $z$, which is assumed to generate both the unlabeled target instances, $x \in T$, and their corresponding source labels along with the relationship between source and target domains (modelled in our algorithm by the relaxed covariate shift assumption). We then learn the target labels, $y = \text{argmax}_{y \in Y} P_T(y|x)$, of the target sample, $x \in T$, by inferring $P_T(y|x)$. Learning target labels through $P_T(y|x)$ is more informative than through $P_T(y|x)$ since $z$ conveys information about both the unlabeled target data, $x$ and the corresponding source labels (source labels of the target data) mapped onto the target via the relaxed covariate shift assumption, or more generally via the assumption(s) on the relationship between source and target labels.

Relaxed Covariate Shift

We do not assume covariate shift, i.e. $l_S(x)$ and $l_T(x)$ do not have to be equal for each $x$. Instead, we construct a model where learning the target hypothesis is based on a nonlinear interaction of two respects: The topology of the unlabeled target data and their source labels along with the introduced probabilistic domain adaptation assumption. In the proposed model, we introduce and employ an assumption that is a relaxation of covariate shift. Even though we focus on this introduced assumption, the proposed model is not exclusively tailored for a solitary particular assumption. On the contrary, other assumptions on the relationship between source and target domains can be modelled, as we briefly show in Section 2 in the supplementary.

Rather than rigidly assuming the standard covariate shift, the relaxed covariate shift assumption states that the uncertainty in the source label assigned to a data instance, $x$, manifested in the source conditional labeling probability, is proportionate to the value of the loss function $L$ between its source and target labeling functions, $L(h_S(x), h_T(x))$. In other words, the more uncertain the source hypothesis about a labeling decision of an instance, the higher the probability its source and target labels are not identical. Assuming a 0-1
loss, \( L(h_S(x), h_T(x)) \), the relaxed covariate shift assumption is formulated as follows:

For \( x \in T, P_S(x) \neq 0 \), let \( L(h_S(x), h_T(x)) \sim Bin(1, p(x)) \), a binomial distribution with a parameter \( p(x) \) referring to the probability of success, i.e. the probability that \( L(h_S(x), h_T(x)) = 1 \) (the probability that the source and target labels of \( x \) are not identical). The value of \( p(x) \) follows a monotonically decreasing function of \( |l_S(x) - 1/k| \), for which we choose the exponential decay function:

\[
p(x) = f(|l_S(x) - 1/k|) = e^{-\lambda(|l_S(x) - 1/k|)}, \quad |l_S(x) - 1/k| \in [0, k^{-1}]
\]

where \( |l_S(x) - 1/k| \) signifies the uncertainty manifested in the source labeling decision, since \( l_S(x) = 1/k \) is the minimum possible value for \( l_S(x) \). Think of the binary classification case as an example, where the closer \( l_S(x) \) to \( 1/k = 0.5 \), the more uncertain the source labeling decision. The extremes in the spectrum of this uncertainty are: i) \( l_S(x) = 1 \) then \( |l_S(x) - 1/k| = 0 \), which denotes total certainty (minimum uncertainty), and ii) \( l_S(x) = 1/k \) leading to \( |l_S(x) - 1/k| = 0 \), denoting maximum uncertainty. Moving on through this spectrum from case i) to case ii), the value of \( p(x) \) keeps diminishing, corresponding to a decline in the probability of having a 1 (rather than 0) loss between the source and target hypotheses for \( x, L(h_S(x), h_T(x)) \). The relaxed covariate shift is a generalization of the standard covariate shift assumption, since setting the probability of a 1 loss, \( L(h_S(x), h_T(x)) = 1 \), to be always zero, \( p(x) = 0 \), for all values of \( |l_S(x) - 1/k| \) results in the standard covariate shift definition.

The measure of uncertainty developed in (1) is chosen because it is sound and also because it well captures the relationships between source and target domains in some applications (demonstrated by the experiments). However, other measures of uncertainty, such as the conditional entropy, \( -\sum_x \sum_y P_S(x) P_T(y|x) \log P_T(y|x) \), can be used instead. The relationship between the loss, \( L(h_S(x), h_T(x)) \), and the uncertainty manifested by the source hypothesis, \( |l_S(x) - 1/k| \) for \( k = 2 \), is illustrated in Figure 1 in the supplementary.

After learning parameters of the source hypothesis, the target sample, \( x \in T \), is given as input to the source hypothesis resulting in \( l_S(x) \) and the corresponding values of \( p(x) \) from (1). Afterwards one learns the latent representation, \( z \), that is assumed to generate both \( x \in T \), and \( p(x) \), where \( x \) conveys information about the topology of the target sample, and \( p(x) \) conveys information about the corresponding source labels along with probabilistic labeling information according to the relaxed probabilistic covariate shift assumption. Based on the latent representation, a grouping followed by a label alignment technique is pursued to extract the target labels. Extending the proposed form of the covariate shift assumption to other relationships between source and target domains, as well as to other loss functions, is not arduous.

**Learning**

The graphical model depicting the proposed model is displayed in Figure 2 in the supplementary. For clarity of presentation, the variable \( x \) (whose marginals in both domains are \( P_S(x) \) and \( P_T(x) \)), and \( y \) (whose conditional distributions in both domains are \( P_S(y|x) = \frac{P_S(y,x)}{P_S(x)} \) and \( P_T(y|x) \) are displayed twice in a slight abuse of notation. To learn the parameter set, \( \alpha \), of the source discriminative classifier (hypothesis), the source training data, \( (x, y) \in S, |S| = n \), is used to learn a discriminative classifier on the source domain (the upper plate of Figure 2 in the supplementary). A probabilistic output for each \( y|x \) is obtained via the learnt source discriminative classifier so that \( l_S(x) \) can be computed, followed by the computation of corresponding values of \( p(x) \) by (1).

Now we have learnt the parameters, \( \alpha \), of the source hypothesis. The intuition behind the target hypothesis is that we have an unlabeled target sample, and we base the learning on two aspects: i) topology of the unlabeled target data instances, and ii) the corresponding source labels of the target sample, which belong to a related (source) domain. The relationship between these two aspects as well as their impact on the sought target label is modelled by the latent variable, \( z \). The latent feature space, \( z \), is a representation that allows for a non-linear transformation of: i) the observed target data, and of ii) the probabilistic labeling information through the corresponding source labels and the relaxed probabilistic covariate shift assumption. The latent feature space, \( z \), represents such components in a more robust and lower-dimensional (than the original space of \( x \) and \( p(x) \) space) space, which potentially makes the target data more easily separable into their labels. The latent variable \( z \) models the relationship between each \( x \in T \) and their corresponding values of \( p(x) \) expressing probabilistic information about the 0-1 loss, \( L(h_S(x), h_T(x)) \).

Instances of the target sample, \( x \in T \), are given as input to the established source hypothesis (with parameters \( \alpha \)) resulting in \( l_S(x) \) for each \( x \in T \). Based on \( l_S(x) \), values of \( p(x) \) are indicated by (1). For each \( x \in T \), \( p(x) \) determines the probability of the target label being different from the corresponding source label, i.e. the probability of \( L(h_S(x), h_T(x)) = 1 \). There is one \( z \) per each \( x \in T \). The generative model utilized to learn the variable \( z \) corresponding to each data point \( x, x \in T \), is as follows (note that \( \theta \) refers to the generative parameters):

\[
P_\theta(z) = N(0, I): \text{standard Gaussian (\( \mu = 0, \sigma^2 = 1 \))}
\]

\[
P_\theta(x, p(x)|z) = f(x, p(x); z, \theta) = N(\mu_\theta(z), \sigma_\theta(z))
\]

The function \( f(x, p(x); z, \theta) \) is a non-linear transformation of \( z \), modelled by a neural network. The nonlinear relationship between \( z \) and \( (x, p(x)) \) renders the exact posterior computation intractable. Since the exact posterior, \( P_\theta(z|x, p(x)) \), is intractable, the variational posterior, \( q_\phi(z|x, p(x)) \) is computed via which approximate samples of \( z \) are used in the subsequent generative process including \( z \) and \( y^T \). The latter is a variable that groups the \( z \) variables, where each \( z \) is a representation of \( x \in T \), and the corresponding \( p(x) \). Regarding the variational posterior, a recognition model (Rezende, Mohamed, and Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013) \( q_\phi(z|x, p(x)) \) is developed for modelling \( z \). Note that \( \phi \) refers to the variational parameters. More details about the variational inference procedure are given in Section . The Gaussian chosen in (2) is
not a limitation of the domain adaptation model; assuming a
Gaussian \( P_{\theta}(z) \) facilitates the inference procedure as will be
shown in (4).

The grouping operation (with parameters \( \beta \) in Figure 2
in the supplementary is applied to \( z \), resulting in \( k \) groups,
where each group is expressed by a different value of the
grouping variable, \( y^U \), using Expectation-Maximization
(EM). By casting the resulting clusters into target labels, the
following label alignment method is pursued: i) We construct
a confusion matrix whose columns are the source labels of
data points \( x_i \in T \), (identified via \( l_S(x_i) \)), and rows are the
clusters, \( y^U \). ii) Out of values of each column, the row (clus-
ter) with the largest number of members is considered the
representative cluster of the respective column (class label),
i.e. the cluster whose members are the target members of
the source label at the column. The intuition behind this method
counts on the validity of the relaxed covariate shift assump-
tion, which considers the label change across domains to be
unlikely in case \( l_S(x) \) is large, i.e. \( l_S(x) \to 1 \). Its intuition
also counts on the general consensus of domain adaptation
that there should still be some similarity between source and
target domains so that an overall change would not span all
data points. Both source and target domains should still be
similar enough to allow for the idea of adaptation across
domains to be brought in, in the first place. Therefore, as-
suming that a majority of the data points in each class do
not change labels across different domains is quite reality-
istic (and still much more flexible than the assumption that
no label change at all can take place across domains). Note
that the term “grouping” is performed only w.r.t. \( z \), which
already had labeling information from \( p(x) \), and in turn from
\( L(h_S(x), h_T(x)) \), and for that it is not grouping (not fully
unsupervised) w.r.t. the observed data, \( x \in T \). The primary
steps of the algorithm, referred to as \text{GenDA}, are shown in
Algorithm 1.

Section 2 of the supplementary explains how some of the
other domain adaptation models can be seen as special cases
of the proposed model.

**Variational Inference**

**Evidence Lower Bound (ELBO)**

Due to the non-linear dependencies between the model com-
ponents, exact computation of the posterior \( P_{\theta}(z|x, p(x)) \)
is intractable. A scalable variational inference technique is
developed and optimized here. It is based on a variational
inference approach recently introduced in Rezende, Mo-
hamed, and Wierstra; Hoffman et al. (2014; 2013), which
takes into account producing a high-fidelity as well as com-
putationally efficient variational approach to estimate pos-
teriors. The pursued variational inference approach is
referred to as a recognition model (Rezende, Mohamed, and
Wierstra 2014; Stuhlmuller, Taylor, and Goodman 2013; King-
ma and Welling 2014). We derive a lower bound on the
marginal likelihood of the proposed generative domain
adaptation model and utilize it in establishing the objective
function used in computing a high-fidelity variational pos-
terior, \( q_{\phi}(z|x, p(x)) \). For a data instance \( x \in T \), and its
	 corresponding \( p(x) \), the variational bound, \( \mathcal{L}(x, p(x)) \) is:

\[
\log \int z P_{\theta}(x, p(x), z) \ dz = \log E_{q_{\phi}(z|x, p(x))} \frac{P_{\theta}(z)P_{\theta}(x, p(x)|z)}{q_{\phi}(z|x, p(x))} \geq \nonumber \\
\mathcal{L}(x, p(x)) = \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}(x_s, p(x_s))
\]

which yields the objective function used in optimizing the
generative and variational parameters, \( \theta \) and \( \phi \), respecti-
vely. For the recognition model of the latent variable, \( z \), we assume
a variational Gaussian \( q_{\phi}(z|x, p(x)) \):

\[
q_{\phi}(z|x, p(x)) = \mathcal{N}(\mu_{\phi}(x, p(x)), \sigma_{\phi}(x, p(x)))
\]

where \( \mu_{\phi}(x, p(x)) \) and \( \sigma_{\phi}(x, p(x)) \) are modelled as neu-
networks. The neural networks used in the generative and
variational inference models are multilayer perceptrons
(MLPs) with the softplus activation function, \( \log(1 + e^x) \). Two
hidden layers are used.

Back to (4), for the optimization of the objective func-
tion, we should differentiate \( \mathcal{L}(x, p(x)) \) w.r.t. generative and
variational parameters, \( \theta \) and \( \phi \), respectively. An
analytical computation of the second term in the third line
of (4), \( D_{KL}(q_{\phi}(z|x, p(x)) \parallel P_{\theta}(z)) \), is possible since both
distributions of the KL-divergence, \( P_{\theta}(z) \) and \( q_{\phi}(z|x, p(x)) \)
(from (2) and (3)) are Gaussians. Regarding the first term
of (4), \( E_{q_{\phi}(z|x, p(x))}\log P_{\theta}(x, p(x)|z) \), the workaroun-
d for the tricky step of taking gradients w.r.t. \( \phi \) is to repara-
terize the latent variable \( z = g_{\phi}(x, p(x), u) \) (Rezende, Mo-
hamed, and Wierstra 2014), where \( g_{\phi}(x, p(x), u) \) is a
deterministic function and all the randomness comes through
\( u \). Since \( z|x, p(x) \) is assumed to be Gaussian, we can apply
a location-scale transformation, \( z = g_{\phi}(x, p(x), u) =

Algorithm 1 Generative Domain Adaptation Algorithm (GenDA)

**Input:** source \((x, y) \in S\), size \( n + t \) target \( x \in T \), size \( m \),
where \((x_i, y_i)\) is a data instance-label pair.
**output:** target label, \( y, x \in T \).
**Source learning:** Learn \( P_S(y|x) \) from \((x, y) \in S\), by establish-
ing \( h_S; \) a source hypothesis.
**Source labels of \( x \in T\):**
- For \( x \in T \), learn source labels using \( h_S(x) \).
- Using the resulting \( l_S(x) \), compute:
  \[ p(x) = P(L(h_S(x), h_T(x)) = 1) \] by (1).
- Initialize \( \theta \) and \( \phi \) (generative \& variational param.)

**repeat**

**Perform minibatch stochastic gradient ascent on**
\( \mathcal{L}(x, p(x)) \) **to learn** \( \theta \) **and** \( \phi \):
- \( z_i \sim q_{\phi}(z|x_i, p(x_i)), \forall x_i \in B, B \) is a random
  minibatch.
- \( \mathcal{L} = \sum_{i=1}^{B} \mathcal{L}(x_i, p(x_i)) \).
- Take derivatives of \( \mathcal{L} \) w.r.t. \( \theta \) and \( \phi \).
- Update \( \theta \) and \( \phi \) accordingly

**until** \( \theta, \phi \) **do not change**
- Learn \( P(y^U|z) \) for \( x \in T \) by the EM algorithm.
- Learn target labels, \( y \in T \) from \( y^U \) by the label
  alignment procedure.

1694
\( \mu_{q}(x, p(x)) + \sigma_{q}(x, p(x)) \times u, u \in N(0, I) \). The first term in (4) can then be reformulated as:

\[
E_{q_{\phi}(z|x,p(x))} \left[ \log P_{h}(x, p(x)|z) \right] = \tag{6} \\
E_{u \sim N(0, I)} \left[ \log P_{h}(x, p(x)) \left[ \mu_{q}(x, p(x)) + \sigma_{q}(x, p(x)) \times u \right] \right]
\]

Gradients of (6) are:

\[
\nabla_{\theta, \phi} E_{q_{\phi}(z|x,p(x))} \left[ \log P_{h}(x, p(x)|z) \right] = \tag{7} \\
E_{u \sim N(0, I)} \left[ \nabla_{\theta, \phi} \left[ \log P_{h}(x, p(x)) \left[ \mu_{q}(x, p(x)) + \sigma_{q}(x, p(x)) \times u \right] \right] \right]
\]

And they can be computed via Monte Carlo estimates of the expectation (Kingma and Welling 2014). Gradients are computed here by stochastic gradient descent (SGD) (Bottou 2010) and AdaGrad (Duchi, Hazan, and Singer 2010).

**A Generalization Bound**

Several generalization bounds for domain adaptation have been introduced into the literature. Some of them are based on specific notions of distance between domains like the \( A \)-distance defined in Kifer, Ben-David, and Gehrke (2004) and used, most notably, in Ben-David et al. (2007), and like the discrepancy distance-based Rademacher complexity bound introduced in Mansour, Mohri, and Rostamizadeh (2009a). We derive a Rademacher complexity bound based on the introduced relaxed covariate shift assumption.

Define the discrepancy distance between the source and target domains as:

\[
\text{dsc-dist}(S, T) = \max_{h, h' \in H} \left| \mathbb{E}_{p_S}(h, h') - \mathbb{E}_{p_T}(h, h') \right| \tag{8}
\]

**Theorem 1.** Let \( h^*_S \in H \) be the best hypothesis in \( H \) at the source domain, i.e. \( h^*_S \in \arg \min_{h \in H} \mathbb{E}_{p_S}(h, h^*_S) \), where \( h^*_S \) is the Bayes Optimal, and similarly define \( h^*_T \) over the target domain. For any target hypothesis, \( h \in H \):

\[
\mathbb{E}_{p_T}(h, h^*_T) - \mathbb{E}_{p_T}(h^*_S, h^*_T) \leq \frac{1 - e^{-\lambda \left( \frac{1}{\lambda} \right)}}{\lambda^2} \left( \frac{(k-1)}{k} + 1 \right) + \text{dsc-dist}(S, T) + \mathbb{E}_{p_S}(h^*_S, h). \tag{9}
\]

**Proof.** The proof of Theorem 1 can be found in Section 5 of the supplementary.

Also, we can see from Theorem 1 that larger values of \( \lambda \) lead to a tighter generalization bound. As such, Theorem 1 supports the conclusions of (1) and Figure 1 in the supplementary that a larger \( \lambda \) leads to a smaller probability of \( L(h^*_S(x), h^*_T(x)) = 1 \), which signifies that the source and target domains are more similar to each other with larger \( \lambda \) values than cases with smaller \( \lambda \):

Larger \( \lambda \) \( \rightarrow \) More similar domains \( \rightarrow \) Tighter bound.

In a way, that can be seen as if the true value of \( \lambda \) between a source and a target domain provides a rough estimate of a distance between the two domains.

**Experiments**

We evaluate the proposed GenDA algorithm on the Amazon reviews dataset, and on the CVC-02 and CVC-04 pedestrian detection datasets. We start with the note that descriptions of some issues related to the three discriminative learners used at the source domain and to setting the parameters of GenDA, the reverse cross-validation method used, and finally example images related to the CVC data, are all left, due to space limitation, to Section 6 in the supplementary.

**Sentiment Analysis**

The dataset used in this experiment is the Amazon reviews dataset (Blitzer, Dredze, and Pereira 2007), which represents one of the benchmark domain adaptation datasets. Its intact version contains more than 340,000 reviews describing 22 product types. Since its original form is immensely unbalanced, the version of the Amazon reviews data used in domain adaptation, which was first introduced by Blitzer, Dredze, and Pereira (2007), consists of 4 product types where each product type refers to a domain. The four domains are: i) books (BK), ii) DVDs (DV), iii) electronics (EL), and iv) kitchen appliances (KA). Each Amazon review originally had a rating from 0 to 5 stars. For comparability, the convention followed in domain adaptation is to change the review ratings into either positive (higher than 3 stars) or negative (less than or equal to 3 stars) (Chen et al. 2012; Glorot, Bordes, and Bengio 2011). Features are pre-processed with standard tf-idf (Salton and Buckley 1988). The selected feature set is composed of 5,000 features of unigrams and bigrams. All these settings are done in order to compare on common grounds with state-of-the-art.

There are 2,000 labeled reviews per each of the 4 domains. Number of unlabeled reviews ranges between 3,586 KA reviews and 5,945 DV reviews. The two classes of the data are balanced in each of the 4 domains, i.e. ratio of positive ratings is 50%. We perform 12 adaptation tasks, moving from one domain to another, e.g. EL \( \rightarrow \) BK. Training on the source domain is performed on 2,000 labeled source data instances (reviews), i.e. the sample \( S \). The target learner is given a set of 2,000 unlabeled target instances. The unlabeled target learning set is used as the sample, \( T \), that is given as input to the source learner so that corresponding values of \( p(x) \) can be computed according to (1), and then the non-linear transformation between the unlabeled sample, \( T \) along with \( p(x) \) and the latent feature space, \( z \), is learnt. Afterwards, Expectation-Maximization (EM) followed by the label alignment method are used to learn the target labels of the sample \( T \), as illustrated in Section 5.

Results of the experiments performed on the 12 Amazon reviews adaptation tasks are displayed in Table 1. Comparisons with state-of-the-art results on the Amazon reviews data by Ganin et al. (2015) are based on classification accuracy. For each adaptation task, we compare to the best classification accuracy obtained by Ganin et al. (2015). A paired t-test with \( p = 0.05 \) is used to identify significance. The proposed algorithm GenDA outperforms the best result by Ganin et al. (2015) in 10 out of the 12 adaptation tasks (significantly so in 8 tasks). Concerning the source classifiers, the SVM classifier is clearly superior to the other two classifiers, and the logistic regression (LR) comes second in most of the tasks.

**Pedestrian Detection**

The task of pedestrian detection (Dollar et al. 2012; Enzweiler and Gavrila 2009) in images is challenging both due to diffi-
Table 1: Classification Accuracy of GenDA using 3 different source classifiers vs. state-of-the-art by taking the best available classification accuracy from Ganin et al. (2015), on the Amazon reviews dataset. Bold denotes the best result significantly outperforms all the competitors for such task.

<table>
<thead>
<tr>
<th>Source-Target</th>
<th>SVM GenDA</th>
<th>NN GenDA</th>
<th>LR GenDA</th>
<th>Ganin</th>
</tr>
</thead>
<tbody>
<tr>
<td>BK-DV</td>
<td>81.2%</td>
<td>72.4%</td>
<td>78.8%</td>
<td>79.9%</td>
</tr>
<tr>
<td>BK-EL</td>
<td>77.4%</td>
<td>75.1%</td>
<td>75.9%</td>
<td>79.2%</td>
</tr>
<tr>
<td>BK-KA</td>
<td>84%</td>
<td>74.7%</td>
<td>81.5%</td>
<td>81.6%</td>
</tr>
<tr>
<td>DV-BK</td>
<td>77.4%</td>
<td>70.2%</td>
<td>73.3%</td>
<td>75.3%</td>
</tr>
<tr>
<td>DV-EL</td>
<td>78.9%</td>
<td>73.9%</td>
<td>77.8%</td>
<td>78.6%</td>
</tr>
<tr>
<td>EL-BK</td>
<td>74.5%</td>
<td>65.5%</td>
<td>69.7%</td>
<td>72.7%</td>
</tr>
<tr>
<td>EL-DV</td>
<td>77.3%</td>
<td>76.1%</td>
<td>71.2%</td>
<td>76.5%</td>
</tr>
<tr>
<td>KA-BK</td>
<td>81%</td>
<td>71.4%</td>
<td>74.1%</td>
<td>72%</td>
</tr>
<tr>
<td>KA-DV</td>
<td>74.4%</td>
<td>70.1%</td>
<td>73.9%</td>
<td>74%</td>
</tr>
<tr>
<td>KA-EL</td>
<td>85.1%</td>
<td>79%</td>
<td>81.7%</td>
<td>84.3%</td>
</tr>
</tbody>
</table>

We compare our results to state-of-the-art unsupervised domain adaptation results on CVC datasets represented by Vazquez, Lopez, and Ponsa (2012; 2013), which are based on transductive SVM. We compare results based on the same metric used in Vazquez, Lopez, and Ponsa (2012; 2013), which is detection error tradeoff (DET) curve (Martin et al. 1997) showing the FPPI (false positive per image) rate vs. false negative (or missed detections) rate, for different thresholds. The smaller the area under the curve, the more accurate the algorithm. We implemented the algorithm in Vazquez, Lopez, and Ponsa (2012) with the help of coding excerpts taken from SVMlight (Joachims 1999a; 1999b).

Figure 1: A DET curve showing FPPI vs. missed detections rate resulting from applying GenDA over UDA, which we believe is mainly due to two reasons: Flexibility of the introduced relaxed covariate shift assumption, and the power of the latent feature representation, Z. Note that we stick to the DET curve, shown in Figure 1, as a metric so that we can compare on common grounds with state-of-the-art. Two examples from the CVC-02 dataset are displayed in Section 7 of the supplementary.

1we use the term “cropped image” to refer to a part of an image on which pedestrian detection has already been applied. It is the same term used in the CVC data descriptions.
References

Adel, T., and Wong, A. 2015. A probabilistic covariate shift assumption for domain adaptation. AAAI.


Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. NIPS.


Chen, M.; Xu, Z.; Weinberger, K.; and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. ICML.


Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentence classification: A deep learning approach. ICML.


Huang, J.; Gretton, A.; Borgwardt, K.; Schoelkopf, B.; and Smola, A. 2006. Correcting sample selection bias by unlabeled data. NIPS.


Joachims, T. 1999b. Transductive inference for text classification using support vector machines. ICML.


Kingma, D., and Welling, M. 2014. Auto-encoding variational Bayes. ICLR.


Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009a. Domain adaptation: Learning bounds and algorithms. COLT.


