The balancing principle for parameter choice in distance-regularized domain adaptation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We address the unsolved algorithm design problem of choosing a justified regularization parameter in unsupervised domain adaptation. This problem is intriguing as no labels are available in the target domain. Our approach starts with the observation that the widely-used approach of minimizing the source error, weighted by a distance measure between source and target feature representations, shares characteristics with regularized ill-posed inverse problems. Regularization parameters in inverse problems can be chosen by the fundamental *principle of balancing* approximation and sampling errors. We use this principle to balance learning errors and domain distance in a target error bound. As a result, we obtain a theoretically justified rule for the choice of the regularization parameter. In contrast to the state of the art, our approach allows source and target distributions with disjoint supports. An empirical comparative study on benchmark datasets underpins the performance of our approach.

1 Introduction

Domain adaptation uses the knowledge in the *source* domain to improve the performance of an algorithm on a related *target* domain [1]. In particular, domain adaptation tackles domain shifts in machine learning applications: Medical diagnostic systems should be adapted to new physical human variations; Industrial quality inspection systems should be accurate for new products; Self-driving cars should be able to adapt to new geographical environments and weather conditions. In this work, we focus on *unsupervised domain adaptation* where labels are only available in the source domain.

There are mainly two types of approaches for unsupervised domain adaptation: *importance weighting* [2, 3, 4, 5, 6] and *feature representation learning* [7, 8, 9, 10, 11, 12]. In this work, we focus on feature representation learning which goes beyond classical importance weighting by allowing a target distribution with support outside of the source distribution. The core idea behind feature representation learning approaches is to map the data into a new feature space where the source and target data representations appear similar, and where enough information is preserved for prediction [13]. The similarity is often realized by *regularization using distance measures* between source and target representations [14, 15, 16, 17, 18, 19, 20]. However, the performance of such methods crucially depends on the choice of the regularization parameter which penalizes the distance. The problem we investigate in this work is to choose this parameter, which is sophisticated without any target labels.

While remarkable theoretical results have been achieved which quantify the generalization ability of domain adaptation models [7, 21, 18, 19, 20], the choice of the regularization parameter which is crucial for finding such models has not systematically been addressed. Even though some parameter choice strategies exist, they are either purely heuristically driven or very limited by their assumptions [22]. Typical approaches are fixing the regularization parameters [11], minimizing the source error [15], balancing the source error and a distance [16] or (importance) weighting the samples by

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

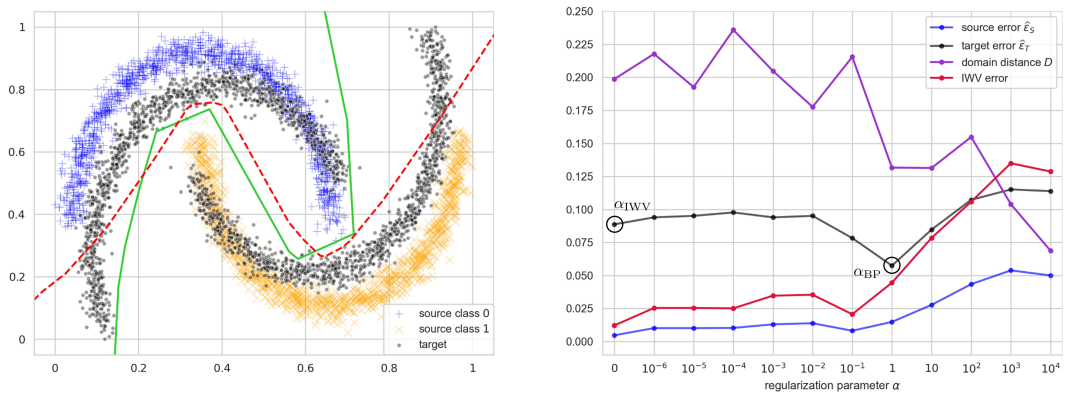

Figure 1: Unsupervised domain adaptation on transformed moons. Left: Target data (dots) partially outside of the support of the source data (blue $+$, orange $\times$). The common assumption of bounded density ratio is violated in large regions. In contrast, all our assumptions are satisfied. Our method identifies the best parameter of a domain adaptation algorithm [15] (green solid) which improves standard training on source data only (red dashed). Right: Regularization parameter (x-axis) which penalizes a distance [15] (purple) which leads to models with different source error (blue) and target error (black). The density ratio weighted error (red) Importance weighted validation (IWV) shows the smallest error (red) for models without domain adaptation ($\alpha_{\mathrm{IWV}} = 0$). In contrast, our approach identifies the optimal parameter ($\alpha_{\mathrm{BP}} = 1$).

the ratio between target and source density [3, 23, 22]. One common problem shared among all these approaches is that they all can fail if the density ratio is unbounded. Such unbounded density ration is typical for many of the high dimensional problems considered in machine learning [18], e.g. see Figure 1. Besides the aforementioned issues, the lack of principled strategies for parameter choice causes misinterpretations in the ranking of domain adaptation methods which are compared by performance, while often relying on different parameter choice strategies.

In this work, we propose a principled method for choosing distance-penalizing parameters in feature representation learning approaches for unsupervised domain adaptation. Our approach starts with the observation that the distance-regularization setting of domain adaptation shares characteristics with regularized ill-posed inverse problems (see Table 1). In inverse problems, the regularization parameter can be optimally chosen by the fundamental *balancing principle* which optimizes an approximation-sample (bias-variance) trade-off [24, 25, 26]. We apply this principle for balancing *domain distance* and *learning errors* of target error bounds. In particular, we approach the problem of non-computable terms in the target error bound by a new algorithmic criterion for approximating the value of balance. We call our method the Balancing Principle for Domain Adaptation (BPDA).

The BPDA is general in the sense that it can be applied based on different target error bounds, e.g. on [7, 21, 18, 19, 20]. To the best of our knowledge, the BPDA is the first principled method for parameter choice in unsupervised domain adaptation that allows an unbounded ratio between target and source density. We provide a bound on the generalization error of the best model corresponding to the parameter chosen by the BPDA. Finally, we empirically investigate the behavior of the BPDA based on two target error bounds, different domain adaptation methods and benchmark datasets. Our results show that BPDA outperforms or is on par with the state of the art on the problem of choosing the regularization parameter, on several domain adaptation methods; applied on different datasets.

2 Summary of results

Notation Let $\mathcal{X} \subset \mathbb{R}^n$ be an *input space* and $\mathcal{Y}$ be a discrete *label space*[1]. Following the classical setting of unsupervised domain adaptation [27], we consider two datasets: A *source dataset* $(\mathbf{x}, \mathbf{y}) = ((x_1, l_S(x_1)), \ldots, (x_s, l_S(x_s))) \in (\mathcal{X} \times \mathcal{Y})^s$ with *inputs* $x_1, \ldots, x_s$ independently drawn according to some source distribution (Borel probability measure) p_S on $\mathcal{X}$ and labeled according to some *labeling function* $l_S : \mathcal{X} \to \mathcal{Y}$, and, an unlabeled *target* dataset $\mathbf{x}' = (x'_1, \ldots, x'_t) \in \mathcal{X}^t$ with

[1]For simplicity, we use labeling functions instead of slightly more general concept of conditional distributions.

elements independently drawn according to some target distribution p_T on $\mathcal{X}$. Throughout this work, we focus on loss functions $L : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ which satisfy the triangle inequality and $L(y, y) = 0$. For example consider the 0-1 loss $L(y_1, y_2) := \mathbb{1}[y_1 \neq y_2]$, where $\mathbb{1}[P]$ is 1 iff the predicate P is true and 0 otherwise, and the quadratic loss function $L(y_1, y_2) := |y_1 - y_2|^2$. We denote the source error by $\varepsilon_S(f) = \varepsilon_S(f, l_S)$ with *cross-error* defined as $\varepsilon_S(f, g) := \mathbb{E}_{x \sim p_S}[L(f(x), g(x))]$ and its empirical sample estimate by $\widehat{\varepsilon}_S(f) = \widehat{\varepsilon}_S(f, l_S)$ with $\widehat{\varepsilon}_S(f, g) := \sum_{i=1}^{s} L(f(x_i), g(x_i))$. We denote the analogously defined target error by $\varepsilon_T(f)$, target cross-error by $\varepsilon_T(f, g)$ and its empirical sample estimate by $\widehat{\varepsilon}_T(f)$ with empirical cross-error $\widehat{\varepsilon}_T(f, g)$.

Learning setup In this work, we focus on feature representation learning algorithms for domain adaptation. These approaches aim at finding two learning models: A *representation* mapping $\phi \in \Phi \subset \{\phi : \mathcal{X} \to \mathcal{R}\}$ into some representation space $\mathcal{R} \subset \mathbb{R}^m$ and a *classifier* $g \in \mathcal{G} \subset \{g : \mathcal{R} \to \mathcal{Y}\}$. Loosely speaking, the aim is to find a mapping ϕ under which the source representations $\phi(\mathbf{x}) := (\phi(x_1), \ldots, \phi(x_s))$ and the target representations $\phi(\mathbf{x}') := (\phi(x'_1), \ldots, \phi(x'_t))$ appear similar, and, at the same time, enough information is preserved for prediction [13] by $g(x)$. A common approach to realize this aim is to solve the following objective function [18]

$$\min_{g \in \mathcal{G}, \phi \in \Phi} \widehat{\varepsilon}_S(g \circ \phi) + \alpha \cdot d(\phi(\mathbf{x}), \phi(\mathbf{x}')) \tag{1}$$

where d is a distance measure between source and target representations and $\alpha \in [0, \infty)$ is a parameter[2]. Good choices for d in Eq. (1) have been identified to be the Wasserstein distance [28, 29], the Maximum Mean Discrepancy [30, 31], moment distances [16, 32], adversarially learned distances [15, 33] and other measures of divergence [34, 35, 18, 19].

Problem For some $\alpha \in [0, \infty)$, let $g_\alpha \circ \phi_\alpha$ denote the minimizer of Eq. (1). Given an increasing sequence of parameters $\alpha_1, \ldots, \alpha_w \in [0, \infty)$ with $\alpha_1 = 0$, the problem studied in this work is to choose the parameter α in the sequence $\alpha_1, \ldots, \alpha_w$ with the lowest target error $\varepsilon_T(g_\alpha \circ \phi_\alpha)$.

Approach Our approach consists in minimizing a target error bound which satisfies the form

$$\varepsilon_T(g_\alpha \circ \phi_\alpha) \leq D(\alpha) + E(\alpha) \tag{2}$$

where $D(\alpha)$ gives a notion of *domain distance* (cf. [27]) by quantifying a distance between source and target data representations and $E(\alpha)$ comprises different *learning errors*. We assume that $E(\alpha)$ is bounded by some constant $B > 0$. The general form in Eq. (2) is satisfied by many error bounds [7, 21, 19, 20] which all can be taken as a basis for our approach. Having fixed a target error bound, the BPDA is realized by the following algorithm.

Algorithm 1: Balancing principle for domain adaptation (BPDA)

Input : Set of parameters $\alpha_1, \ldots, \alpha_w \in [0, \infty)$ with $\alpha_1 = 0$ and minimizers $f_1 := g_{\alpha_1} \circ \phi_{\alpha_1}, \ldots, f_w := g_{\alpha_w} \circ \phi_{\alpha_w}$ of Eq. (1).
Output : Parameter α_{BP} solving the problem above.
Initialization : $S = \{\}$
for $i = 1, \ldots, w$ **do**
  Compute empirical cross-error $\widehat{\varepsilon}_T(f_i, f_j)$ and domain distance $D(\alpha_j)$ for all $j = 1, \ldots, i-1$.
  if $\widehat{\varepsilon}_T(f_i, f_j) \leq D(\alpha_j) \left(2 + \frac{2B}{D(0)}\right)$ *for all* $j = 1, \ldots, i-1$ **then**
    $S := S \cup \{\alpha_i\}$
  end
end
return : $\alpha_{\mathrm{BP}} := \max S$

Properties of Algorithm 1 The BPDA has the following striking properties.

- The BPDA is a general procedure which can be instantiated by any error bound of the form in Eq. (2). See Section 4 and Section 5 for its application based on two different target error bounds [27, 19].
- In contrast to state-of-the-art methods, the BPDA does not assume a target labeling function l_T that is equal to the source labeling function l_S (covariate-shift assumption) and it does

[2] For simplicity we omit further regularization of ϕ and g.

not assume that the ratio between target and source density is bounded (sample selection bias assumption).

- The learning model $g_{\alpha_{\mathrm{BP}}} \circ \phi_{\alpha_{\mathrm{BP}}}$ identified by the BPDA satisfies a generalization guarantee, see Section 4.

Besides the properties above, the BPDA outperforms or is on par with the state of the art on the problem of choosing the regularization parameter, on several domain adaptation methods; applied on different datasets, see Section 5.

Related work Approaches which follow Eq. (1) are sometimes interpreted as learning *domain-invariant* representation. Note that minimizing $d(\phi(\mathbf{x}), \phi(\mathbf{x}'))$ in Eq. (1) to achieve $\phi(\mathbf{x}) = \phi(\mathbf{x}')$ differs from the conception as regularization [36, 37]. In fact, minimization of the distance means unjustified over-penalization which might lead to deteriorated performance [18]. Interestingly, our interpretation as regularization problem opens up a powerful toolbox of mathematical techniques. Our approach takes up the technique of balancing stability and approximation in regularized ill-posed inverse problems. The balancing principle has its origins in [24] devoted to nonparametric regression estimation and has been introduced in the context of ill-posed problems [25] and in supervised learning with kernels [38]. Following this line of research, we propose to apply the mathematical techniques underlying the balancing principle in the context of domain adaptation. The most related principled parameter choice methods in the context of unsupervised domain adaptation are importance-weighted cross-validation [3] and its extensions [23, 22]. In contrast to these methods, our method is not restricted by the assumption of a bounded ratio between target and source density. One empirically driven method which is related to ours is [16] which aims at balancing source error and the distance between source and target data representations. However, this method is not theoretically justified as it ignores the minimal combined error of a classifier on representations as defined in [7]. Our method relies on bounds on the target error such as e.g. [7, 21, 18, 19, 20].

3 Preliminaries

Sampling error bound Throughout this work, we generically denote an upper bound on the sampling error by $\eta_{z,\mathcal{G},\delta} \in [0, \infty)$, which is assumed to hold true with probability at least $1 - \delta$:

$$|\varepsilon_T(f, g) - \widehat{\varepsilon}_T(f, g)| \leq \eta_{t,\mathcal{G},\delta} \tag{3}$$

The bound $\eta_{t,\mathcal{G},\delta}$ depends on the sample size t, the function class $\mathcal{G}$ and the constant δ, and, it is assumed to satisfy $\eta_{t,\mathcal{G},\delta} \to 0$ for $t \to \infty$.

Two accompanying target error bounds In the seminal works [7, 27] binary classification $\mathcal{Y} := \{0, 1\}$ with 0-1 loss is considered and it is shown that the following bound holds for all $\phi \in \Phi$, $g \in \mathcal{G}$ and datasets of equal size $s = t$ with probability at least $1 - \delta$:

$$\varepsilon_T(g \circ \phi) \leq \varepsilon_S(g \circ \phi) + \lambda_{\mathcal{G}}(\phi) + \widehat{d}_{\mathcal{G}}(\phi(\mathbf{x}), \phi(\mathbf{x}')) + \eta_{t,\mathcal{G},\delta} \tag{4}$$

where

$$\widehat{d}_{\mathcal{G}}(\mathbf{x}, \mathbf{x}') := 2 \left(1 - \min_{g \in \mathcal{G}} \left[\frac{1}{s} \sum_{i=1}^{s} \mathbb{1}[g(x_i) = 0] - \frac{1}{s} \sum_{i=1}^{s} \mathbb{1}[g(x'_i) = 1] \right] \right) \tag{5}$$

is the empirical $\mathcal{G}$*-divergence* [39] which estimates the *generalized Kolmogorov-Smirnov distance* [40] and $\lambda_{\mathcal{G}}(\phi) := \inf_{g \in \mathcal{G}} (\varepsilon_S(g \circ \phi) + \varepsilon_T(g \circ \phi))$ is the minimum possible *combined error* determined by the application of $\mathcal{G}$. Using Eq. (3), the source error can be further upper bounded by the empirical source error. However, the term $\lambda_{\mathcal{G}}(\phi)$ cannot be estimated based on given datasets as it depends on the unknown labeling function l_T.

In [19], the bound in Eq. (4) is generalized to multiple classes $\mathcal{Y} := \{1, \ldots, k\}$ and scoring functions $\mathcal{H} \subset \{h : \mathcal{R} \to \mathbb{R}^k\}$ for which the output on each dimension indicates the confidence of prediction. For some $h \in \mathcal{H}$, let us denote by $g_h : \mathcal{R} \to \mathcal{Y}, x \mapsto \arg\max_{i \in \{1,\ldots,k\}} h^{(i)}(\phi(x))$ with $h^{(i)}(z)$ being the i-th component of $h(z)$. Let further ε_T denote the target error based on 0-1 loss. Then, with probability at least $1 - \delta$

$$\varepsilon_T(g_h \circ \phi) \leq \varepsilon_S^{(\rho)}(h \circ \phi) + \lambda_{\mathcal{H}}(\phi) + \widehat{d}_{h,\mathcal{H}}^{(\rho)}(\phi(\mathbf{x}), \phi(\mathbf{x}')) + \eta_{s+t,\mathcal{H},\delta} \tag{6}$$

where $\varepsilon_S^{(\rho)}(h) := \int_{\mathcal{X}} \Lambda_\rho \circ \rho_h(x, l_S)\, \mathrm{d}p_S(x)$ with $\rho_h(x, l_S) := \frac{1}{2}(h^{(l_S(x))}(x) - \max_{y \neq l_S(x)} h^{(y)}(x))$ and $\Lambda_\rho(x)$ being $1 - \frac{x}{\rho}$ if $0 \leq x < \rho$, 0 if $\rho \leq x$ and 1 otherwise. Further, the target terms T, p_T, l_T are defined analogously to S, p_S, l_S, respectively; the combined error is defined by $\lambda_{\mathcal{H}}(\phi) := \inf_{h \in \mathcal{H}}(\varepsilon_S^{(\rho)}(h \circ \phi) + \varepsilon_T^{(\rho)}(h))$, and the *empirical margin disparity discrepancy* (MDD) is given by

$$\widehat{d}_{h,\mathcal{H}}^{(\rho)}(\mathbf{x}', \mathbf{x}) := \max_{h' \in \mathcal{H}} \left[\frac{1}{s} \sum_{i=1}^{s} \Lambda_\rho \circ \rho_{h'}(x_i, g_h) - \frac{1}{t} \sum_{i=1}^{t} \Lambda_\rho \circ \rho_{h'}(x_i', g_h) \right] \tag{7}$$

Similarly to Eq. (4), the source error can be further estimated by an empirical error and the combined error $\lambda_{\mathcal{H}}(\phi)$ cannot be estimated based on given data.

Balancing principle for regularized inverse problems Let $\mathcal{H}$ and $\mathcal{K}$ be two Hilbert spaces and $V : \mathcal{H} \to \mathcal{K}$ be a linear operator. Then, the linear *inverse problem* associated to a given *datum* $g \in \mathcal{K}$ is to find some function f satisfying $Vf = g$, see e.g. [41, 42, 36, 37] and references therein. In general, the above problem is ill-posed, i.e. a solution does either not exist, is not unique or does not depend continuously on g. Existence and uniqueness can be approached by using the the following minimizer as approximation of f

$$f_{\mathcal{H}} \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \|Vf - g\|_{\mathcal{K}}^2$$

However, especially in the case of a noisy operator $\widehat{V}$ and noisy data $\widehat{g}$, continuous dependency on data becomes an important issue which can be restored using Tikhonov regularization [41]

$$f_\alpha \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \|\widehat{V} f - \widehat{g}\|_{\mathcal{K}}^2 + \alpha \|f\|_{\mathcal{H}}^2 \tag{8}$$

In many cases, probabilistic bounds on the error can be proven

$$\|f_\alpha - f_{\mathcal{H}}\|_{\mathcal{H}} \leq S(\alpha) + A(\alpha) \tag{9}$$

where $S(\alpha)$ is called *sampling error* made by considering noisy approximations $\widehat{V}$ and $\widehat{g}$ of V and g, respectively, and, $A(\alpha)$ is called *approximation error* originating from adding the regularizer $\|f\|_{\mathcal{H}}^2$. Commonly, $S(\alpha)$ decreases while $A(\alpha)$ increases with increasing α. This fact motivates the so-called *balancing principle* which aims at computing α^* such that $S(\alpha^*) = A(\alpha^*)$ [24, 25, 38, 37, 26]. As a result, the balancing principle provides a procedure for approximating α^* without having access to the values of $A(\alpha)$. The balancing principle obtains optimal error rates in many settings [43, 44, 45, 38, 37, 26].

4 Balancing principle for domain adaptation

In the following, we present the mathematical foundations of Algorithm 1 in two steps. In a first step, in Subsection 4.1, we state our assumptions and detail our idea of choosing the value of balance between terms in a target error bound. In a second step, in Subsection 4.2, we propose an algorithmic criterion, the *balancing principle estimate* for approximating the theoretical choice of the balancing value without target labels. Based on this criterion, we explain why we expect Algorithm 1 to be accurate.

4.1 Balancing terms in a target error bound

Assumptions In the following, let $g_\alpha \circ \phi_\alpha$ be a minimizer of Eq. (1) and assume that a target error bound holds which is of the form given by Eq. (2). Our approach is based on the plausible assumptions (a) that the function $\alpha \mapsto E(\alpha)$ is continuous and bounded by some constant $B > 0$, and (b) that the function $\alpha \mapsto D(\alpha)$ is continuous, non-increasing and non-degenerate, i.e. $D(0) > 0$. We further make the technical assumption (c) that $\lim_{\alpha \to \infty} \frac{D(\alpha)}{D(0)} \leq \sup_{\alpha \in [0,\infty)} \frac{E(\alpha)}{B}$. Note that (c) is satisfied in most standard cases, where $D(\alpha) \to 0$ for $\alpha, t \to \infty$. For example consider the two accompanying target error bounds in Section 3 as discussed at the end of this section.

Bridging regularized inverse problems and domain adaptation Under the assumptions above, our domain adaptation learning setup in Section 2 shares characteristics with the setting of regularized inverse problems as described in Section 3. Indeed, the Tikhonov regularizer $\|f\|_{\mathcal{H}}^2$ in Eq. (8) is similarly applied as the distance-regularizer $d(\phi(\mathbf{x}), \phi(\mathbf{x}'))$ in Eq. (1). In addition, error bounds for inverse problems as given by Eq. (9) show a similar form as target error bounds following Eq. (2). The sampling error $S(\alpha)$ in Eq. (9) decreases similarly to the domain distance in Eq. (2) and the approximation error $A(\alpha)$ in Eq. (9) cannot be estimated similarly to the learning errors $E(\alpha)$ in Eq. (2). However, in the domain adaptation setting, the term $E(\alpha)$ does not necessarily increase. We approach this issue by considering the *least non-decreasing majorant* of $E(\alpha)$ instead.

Definition 1 (Least non-decreasing majorant [46])**.** *The least non-decreasing majorant of* $E(\alpha)$ *is given by* $\overline{E}(\alpha) := \sup_{\beta \in [0,\alpha]} E(\beta)$.

Further upper bounding Eq. (2) by

$$\varepsilon_T(g_\alpha \circ \phi_\alpha) \leq D(\alpha) + E(\alpha) \leq D(\alpha) + \overline{E}(\alpha) \tag{10}$$

results in a form which satisfies all properties needed to apply the balancing principle (Table 1).

Table 1: Correspondences between regularized inverse problems and domain adaptation which allow to apply the balancing principle. See Section 3 for details on the inverse problem setting.

Tikhonov-regularized inverse problem	**Distance-regularized domain adaptation**
$f_\alpha \in \arg\min_{f \in \mathcal{H}} \|\widehat{A}f - \widehat{g}\|_{\mathcal{K}}^2 + \alpha \|f\|_{\mathcal{H}}^2$	$g_\alpha \circ \phi_\alpha \in \arg\min_{g \in \mathcal{G}, \phi \in \Phi} \widehat{\varepsilon}_S(g \circ \phi) + \alpha\, d(\phi(\mathbf{x}), \phi(\mathbf{x}'))$
$\|f_\alpha - f_{\mathcal{H}}\|_{\mathcal{H}} \leq S(\alpha) + A(\alpha)$	$\varepsilon_T(g_\alpha \circ \phi_\alpha) \leq D(\alpha) + E(\alpha)$
decreasing sampling error $S(\alpha)$	decreasing domain distance $D(\alpha)$
increasing approximation error $A(\alpha)$	bounded learning errors $E(\alpha)$
$A(\alpha)$ not estimable	$E(\alpha)$ not estimable
balance $A(\alpha^*) = S(\alpha^*)$	balance $\frac{D(\alpha^*)}{D(0)} = \frac{\overline{E}(\alpha^*)}{B}$

Balancing value for domain adaptation Having identified the shared characteristics between regularized inverse problems and domain adaptation, we now apply the techniques underlying the balancing principle to domain adaptation. We define the *balancing value* α^* as achieving

$$\frac{D(\alpha^*)}{D(0)} = \frac{\overline{E}(\alpha^*)}{B} \tag{11}$$

The value α^* of balance exists due to our assumptions. See Figure 2 for an illustration. Algorithm 1 approximates α^*. Note that if the bound in Eq. (10) is tight and its minimum $\inf_{\alpha \in [0,\infty)} D(\alpha) + \overline{E}(\alpha)$ is achieved then the balancing value α^* provides an optimal target error rate for $g_{\alpha^*} \circ \phi_{\alpha^*}$, or more precisely, it provides an error bound, which is only by a constant factor worse than the optimal one. See the supplementary material for a proof. This optimality property is shared with related regularization settings [43, 44, 45, 38, 26].

Two accompanying target error bounds Let us now discuss the reasoning above based on the two error bounds described in Section 3. First, recall the target error bound of [27] in Eq. (4). If we take $D(\alpha) := \widehat{d}_{\mathcal{G}}(\phi_\alpha(\mathbf{x}), \phi_\alpha(\mathbf{x}')) + \eta_{t,\mathcal{G},\delta}$ and $E(\alpha) := \varepsilon_S(g_\alpha \circ \phi_\alpha) + \lambda_{\mathcal{G}}(\phi_\alpha)$, it is natural to assume $D(\alpha)$ to decrease with α, especially for adversarial approaches which penalize the empirical $\mathcal{G}$-divergence $\widehat{d}_{\mathcal{G}}(\phi(\mathbf{x}), \phi(\mathbf{x}'))$, see [15]. It also holds that $E(\alpha) \leq 3$ for all $\alpha \in [0, \infty)$. For the balancing value to exist for large sample size t, we need to verify that $\frac{D(\alpha)}{D(0)} \to 0$ for $t, \alpha \to \infty$. However, this is satisfied for most function classes Φ, since it can be assumed that the constant function $\phi : x \mapsto c \in \mathbb{R}$ which achieves $\widehat{d}_{\mathcal{G}}(\phi(\mathbf{x}), \phi(\mathbf{x}')) = 0$ is contained in Φ. That is, $\widehat{d}_{\mathcal{G}}(\phi(\mathbf{x}), \phi(\mathbf{x}')) = 0$ can be achieved for $\alpha \to \infty$ and consequently $\frac{D(\alpha)}{D(0)} \to 0$ for $t, \alpha \to \infty$.

Consider now the bound in Eq. (6). By using $D(\alpha) := \widehat{d}_{h,\mathcal{H}}^{(\rho)}(\phi_\alpha(\mathbf{x}), \phi_\alpha(\mathbf{x}')) + \eta_{s+t,\mathcal{H},\delta}$, $E(\alpha) := \varepsilon_S^{(\rho)}(h \circ \phi) + \lambda_{\mathcal{H}}(\phi)$ and $B = 3$, all assumptions above are naturally satisfied similarly to the target error bound in Eq. (4) and the balancing value exists. In Section 5, we empirically investigate the performance of our method based on our two accompanying target error bounds. However, let us first show how we can overcome the problem of the non-computable term $E(\alpha)$.

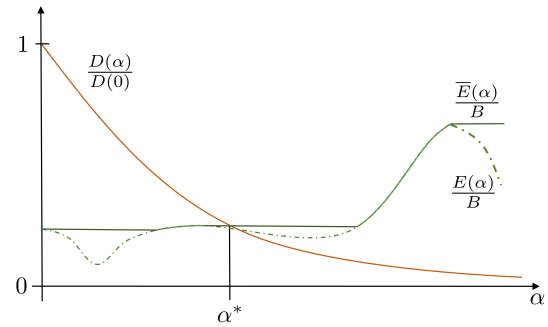

Figure 2: The BPDA in Algorithm 1 approximates α^* which balances $\frac{D(\alpha)}{D(0)}$ (orange) and the non-decreasing majorant $\frac{\overline{E}(\alpha)}{B}$ (green solid) of $\frac{E(\alpha)}{B}$ (green dashed).

4.2 Approximation of balancing value

Unfortunately, $E(\alpha)$ in Eq. (1) usually contains information about the target labeling function l_T and α^* can therefore not be calculated directly.

Balancing principle estimate We propose the following criterion for estimating α^* which is realized by the BPDA in Algorithm 1.

Definition 2 (Balancing principle estimate)**.** *Let $\alpha_1, \ldots, \alpha_w, w \in \mathbb{N}$ with $\alpha_1 = 0$ be an increasing sequence of values in $[0, \infty)$ and denote by $f_i := g_{\alpha_i} \circ \phi_{\alpha_i}$. Then, the balancing principle estimate is*

$$\alpha_{\mathrm{BP}} := \max \left\{ \alpha_i \,\middle|\, \forall j \in \{1, \ldots, i-1\} : \widehat{\varepsilon}_T(f_i, f_j) \leq D(\alpha_j) \left(2 + \frac{2B}{D(0)}\right) + \eta_{t,\mathcal{G},\delta} \right\} \tag{12}$$

The balancing principle estimate in Definition 2 is based on checking an upper bound on the empirical cross-error $\widehat{\varepsilon}_T(f_i, f_j)$ between two models f_i, f_j resulting from two different values α_i, α_j of the regularization parameter, respectively. The empirical cross-error does not contain information about the unknown target labels $l_T(x'_1), \ldots, l_T(x'_t)$ and can be computed based on available data. The main reason why we expect the balancing principle estimate α_{BP} to be near to the balancing value in Eq. (11) can be explained as follows.

Lemma 1. *Let $\delta \in (0, 1)$, $\alpha, \beta \in [0, \infty)$ and denote by $f_\alpha := g_\alpha \circ \phi_\alpha$. If $0 \leq \alpha \leq \beta \leq \alpha^*$ then the following holds with probability at least $1 - \delta$:*

$$\widehat{\varepsilon}_T(f_\alpha, f_\beta) \leq D(\alpha) \left(2 + \frac{2B}{D(0)}\right) + \eta_{t,\mathcal{G},\delta} \tag{13}$$

Lemma 1 (see the supplementary material for its proof) shows that the inequality in Eq. (12) is satisfied if $\alpha_j \leq \alpha_i \leq \alpha^*$. This implies that $\alpha_i > \alpha^*$ if the criterion is violated for some $\alpha_j \leq \alpha_i$. Consequently, the maximum α_{BP} as defined in Eq. (12) from an increasing sequence $\alpha_1, \ldots, \alpha_w$ which violates Eq. (13) for some $j \in \{1, \ldots, i-1\}$ can be assumed to be near to α^*.

Generalization guarantee The model $g_{\alpha_{\mathrm{BP}}} \circ \phi_{\alpha_{\mathrm{BP}}}$ obtained by the BPDA in Algorithm 1 satisfies the following generalization guarantee.

Theorem 1. *Let $\delta \in (0, 1)$ and $\alpha_1, \ldots, \alpha_w \in [0, \infty), \alpha_1 = 0$ be an increasing sequence such that*

$$D(\alpha_l) \leq q \cdot D(\alpha_{l+1}) \tag{14}$$

for all $l \in \{1, \ldots, w-1\}$ and some $q > 1$. Then, with probability at least $1 - \delta$

$$\varepsilon_T(g_{\alpha_{\mathrm{BP}}} \circ \phi_{\alpha_{\mathrm{BP}}}) \leq D(\alpha^*) \left(3 + \frac{3B}{D(0)}\right) q + \eta_{t,\mathcal{G},\delta} \tag{15}$$

Theorem 1 shows that the target generalization error of the BPDA has the same error rate as $D(\alpha^*)$ for $t \to \infty$. Moreover, the rate of the error $g_{\alpha_{\mathrm{BP}}} \circ \phi_{\alpha_{\mathrm{BP}}}$ is optimal if the bound in Eq. (2) is tight and

its minimum is achieved, or more precisely, it is only by a constant factor worse than the optimal one (see the supplementary material for a proof).

It is well known that appropriate assumptions as given above are needed for successful domain adaptation [47, 18]. However, in practice, target error bounds are often not tight and optimality cannot be guaranteed. In the following section, we therefore investigate the performance of our method on benchmark datasets based on our two accompanying target error bounds.

5 Empirical Evaluations

We empirically investigate the performance of our approach based on two target error bounds, two parameter selection methods, three datasets and different domain adaptation methods.

5.1 Setup

Given a domain adaptation algorithm that follows Eq. (1), the goal is to identify the regularization parameter α from the sequence $0, 10^{-6}, 10^{-5}, \ldots, 10^{3}, 10^{4}$ which leads to the smallest target error of the model learned by the algorithm[3].

Datasets We rely on one academic example which we call *Transformed Moons*. Transformed Moons shows a density ratio that is unbounded in large regions, see Figure 1. We also use the *Amazon Reviews* dataset [48]. The dataset contains text reviews for four categories of products: books (B), DVDs (D), electronics (E), and kitchen appliances (K). Reviews are encoded in 5000 dimensional feature vectors of bag-of-words unigrams and bigrams with binary labels: label 0 if the product is ranked by 1 to 3 stars, and label 1 if the product is ranked by 4 or 5 stars. From the four categories we obtain twelve domain adaptation tasks where each category serves once as source domain and once as target domain. We use the same data splits as previous works [49, 50, 15]. Thus, we have 2000 labeled source examples and 2000 unlabeled target examples for training, and between 3000 and 6000 examples for testing. Our third dataset is the *DomainNet–2019* dataset consisting of six different domains, each having 345 classes, and an average count of 288 images per class, i.e. around 0.6 million images [51]. However, our focus is not on large-scale domain adaptation, but rather on ranking model selection methods; hence, we propose a reduced version of the DomainNet–2019 dataset. In particular, we reduce the number of classes to five. We call our new dataset *MiniDomainNet*. See the supplementary material for the dataset statistics.

Balancing principle for domain adaptation We apply the BPDA in Algorithm 1 based on the two accompanying target error bounds described in Section 3. On the two-class datasets Transformed Moons and Amazon Reviews, we apply the BPDA based on Eq. (4). We set the domain distance $D(\alpha) := \widehat{d}_{\mathcal{G}}(\phi_\alpha(\mathbf{x}), \phi_\alpha(\mathbf{x}')) + \eta_{t,\mathcal{G},\delta}$, approximate its value by a classifier as proposed in [15, Subsection 3.2]. following [27, Subsection 7.2]. In our experiments on the multi-class dataset MiniDomainNet, we apply the BPDA based on Eq. (6). We define $D(\alpha) := \widehat{d}^{(\rho)}_{h,\mathcal{H}}(\phi_\alpha(\mathbf{x}), \phi_\alpha(\mathbf{x}')) + \eta_{s+t,\mathcal{H},\delta}$ and approximate its value as proposed in [19, Subsection 4.2]. In both variants of the BPDA, one each bound, we follow the argument of [27, Subsection 7.2] to have enough unlabeled data to evaluate the bounds without considering the finite sample error term $\eta_{t,\mathcal{G},\delta}$. We repeat each domain adaptation model training several times and check the if-statement in Algorithm 1 for each repetition. We define it to be violated, if it is violated in at least one of the repetitions.

Parameter choice baselines We compare our parameter choice approach to four baselines. The first baseline is training on source data only (SO). The second baseline is the best target error (TB) and it serves as a lower bound for the error. The third baseline is *importance weighted validation* [3] (IWV). We follow [22] and use *held-out* validation, i.e. we hold out a part of the training data as validation set, and we compute the importance weights based on this validation set. We also follow [52] and [22, Subsection 4.3] to estimate the importance weight by a classifier trained to separate source from target data. The classifier is tuned separately for each task and dataset such that its validation misclassification error is at most 0.05. For MiniDomainNet, we compute the importance weight based on the features of the pre-trained ResNet-18 [53]. The fourth baseline is *deep embedded validation* [22] (DEV) which solves the unbounded variance issue in IWV. Following [22], DEV is applied on the features of the neural networks.

[3]. The source-code can be found at `https://anonymous.4open.science/r/bpda-1D3D`

Domain adaptation methods In our experiments, we use three domain adaptation techniques. Domain-adversarial neural networks [15] (DANN), Maximum Mean Discrepancy [54, 14] (MMD) and Central Moment Discrepancy [17] (CMD). The details of all neural network architectures used, as well as the training strategy and hyperparameters are provided in the supplementary material.

5.2 Results

Average values of the results of our experiments on Transformed Moons and Amazon Review are summarized in Table 2. For the full tables on Transformed Moons, Amazon Reviews and MiniDomainNet, see the supplementary material. Although making no assumptions on the ratio between target and source density, the BPDA outperforms related parameter choice methods (IWV, DEV) on almost all tasks on Transformed Moons and Amazon Reviews, and, obtains competitive results on MiniDomainNet.

Table 2: Average target misclassification error with best values in bold. See the supplementary material for the full results for all datasets, domain adaptation tasks and random repetitions.

	Transformed Moons					Amazon Reviews				
Method	SO	IWV	DEV	BPDA	TB	SO	IWV	DEV	BPDA	TB
MMD	0.21	0.20	0.34	**0.16**	0.16	0.27	0.25	0.25	**0.22**	0.21
DANN	0.18	0.18	0.17	**0.12**	0.12	0.28	0.28	0.32	**0.27**	0.24
CMD	0.21	0.20	**0.19**	**0.19**	0.18	0.28	0.24	0.25	**0.21**	0.21
Avg.	0.20	0.19	0.23	**0.16**	0.15	0.28	0.26	0.27	**0.23**	0.22

6 Conclusion and Future Work

One widely-used technique for unsupervised domain adaptation is to map the data into a new feature space where the source and target data representations appear similar, and where enough information is preserved for prediction [13]. The similarity is often realized by minimizing the source error weighted by a distance measure between source and target representations. One common interpretation of this approach is to learn *domain-invariant* representations. However, there is a subtle difference between *distance-regularization* by penalizing the error minimization with a distance as above, and *distance-minimization* which results in domain-invariant representations. The latter can even lead to deteriorated performance [18]. In contrast, the interpretation as regularization problem opens up a powerful toolbox of mathematical techniques. We take up the technique of balancing stability and approximation in the regularization of ill-posed inverse problems, to tackle the problem of choosing the crucial regularization parameter in distance-regularized domain adaptation. Our approach overcomes the lack of target labels, it satisfies a generalization guarantee, and is (to the best of our knowledge) the first theoretically justified method that allows source and target distributions with disjoint supports. Finally, our approach outperforms or is on par with the state of the art on the problem of choosing the regularization parameter, on several domain adaptation methods; applied on different datasets.

Broader Impact Many industrial applications, such as manufacturing, personalized medicine or analytical chemistry, have to handle problems of domain shift together with issues of data limitations. These areas can profit from our research, as it provides a principled way of choosing a justified regularization parameter of unsupervised domain adaptation algorithms. Our method provides a high level of trust by applying mathematical techniques with guarantees originally developed in the area of ill-posed inverse problems. However, in domain adaptation in general, one critical point is the bias stored in the source data (e.g. past product, previous patient data, previous spectroscope), since knowledge in this data is used to improve the learning on the unlabeled target data (e.g. new product, new patient data, new spectroscope). If such a bias is present in one of the source domains, the predictions in new target domains might always suffer from this bias. We therefore suggest to not only look at new data in new target domains but more importantly consider dataset bias in already collected labeled source data.

References

[1] Sinno J Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineeringn*, 22(10):1345–1359, 2010.

[2] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[3] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

[4] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, pages 601–608, 2007.

[5] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

[6] Corinna Cortes, Mehryar Mohri, and Andrés M Medina. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

[7] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 137–144, 2007.

[8] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, 2007.

[9] Sinno J Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[10] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, pages 222–230, 2013.

[11] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[12] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *Proceedings of the International Conference on Machine Learning*, pages 2839–2848, 2016.

[13] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[14] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning*, pages 97–105, 2015.

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

[16] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 443–450, 2016.

[17] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *International Conference on Learning Representations*, 2017.

[18] Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 527–536, 2019.

[19] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *Proceedings of the International Conference on Machine Learning*, pages 7404–7413, 2019.

[20] Victor Bouvier, Philippe Very, Clément Chastagnol, Myriam Tami, and Céline Hudelot. Robust domain adaptation: Representations, weights and inductive bias. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2020.

[21] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *Proceedings of the Annual Conference on Learning Theory*, 2009.

[22] Kaichao You, Ximei Wang, Mingsheng Long, and Michael Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, pages 7124–7133. PMLR, 2019.

[23] Wouter M Kouw, Jesse H Krijthe, and Marco Loog. Robust importance-weighted cross-validation under sample selection bias. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2019.

[24] OV Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

[25] Alexander Goldenshluger and Sergei V Pereverzev. Adaptive estimation of linear functionals in hilbert scales from indirect white noise observations. *Probability Theory and Related Fields*, 118(2):169–186, 2000.

[26] Shuai Lu, Peter Mathé, and Sergei V Pereverzev. Balancing principle in supervised learning for a general regularization scheme. *Applied and Computational Harmonic Analysis*, 48(1):123–148, 2020.

[27] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[28] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[29] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[30] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 769–776, 2013.

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[32] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6629–6640, 2017.

[33] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[34] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the International Conference on Machine Learning*, pages 10–18, 2013.

[35] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2015.

[36] Heinz W Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[37] Shuai Lu and Sergei V Pereverzev. *Regularization theory for ill-posed problems: selected topics*, volume 58. Walter de Gruyter, 2013.

[38] Ernesto De Vito, Sergei V Pereverzyev, and Lorenzo Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10(4):455–479, 2010.

[39] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases-Volume 30*, pages 180–191. VLDB Endowment, 2004.

[40] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

[41] Andrey N Tikhonov and Vasiliy Y Arsenin. Solutions of ill-posed problems. *New York*, 1:30, 1977.

[42] CW Groetsch. The theory of tikhonov regularization for fredholm equations. *104p, Boston Pitman Publication*, 1984.

[43] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

[44] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[45] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[46] Gord Sinnamon. Transferring monotonicity in weighted norm inequalities. *Collectanea Mathematica*, pages 181–216, 2003.

[47] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[48] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006.

[49] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. *Proceedings of the International Conference on Machine Learning*, pages 767–774, 2012.

[50] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair auto encoder. *International Conference on Learning Representations*, 2016.

[51] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

[52] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning*, pages 81–88, 2007.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[54] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2006.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Subsection 4.2.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See paragraph **Assumptions** in Subsection 4.1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the supplementary results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 5.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Subsection 5.1 and the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See the supplementary material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See the supplementary material.
   (b) Did you mention the license of the assets? [Yes] See the supplementary material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See the source code referenced in Section 5.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 5.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]