Best-Effort Adaptation

Pranjal Awasthi[1], Corinna Cortes[2] and Mehryar Mohri[2,3*]

[1] Google Research, Mountain View, CA, USA .
[2] Google Research, 111 8th Avenue, New York, 10011, NY, USA .
[3] Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, 10012, NY, USA .

*Corresponding author(s). E-mail(s): mohri@google.com;
Contributing authors: pranjalawasthi@google.com;
corinna@google.com;

Abstract

We study a problem of *best-effort adaptation* motivated by several applications and considerations, which consists of determining an accurate predictor for a target domain, for which a moderate amount of labeled samples are available, while leveraging information from another domain for which substantially more labeled samples are at one's disposal. We present a new and general discrepancy-based theoretical analysis of sample reweighting methods, including bounds holding uniformly over the weights. We show how these bounds can guide the design of learning algorithms that we discuss in detail. We further show that our learning guarantees and algorithms provide improved solutions for standard domain adaptation problems, for which few labeled data or none are available from the target domain. We finally report the results of a series of experiments demonstrating the effectiveness of our best-effort adaptation and domain adaptation algorithms, as well as comparisons with several baselines. We also discuss how our analysis can benefit the design of principled solutions for *fine-tuning*.

Keywords: Domain adaptation, Distribution shift, ML fairness.

1 Introduction

Consider the following adaptation problem that frequently arises in applications. Suppose we have access to a fair amount of labeled data from a target domain $\mathcal{P}$ and to a significantly larger amount of labeled data from a different domain $\mathcal{Q}$. How can

we best exploit both collections of labeled data to come up with as accurate a predictor as possible for the target domain $\mathcal{P}$? We will refer to this problem as the *best-effort adaptation problem* since we seek the best method to leverage the additional labeled data from $\mathcal{Q}$ to come up with a best predictor for $\mathcal{P}$. One would imagine that the data from $\mathcal{Q}$ should be helpful in improving upon the performance obtained by training only on the $\mathcal{P}$ data, if $\mathcal{Q}$ is not too different from $\mathcal{P}$. The question is how to measure this difference and account for it in the learning algorithm. This best-effort problem differs from standard domain adaptation problems where typically very few or no labeled data from the target is at one's disposal.

Best-effort adaptation can also be motivated by fairness considerations, such as racial disparities in automated speech recognition (Koenecke et al., 2020). A significant gap has been reported for the accuracy of speech recognition systems when tested on speakers of vernacular English versus non-vernacular English speakers. In practice, there is a substantially larger amount of labeled data available for the non-vernacular domain since it represents a larger population of English speakers. As a result, it might not be possible, with the training data in hand, to achieve an accuracy for vernacular speech similar to the one achieved for non-vernacular speech. Such a recognition system might therefore have only one method for equalizing accuracy between these populations: namely, degrading the system's performance on the larger population. Alternatively, one could instead formulate the problem of maximizing the performance of the system on the vernacular speakers, leveraging *all* the data available at hand to find the *best-effort* predictor for vernacular speakers.

Here, we present a detailed study of best-effort adaptation, including a new and general theoretical analysis of reweighting methods using the notion of discrepancy, as well as new algorithms and empirical evaluations. We further show how our analysis can be extended to that of domain adaptation problems, for which we also design new algorithms and report experimental results.

There is a very broad literature dealing with adaptation solutions for distinct scenarios and we cannot present a comprehensive survey here. Instead, we briefly discuss here the most closely related work and give a detailed discussion of previous work in Section 7. We also refer the reader to papers such as (Pan and Yang, 2009; Wang and Deng, 2018). Let us add that similar scenarios to best-effort adaptation have been studied in the past under some different names such as *inductive transfer* or *supervised domain adaptation* but with the assumption of much smaller labeled data from the target domain (Garcke and Vanck, 2014; Hedegaard et al., 2021).

The work we present includes a significant theoretical component and benefits from prior theoretical analyses of domain adaptation. The theoretical analysis of domain adaptation was initiated by Kifer et al. (2004) and Ben-David et al. (2006) with the introduction of a d_A*-distance* between distributions. The authors used this notion to derive VC-dimension learning bounds for the zero-one loss, which was elaborated on in subsequent works (Blitzer et al., 2008; Ben-David et al., 2010). Later, Mansour et al. (2009a) and Cortes and Mohri (2011, 2014) presented a general analysis of single-source adaptation for arbitrary loss functions, where they introduced the notion of *discrepancy*, a divergence measure adequately aligned with domain adaptation. Discrepancy coincides with the d_A-distance in the special case

of the zero-one loss. It takes into account the loss function and hypothesis set and, importantly, can be estimated from finite samples. The authors gave a discrepancy minimization algorithm based on a reweighting of the losses of sample points. We use their notion of discrepancy in our new analysis. Cortes et al. (2019) presented an extension of the discrepancy minimization algorithm based on the so-called *generalized discrepancy*, which both incorporates a hypothesis-dependency and works with a less conservative notion of *local discrepancy* defined by a supremum over a subset of the hypothesis set. The notion of local discrepancy has been since adopted in several recent publications, in the study of active learning or adaptation (de Mathelin et al., 2022; Zhang et al., 2019c, 2020) and is also used in part of our analysis.

While our main motivation is best-effort adaptation, in Section 3, we present a general analysis that holds for *all sample reweighting methods*. Our theoretical analysis and learning bounds are new and are based on the notion of discrepancy. They include learning guarantees holding uniformly with respect to the weights, as well as a lower bound suggesting the importance of the discrepancy term in our bounds. Our theory guides the design of principled learning algorithms for best-effort adaptation, BEST and SBEST, that we discuss in detail in Section 4. This includes our estimation of the discrepancy terms via DC-programming (Appendix A.3).

In Section 5, we further show how our analysis can be extended to the case where few labeled data or none are available from the target domain, that is the scenario of (unsupervised or weakly supervised) domain adaptation. Here too, we derive new discrepancy-based learning bounds based on reweighting, including uniform bounds with respect to the weights (Section 5.1). Interestingly, here, an additional set of sample weights naturally appears in the analysis, to account for the absence of labels from the target. Our theoretical analysis leads to the design of a new adaptation algorithms, BEST-DA (Section 5.2). We further discuss in detail how in this scenario labeled discrepancy terms can be upper-bounded in terms of unlabeled ones, including unlabeled local discrepancies, and how some additional amount of labeled data can be beneficial (Section 5.3).

In Section 6, we report the results of experiments with both our best-effort adaptation algorithms and our domain adaptation algorithms demonstrating their effectiveness, as well as comparisons with several baselines. This includes a discussion and empirical analysis of how our results benefit the design of principled solutions for *fine-tuning* and other few-shot algorithms. We start with the introduction of some preliminary definitions and concepts related to adaptation (Section 2).

2 Preliminaries

We write $\mathcal{X}$ to denote the input space and $\mathcal{Y}$ the output space. In the regression setting, $\mathcal{Y}$ is assumed to be a measurable subset of $\mathbb{R}$. We use $\mathcal{H}$ to represent a hypothesis set of functions from $\mathcal{X}$ to $\mathcal{Y}$, and $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function, with values in $[0, 1]$.

We will study problems with a source domain $\mathcal{Q}$ and target domain $\mathcal{P}$, where $\mathcal{Q}$ and $\mathcal{P}$ are distributions over $\mathcal{X} \times \mathcal{Y}$. We will denote by $\widehat{\mathcal{Q}}$ the empirical distribution associated to a sample S of size m drawn from $\mathcal{Q}^m$ and similarly by $\widehat{\mathcal{P}}$ the empirical distribution associated to a sample S' of size n drawn from $\mathcal{P}^n$. We will denote by

$\mathcal{Q}_X$ and $\mathcal{P}_X$ the marginal distributions of $\mathcal{Q}$ and $\mathcal{P}$ on $\mathcal{X}$. We will denote by $\mathcal{L}(\mathcal{P}, h)$ the population loss of a hypothesis over $\mathcal{P}$ defined as: $\mathcal{L}(\mathcal{P}, h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell((x), y)]$.

Several notions of discrepancy have been shown to be adequate measures between distributions for adaptation problems (Kifer et al., 2004; Mansour et al., 2009a; Mohri and Muñoz Medina, 2012; Cortes and Mohri, 2014; Cortes et al., 2019). We will denote by $\mathrm{dis}(\mathcal{P}, \mathcal{Q})$ the *labeled discrepancy* of $\mathcal{P}$ and $\mathcal{Q}$, also called $\mathcal{Y}$-discrepancy in (Mohri and Muñoz Medina, 2012; Cortes et al., 2019) and defined by:

$$\mathrm{dis}(\mathcal{P}, \mathcal{Q}) = \sup_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}}[\ell(h(x), y)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{Q}}[\ell(h(x), y)]. \tag{1}$$

We omit the use of absolute values around the difference of expectations, unlike the original discrepancy definitions in prior work, as the one-sided definition is sufficient for our analysis.

By definition, computing the labeled discrepancy assumes access to labels from both $\mathcal{P}$ and $\mathcal{Q}$. In contrast, the *unlabeled discrepancy*, denoted by $\overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q})$, requires no access to such labels

$$\overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q}) = \sup_{h,h'\in\mathcal{H}} \mathop{\mathbb{E}}_{x\sim\mathcal{P}_X}[\ell(h(x), h'(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{Q}_X}[\ell(h(x), h'(x))]. \tag{2}$$

As shown by Mansour et al. (2009a), the unlabeled discrepancy can be accurately estimated from finite (unlabeled) samples from $\mathcal{Q}_X$ and $\mathcal{P}_X$ when $\mathcal{H}$ admits a favorable Rademacher complexity, for example a finite VC-dimension. The unlabeled discrepancy is a divergence measure tailored to (unsupervised) adaptation that can be upper bounded by the ℓ_1-distance. It coincides with the so-called d_A*-distance* introduced by Kifer et al. (2004) in the special case of the zero-one loss. We will also be using the finer notion of *local labeled discrepancy* for some suitably chosen subsets $\mathcal{H}_1$ and $\mathcal{H}_2$ of $\mathcal{H}$:

$$\overline{\mathrm{dis}}_{\mathcal{H}_1\times\mathcal{H}_2}(\mathcal{P}, \mathcal{Q}) = \sup_{(h,h')\in\mathcal{H}_1\times\mathcal{H}_2} \mathop{\mathbb{E}}_{x\sim\mathcal{P}_X}[\ell(h(x), h'(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{Q}_X}[\ell(h(x), h'(x))]. \tag{3}$$

Local discrepancy (Cortes et al., 2019) is defined by a supremum over smaller sets and is thus a more favorable quantity.

We further extend all the discrepancy definitions just presented to the case where $\mathcal{P}$ and $\mathcal{Q}$ are finite signed measures over $\mathcal{X}\times\mathcal{Y}$, using the same expressions as above. In particular, for any two distributions $\mathcal{P}, \mathcal{Q}$ and real numbers a and b, we extend the definition of labeled discrepancy to that of $a\mathcal{P}$ and $b\mathcal{Q}$ as follows:

$$\mathrm{dis}(a\mathcal{P}, b\mathcal{Q}) = \sup_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}}[a\,\ell(h(x), y)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{Q}}[b\,\ell(h(x), y)]. \tag{4}$$

We also abusively extend the definition of discrepancy to distributions over sample indices. As an example, given the samples S and S' and a distribution q over their $[m+n]$ indices, we define the discrepancy $\mathrm{dis}(\widehat{\mathcal{P}}, \mathsf{q})$ as follows: $\mathrm{dis}(\widehat{\mathcal{P}}, \mathsf{q}) = \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=m+1}^{n} \ell(h(x_i), y_i) - \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i)$.

In our analysis of generalization we use an extended version of the standard notion of Rademacher complexity of a hypothesis set $\mathcal{H}$, $\mathfrak{R}_m(\mathcal{H})$, which is defined as

follows for i.i.d. samples $S = (x_1, \ldots, x_m)$ of size m (Koltchinskii and Panchenko, 2002; Bartlett and Mendelson, 2002; Mohri et al., 2018):

$$\mathfrak{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}, S}\left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i)\right],$$

where σ_is are independent uniform random variables taking values in $\{-1, +1\}$. The expectation is taken over σ_is and the draw of an i.i.d. sample S of size m from the distribution considered.

3 Discrepancy-based generalization bounds

There are many algorithms in adaptation based on various methods for reweighting sample losses and it is natural to seek a similar solution for best-effort adaptation (see Section 7). In this section, we present a general theoretical analysis covering all such sample reweighting methods. We introduce new discrepancy-based generalization bounds, including learning bounds holding uniformly over the weights. Next, we compare them with existing guarantees and derive a simplified corollary.

3.1 General learning bounds for sample reweighting methods

We assume that the learner has access to a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn from $\mathcal{Q}^m$ and a labeled sample $S' = ((x_{m+1}, y_{m+1}), \ldots, (x_{m+n}, y_{m+n}))$ drawn from $\mathcal{P}^n$. In the problems we consider, we typically have $m \gg n$, but our analysis applies is general. For a non-negative vector q in $[0, 1]^{[m+n]}$, we denote by $\overline{\mathsf{q}}$ the *total weight* on the first m points: $\overline{\mathsf{q}} = \sum_{i=1}^{m} \mathsf{q}_i$ and by $\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H})$ the q-*weighted Rademacher complexity*:

$$\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) = \mathop{\mathbb{E}}_{S, S', \boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m+n} \sigma_i \mathsf{q}_i \ell(h(x_i), y_i)\right], \tag{5}$$

where the Rademacher variables σ_i are independent random variables distributed uniformly over $\{-1, +1\}$. The q-weighted Rademacher complexity is a natural extension of the Rademacher complexity taking into consideration distinct weights assigned to sample points. It can be upper-bounded as follows in terms of the (unweighted) Rademacher complexity: $\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) \leq \|\mathsf{q}\|_\infty (m+n)\, \mathfrak{R}_{m+n}(\ell \circ \mathcal{H})$, with equality for uniform weights (see Lemma 10, Appendix A).

The following is a general learning guarantee expressed in terms of the weights q. Note that we do not require q to be a distribution over $[m+n]$, that is $\|\mathsf{q}\|_1$ may not equal one.

Theorem 1 *Fix a vector* q *in* $[0, 1]^{[m+n]}$. *Then, for any* $\delta > 0$, *with probability at least* $1 - \delta$ *over the draw of an i.i.d. sample* S *of size* m *from* $\mathcal{Q}$ *and an i.i.d. sample* S' *of size* n *from* $\mathcal{P}$,

the following holds for all $h \in \mathcal{H}$:

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \mathrm{dis}\Big(\big[(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}\big]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}\Big) + 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) + \|\mathsf{q}\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{2}}.$$

Note that when q is a distribution, the term $\mathrm{dis}(\big[(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}\big]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q})$ admits the following simpler form, since $\|\mathsf{q}\|_1 = 1$:

$$\mathrm{dis}(\big[(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}\big]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) = \mathrm{dis}(\overline{\mathsf{q}}\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) = \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}). \tag{6}$$

where the last equality holds by the definition (4). The following theorem shows that that the bound of Theorem 1 is tight as a function of the discrepancy term when q is a distribution, which emphasizes the crucial significance of this term. The proofs for both theorems are given in Appendix A.

Theorem 2 *Fix a distribution q in the simplex Δ_{m+n}. Then, for any $\epsilon > 0$, there exists $h \in \mathcal{H}$ such that, for any $\delta > 0$, the following lower bound holds with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{Q}$ and an i.i.d. sample S' of size n from $\mathcal{P}$:*

$$\mathcal{L}(\mathcal{P}, h) \geq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) - 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) - \|\mathsf{q}\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{2}} - \epsilon.$$

In particular, for $\|\mathsf{q}\|_2, \mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) \in O\Big(\frac{1}{\sqrt{m+n}}\Big)$, we have:

$$\mathcal{L}(\mathcal{P}, h) \geq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \Omega\left(\frac{1}{\sqrt{m+n}}\right).$$

The bound of Theorem 1 cannot be used to choose q since it holds for a fixed choice of that vector. A standard way to derive a uniform bound over q is via covering numbers. That requires applying the union bound to the centers of an ϵ-covering of $[0,1]^{[m+n]}$ for the ℓ_1 distance. But, the corresponding covering number $\mathcal{N}_1$ would be in $O((1/\epsilon)^{m+n})$, resulting in an uninformative bound, even for $\|\mathsf{q}\|_2 = 1/\sqrt{m+n}$, since $\sqrt{\log \mathcal{N}_1/m+n}$ would be a constant. Instead, we present an alternative analysis, generalizing Theorem 1 to hold uniformly over q in $\big\{\mathsf{q}\colon \|\mathsf{q} - \mathsf{p}^0\|_1 < 1\big\}$, where p^0 could be interpreted as a reference (or ideal) reweighting choice. The proof is presented in Appendix A.

Theorem 3 *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{Q}$ and an i.i.d. S' of size n from $\mathcal{P}$, the following holds for all $h \in \mathcal{H}$ and $\mathsf{q} \in \big\{\mathsf{q}\colon \|\mathsf{q} - \mathsf{p}^0\|_1 < 1\big\}$:*

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \mathrm{dis}\Big(\big[(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}\big]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}\Big) + \mathrm{dis}(\mathsf{p}^0, \mathsf{q})$$
$$+ 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) + 7\|\mathsf{q} - \mathsf{p}^0\|_1 + \Big[\|\mathsf{q}\|_2 + 2\|\mathsf{q} - \mathsf{p}^0\|_1\Big]\left[\sqrt{\log \log_2 \tfrac{2}{1 - \|\mathsf{q} - \mathsf{p}^0\|_1}} + \sqrt{\tfrac{\log \frac{2}{\delta}}{2}}\right].$$

Note that for $\mathsf{q} = \mathsf{p}^0$, the bound coincides with that of Theorem 1.

Learning bounds insights. Theorems 1 and 3 provide general guarantees for best-effort adaptation. They suggest that, for adaptation to succeed via sample reweighting, a favorable balance of *several key terms* is important. The first term suggests minimizing the q-weighted empirical loss. However, the bound advises against doing so at the price of assigning non-zero weights only to a small fraction of the points since that would increase the $\|\mathsf{q}\|_2$ term. In fact, a comparison with the familiar inverse of square-root of the sample size term appearing in other bounds suggests interpreting $\left(1/\|\mathsf{q}\|_2^2\right)$ as the *effective sample size*. As already indicated, when q is a distribution, the second term admits the following simpler form: $\mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) = \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q})$ (Equation 6). Thus, the second term of these bounds suggests allocating less weight to the points drawn from $\mathcal{Q}$, when the discrepancy $\mathrm{dis}(\mathcal{P}, \mathcal{Q})$ is large. The weighted discrepancy term $\mathrm{dis}(\mathsf{p}^0, \mathsf{q})$ and the ℓ_1-distance $\|\mathsf{q} - \mathsf{p}_0\|_1$ in Theorem 3 both press q to be chosen relatively closer to the reference p^0. Finally, the Rademacher complexity term is a familiar measure of the complexity of the hypothesis set, which here additionally takes into consideration the weights.

3.2 Discussion of learning bound of Theorem 1 and comparisons

Here, we compare the bound of Theorem 1 with some existing discrepany-based ones and show how they can be recovered as special cases. In particular, we show that the discrepancy-based bound of Cortes et al. (2019), which is the basis for the discrepancy minimization algorithm of Cortes and Mohri (2014), is always an upper bound on a special case (specific choice of the weights) of the bound of Theorem 1.

It is instructive to examine some special cases for the choice of q, which will demonstrate how our guarantees can recover several previous bounds as a special case. Since our algorithms seek to choose the best weight (and best hypothesis) based on these bounds, this shows that their search space includes that of algorithms based on those previous bounds.

We note that assigning non-uniform weights to the points in S should not be viewed as unnatural, even though the points are sampled from the same distribution. This is because these weights serve to make the q-weighted empirical loss closer to the empirical loss for the target sample. As an example, importance weighting seeks distinct weights for each point based on the source and target distributions. Nevertheless, we discuss a simple α-*reweighting* method, which allocates uniform weights to source points. We show that, under some assumptions, even for this very simple choice of the weights, the learning bound can be more favorable than the one for training only on target samples.

q chosen uniformly on S. For q chosen to be the uniform distribution on S, we have $\overline{\mathsf{q}} = 1$, $\|\mathsf{q}\|_2 = \frac{1}{\sqrt{m}}$, and the bound coincides with the labeled discrepancy-based bound for $\mathcal{P}$ of Cortes et al. (2019)[Prop. 5; Eq. (9)]. Indeed, for q chosen to be supported only on S, the theorem gives a q-discrepancy domain adaptation bound from $\mathcal{Q}$ to $\mathcal{P}$, in terms of a q-Rademacher complexity and $\|\mathsf{q}\|_2$.

q chosen uniformly on S'. Here $\overline{\mathsf{q}} = 0$, $\|\mathsf{q}\|_2 = \frac{1}{\sqrt{n}}$, and the bound coincides with the standard Rademacher complexity bound for $\mathcal{P}$ for learning from a labeled sample of size n:

$$\mathcal{L}(\mathcal{P}, h) \leq \frac{1}{n} \sum_{i=m+1}^{m+n} \ell(h(x_i), y_i) + 2\mathfrak{R}_n(\ell \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \tag{7}$$

Here, $\mathfrak{R}_n(\ell \circ \mathcal{H})$ is the standard Rademacher complexity defined as in (5) where the expectation is over S' and q is the uniform distribution over S'. Thus, for q minimizing the right-hand side of the bound of the theorem, the learning bound is at least as favorable as one restricted to learning from the labeled points from $\mathcal{P}$. But the bound also demonstrates that it is possible to do better than just learning from $\mathcal{P}$. In fact, for $\mathcal{Q} = \mathcal{P}$, we have $\mathrm{dis}(\mathcal{P}, \mathcal{Q}) = 0$, and q can be chosen to be uniform over $T = (S, S')$, thus $\|\mathsf{q}\|_2 = \frac{1}{\sqrt{m+n}}$. The bound then coincides with the standard Rademacher complexity bound for a sample of size $m+n$ for the distribution $\mathcal{P}$. More generally, such a bound holds for any two distributions $\mathcal{P}$ and $\mathcal{Q}$ with $\mathrm{dis}(\mathcal{P}, \mathcal{Q}) = 0$.

The learning bound (7) can be straightforwardly upper-bounded by the weighted discrepancy bound of Cortes et al. (2019)[Prop. 5; Eq. (10)], for any p with support S:

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m} \mathsf{p}_i \ell(h(x_i), y_i) + \mathrm{dis}(\widehat{\mathcal{P}}, \mathsf{p}) + 2\mathfrak{R}_n(\ell \circ \mathcal{H}) + \left[\frac{\log \frac{1}{\delta}}{2n}\right]^{\frac{1}{2}}, \tag{8}$$

using the inequality

$$\mathcal{L}(\widehat{\mathcal{P}}, h) \leq \sum_{i=1}^{m} \mathsf{p}_i \ell(h(x_i), y_i) + \mathrm{dis}(\widehat{\mathcal{P}}, \mathsf{p}),$$

which holds for any p, by definition of the discrepancy. Thus, there is a specific choice of the weights in our bound that makes it a lower bound for that of Cortes et al. (2019), regardless of how the weights p are chosen in their bound (the inequality holds uniformly over p). Our algorithm seeks the best choice of the weights in our bound, for which our bound is thus guaranteed to be a lower bound for that of Cortes et al. (2019), regardless of how the weights p are chosen in their bound.

The weighted-discrepancy minimization algorithm of Cortes and Mohri (2014) is based on a two-stage minimization of (8) and in that sense is sub-optimal compared to an algorithm seeking to minimize the bound of Theorem 1.

q chosen uniformly α-weighted on S. Let $d = \mathrm{dis}(\mathcal{P}, \mathcal{Q})$, $\widehat{d}$ and $\widehat{d} = \mathrm{dis}(\widehat{\mathcal{Q}}, \widehat{\mathcal{P}})$. Consider the following simple, and in general suboptimal, choice of q as a distribution defined by:

$$\overline{\mathsf{q}} = \frac{\alpha m}{m+n} \qquad \mathsf{q}_i = \begin{cases} \frac{\overline{\mathsf{q}}}{m} = \frac{\alpha}{m+n} & \text{if } i \in [m]; \\ \frac{1-\overline{\mathsf{q}}}{n} = \frac{m(1-\alpha)+n}{(m+n)n} & \text{otherwise,} \end{cases}$$

where $\alpha = \Psi(1-d)$ for some non-decreasing function Ψ with $\Psi(0) = 0$ and $\Psi(1) = 1$. We will compare the right-hand side of the bound of Theorem 1, which we denote by B, with its right-hand side B_0 for q chosen to be uniform over S' corresponding to supervised learning on just S':

$$B_0 = \mathcal{L}(\widehat{\mathcal{P}}, h) + 2\mathfrak{R}_n(\ell \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

In Appendix A.5, we show that under some assumptions, we have $B - B_0 \leq 0$. Thus, even for this sub-optimal choice of $\overline{\mathsf{q}}$, under those assumptions, the guarantee of the theorem is then strictly more favorable than the one for training on S' only, uniformly over $h \in \mathcal{H}$.

3.3 Corollary

Theorem 3 suggests choosing $h \in \mathcal{H}$ and $\mathsf{q} \in \{\mathsf{q}: \|\mathsf{q} - \mathsf{p}^0\|_1 < 1\}$ to minimize the right-hand side of the inequality and seek the best balance between these key terms. This guides the design of our learning algorithms. The following corollary provides a slightly simplified version of Theorem 3 (see Appendix A).

Corollary 4 *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{Q}$ and an i.i.d. sample S' of size n from $\mathcal{P}$, the following holds for all $h \in \mathcal{H}$ and $\mathsf{q} \in \{\mathsf{q}: \|\mathsf{q} - \mathsf{p}^0\|_1 < 1\}$:*

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \mathrm{dis}(\mathsf{p}^0, \mathsf{q}) + 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H})$$
$$+ 8\|\mathsf{q} - \mathsf{p}^0\|_1 + \left[\|\mathsf{q}\|_2 + 2\|\mathsf{q} - \mathsf{p}^0\|_1\right]\left[\sqrt{\log \log_2 \frac{2}{1 - \|\mathsf{q} - \mathsf{p}^0\|_1}} + \sqrt{\frac{\log \frac{2}{\delta}}{2}}\right].$$

4 Best-Effort adaptation algorithms

In this section, we describe new learning algorithms for best-effort adaptation directly benefiting from the theoretical analysis of the previous section.

Optimization problem, BEST and SBEST algorithms. The previous section suggests seeking $h \in \mathcal{H}$ and $\mathsf{q} \in [0,1]^{m+n}$ to minimize the bound of Theorem 3 or that of Corollary 4. To simplify the discussion, we will focus on the algorithm derived from Corollary 4. A similar but finer algorithm consists instead of using directly Theorem 3.

Assume that $\mathcal{H}$ is a subset of a normed vector space and that the Rademacher complexity term can be bounded by an upper bound on the norm squared $\|h\|^2$. Then, using the shorthand $d_i = \mathrm{dis}(\mathcal{P}, \mathcal{Q}) 1_{i \in [m]}$, the optimization problem can be written as:

$$\min_{h \in \mathcal{H}, \mathsf{q} \in [0,1]^{m+n}} \sum_{i=1}^{m+n} \mathsf{q}_i [\ell(h(x_i), y_i) + d_i] + \mathrm{dis}(\mathsf{p}^0, \mathsf{q}) + \lambda_\infty \|\mathsf{q}\|_\infty \|h\|^2$$

$$+ \lambda_1 \|\mathsf{q} - \mathsf{p}^0\|_1 + \lambda_2 \|\mathsf{q}\|_2^2,$$

where λ_1, λ_2 and λ_∞ are non-negative hyperparameters. A natural choice for p^0 in our scenario is the uniform distribution over S', which is the empirical distribution in the absence of any point from a different distribution $\mathcal{Q}$, or an appropriate mixture of the empirical distribution over S and the empirical distribution over S'. We will refer by BEST to an algorithm seeking to minimize this objective. We will also consider a simpler version of our algorithm, SBEST, where we upper-bound $\mathrm{dis}(\mathsf{p}^0, \mathsf{q})$ by $\|\mathsf{q} - \mathsf{p}^0\|_1$, in which case this additional term is subsumed by the existing one with λ_1 factor.

When the loss function ℓ is convex with respect to its first argument, the objective function is convex in h and in q. In particular, $\mathrm{dis}(\mathsf{p}^0, \mathsf{q})$ is a convex function of q as a supremum of convex functions (affine functions in q): $\mathrm{dis}(\mathsf{p}^0, \mathsf{q}) = \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{m+n} (\mathsf{p}_i^0 - \mathsf{q}_i) \ell(h(x_i), y_i) \right\}$. But, the objective function is not jointly convex.

Alternating minimization solution. One method for solving the problem consists of alternating minimization (or block coordinate descent), that is of minimizing the objective over $\mathcal{H}$ for a fixed value of q and next of minimizing with respect to q for a fixed value of h. In general, this method does not benefit from convergence guarantees, although there is a growing body of literature proving guarantees under various assumptions (Grippo and Sciandrone, 2000; Li et al., 2019; Beck, 2015).

DC-programming solution. An alternative solution consists of casting the problem as an instance of DC-programming (difference of convex) by rewriting the objective as a difference. Note that for any non-negative and convex function f and any non-decreasing and convex function Ψ defined over $\mathbb{R}_+$, $\Psi \circ f$ is convex: for all $(x, x') \in \mathcal{X}^2$ and $\alpha \in [0, 1]$,

$$\begin{aligned}(\Psi \circ f)(\alpha x + (1-\alpha)x') &\leq \Psi(\alpha f(x) + (1-\alpha) f(x')) \\ &\leq \alpha (\Psi \circ f)(x) + (1-\alpha)(\Psi \circ f)(x'),\end{aligned}$$

where the first inequality holds by the convexity of f and the non-decreasing property of Ψ and the last one by the convexity of Ψ. In particular, for any non-negative and convex function f, f^2 is convex. Thus, we can rewrite the non-jointly convex terms of the objective as the following DC-decompositions:

$$\begin{aligned}\mathsf{q}_i \ell(h(x_i), y_i) &= \frac{1}{2}\Big[\big[\mathsf{q}_i + u\big]^2 - \big[\mathsf{q}_i^2 + u^2\big]\Big], \\ \|\mathsf{q}\|_\infty \|h\|^2 &= \frac{1}{2}\Big[\big[\|\mathsf{q}\|_\infty + \|h\|^2\big]^2 - \big[\|\mathsf{q}\|_\infty^2 + \|h\|^2\big]\Big],\end{aligned}$$

where $u = \ell(h(x_i), y_i)$. We can then use the DCA algorithm of Tao and An (1998), (see also Tao and An (1997)), which in our differentiable case coincides with the CCCP algorithm of Yuille and Rangarajan (2003), further analyzed by Sriperumbudur et al. (2007). The DCA algorithm guarantees convergence to a critical point.

The global optimum can be found by combining DCA with a branch-and-bound or cutting plane method (Tuy, 1964; Horst and Thoai, 1999; Tao and An, 1997). We also present a solution based on convex optimization in the case of the squared loss with a linear or kernel-based hypothesis set (Appendix A.2).

Discrepancy estimation. Our algorithm requires estimating the discrepancy terms d_i. We discuss our DC-programming solution to this problem in detail in Appendix A.3.

As already pointed, our learning bounds are general and can be used for the analysis of various specific reweighting methods with bounded weights, including discrepancy minimization (Cortes and Mohri, 2014), KMM (Huang et al., 2006), KLIEP (Sugiyama et al., 2007), importance weighting (Cortes et al., 2010), when the weights are bounded, and many others. However, unlike our algorithms, which simultaneously learn the weights and the hypothesis and directly benefit from the learning bounds of the previous section, these algorithms typically consist of two stages and do not exploit the guarantees discussed: in the first stage, they determine some weights q, irrespective of the labeled samples and the empirical loss; in the second stage, they use these weights to learn a hypothesis minimizing the q-weighted empirical loss. Additionally, some methods admit other specific drawbacks. For example, it was shown by Cortes et al. (2010), both theoretically and empirically, that, in general, importance weighting may not succeed. Note also that the method relies only on the ratio of the densities and does not take into account, unlike the discrepancy, the hypothesis set and the loss function.

5 Domain adaptation

The analysis of Section 3 can also be used to derive general discrepancy-based guarantees for domain adaptation, where the learner has access to few or no labeled points from the target domain. In this section, we analyze the case where the input points in S' are unlabeled. Our analysis can be straightforwardly extended to the case where a small fraction of the labels in S' are available. Our theoretical analysis leads to the design of new algorithms for domain adaptation.

5.1 Domain adaptation generalization bounds

For convenience, in this section, we will use a different notation for the weights on S and S': $\mathsf{q} \in [0,1]^m$ for the weights on S, $\mathsf{q}' \in [0,1]^n$ for the weights on S'. The labels of the points in S' appear in the first term of the bound of Theorem 1, the q-weighted empirical loss. Since they are not available, we upper-bound the empirical loss in terms of a p-weighted empirical loss and a discrepancy term:

$$\sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), y_i) + \sum_{i=1}^{n} \mathsf{q}'_i \ell(h(x_{m+i}), y_{m+i}) \leq \sum_{i=1}^{m} (\mathsf{q}_i + \mathsf{p}_i) \ell(h(x_i), y_i) + \mathrm{dis}(\mathsf{q}', \mathsf{p}), \quad (9)$$

for any weight vector $\mathsf{p} \in [0,1]^m$. This yields immediately the following theorem.

Theorem 5 *Fix the vectors* q *in* $[0,1]^{[m]}$ *and* $\mathsf{q}' \in [0,1]^n$. *Then, for any* $\delta > 0$, *with probability at least* $1-\delta$ *over the draw of an i.i.d. sample* S *of size* m *from* $\mathcal{Q}$ *and an i.i.d. sample* S' *of size* n *from* $\mathcal{P}$, *the following holds for all* p *in* $[0,1]^{[m]}$ *and* $h \in \mathcal{H}$:

$$\mathcal{L}(\mathcal{P},h) \leq \sum_{i=1}^{m}(\mathsf{q}_i+\mathsf{p}_i)\ell(h(x_i),y_i) + \mathrm{dis}(\mathsf{q}',\mathsf{p}) + \mathrm{dis}\Big(\big[1-\|\mathsf{q}'\|_1\big]\mathcal{P}, \|\mathsf{q}\|_1\mathcal{Q}\Big) + 2\mathfrak{R}_{(\mathsf{q},\mathsf{q}')}(\ell\circ\mathcal{H}) + \sqrt{\frac{\big(\|\mathsf{q}\|_2^2+\|\mathsf{q}'\|_2^2\big)\log\frac{1}{\delta}}{2}}.$$

Let (q,q') denote the vector in $[0,1]^{m+n}$ formed by appending q' to q. The learning bound of Theorem 5 can be extended to hold uniformly over all p in $[0,1]^{[m]}$ and (q,q') in $\big\{(\mathsf{q},\mathsf{q}') \in [0,1]^m \times [0,1]^n \colon \|(\mathsf{q},\mathsf{q}') - \mathsf{p}^0\|_1 < 1\big\}$, where p^0 is a reference (or ideal) reweighting choice over the $(m+n)$ points.

Theorem 6 *For any* $\delta > 0$, *with probability at least* $1-\delta$ *over the drawn of an i.i.d. sample* S *of size* m *from* $\mathcal{Q}$ *and an i.i.d. sample* S' *of size* n *from* $\mathcal{P}$, *the following holds for all* $h \in \mathcal{H}$, $\mathsf{q} \in \big\{\mathsf{q} \colon \|(\mathsf{q},\mathsf{q}') - \mathsf{p}^0\|_1 < 1\big\}$ *and all* $\mathsf{p} \in [0,1]^m$:

$$\begin{aligned}\mathcal{L}(\mathcal{P},h) \leq{}& \sum_{i=1}^{m}(\mathsf{q}_i+\mathsf{p}_i)\ell(h(x_i),y_i) + \mathrm{dis}(\mathsf{q}',\mathsf{p}) \\ &+ \mathrm{dis}\Big(\big[1-\|\mathsf{q}'\|_1\big]\mathcal{P}, \|\mathsf{q}\|_1\mathcal{Q}\Big) \\ &+ \mathrm{dis}(\mathsf{p}^0,(\mathsf{q},\mathsf{q}')) + 2\mathfrak{R}_{(\mathsf{q},\mathsf{q}')}(\ell\circ\mathcal{H}) + 7\|(\mathsf{q},\mathsf{q}')-\mathsf{p}^0\|_1 \\ &+ \Big[\|\mathsf{q}\|_2 + 2\|(\mathsf{q},\mathsf{q}')-\mathsf{p}^0\|_1\Big]\left[\sqrt{\log\log_2\frac{2}{1-\|(\mathsf{q},\mathsf{q}')-\mathsf{p}^0\|_1}} + \sqrt{\frac{\log\frac{2}{\delta}}{2}}\right].\end{aligned}$$

Proof The proof follows immediately by applying inequality (9), which holds for all $\mathsf{p} \in [0,1]^m$, to the bound of Theorem 3. □

Corollary 7 *For any* $\delta > 0$, *with probability at least* $1-\delta$ *over the draw of an i.i.d. sample* S *of size* m *from* $\mathcal{Q}$ *and an i.i.d. sample* S' *of size* n *from* $\mathcal{P}$, *the following holds for all* $h \in \mathcal{H}$, $\mathsf{q} \in \big\{\mathsf{q} \colon \|(\mathsf{q},\mathsf{q}') - \mathsf{p}^0\|_1 < 1\big\}$ *and all* $\mathsf{p} \in [0,1]^m$:

$$\begin{aligned}\mathcal{L}(\mathcal{P},h) \leq{}& \sum_{i=1}^{m}(\mathsf{q}_i+\mathsf{p}_i)\ell(h(x_i),y_i) + \mathrm{dis}(\mathsf{q}',\mathsf{p}) \\ &+ \|\mathsf{q}\|_1\mathrm{dis}(\mathcal{P},\mathcal{Q}) \\ &+ \mathrm{dis}(\mathsf{p}^0,(\mathsf{q},\mathsf{q}')) + 2\mathfrak{R}_{(\mathsf{q},\mathsf{q}')}(\ell\circ\mathcal{H}) + 8\|(\mathsf{q},\mathsf{q}')-\mathsf{p}^0\|_1 \\ &+ \Big[\|\mathsf{q}\|_2 + 2\|(\mathsf{q},\mathsf{q}')-\mathsf{p}^0\|_1\Big]\left[\sqrt{\log\log_2\frac{2}{1-\|(\mathsf{q},\mathsf{q}')-\mathsf{p}^0\|_1}} + \sqrt{\frac{\log\frac{2}{\delta}}{2}}\right].\end{aligned}$$

Proof The result follows Theorem 6 and the application of the upper bound used in the proof of Corollary 1. □

Note that, here, both p and q' can be chosen to make the weighted-discrepancy term $\mathrm{dis}(\mathsf{q}', \mathsf{p})$ smaller. Several of the comments on Theorem 1 similarly apply here. In particular, it is worth pointing out that the learning bound of Cortes et al. (2019) can be recovered for a specific choice of the weights. This holds even in the special case where $\mathsf{q} = 0$ and where q' is a distribution:

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m} \mathsf{p}_i \ell(h(x_i), y_i) + \mathrm{dis}(\mathsf{q}', \mathsf{p}) + 2\mathfrak{R}_{\mathsf{q}'}(\ell \circ \mathcal{H}) + \|\mathsf{q}'\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{2}}.$$

In that case, choosing q' to be the empirical distribution on S' leads to the bound of Cortes et al. (2019) (see also inequality (8), in Appendix 3.2). An alternative choice of the weights may lead to a smaller discrepancy term $\mathrm{dis}(\mathsf{q}', \mathsf{p})$ and a better guarantee overall. Our learning algorithm will seek an optimal choice for the weights.

The discrepancy quantities appearing in the bound of the theorem cannot be estimated in the absence of labels for S'. Thus, we need to resort to upper-bounds expressed in terms of unlabeled discrepancies, using only unlabeled data from $\mathcal{P}$. A detailed analysis is presented in Section 5.3.

5.2 Domain adaptation BEST-DA algorithm

The analysis of the previous section suggests seeking $h \in \mathcal{H}$, q and p in $[0,1]^m$ and q' in $[0,1]^n$ to minimize the bound of Theorem 6 or that of Corollary 7. As in Section 4, assume that $\mathcal{H}$ is a subset of a normed vector space and that the Rademacher complexity term can be bounded in terms of an upper bound on the norm squared $\|h\|^2$. Then, the optimization problem corresponding to Corollary 7 can be written as follows:

$$\min_{\substack{h \in \mathcal{H}, \mathsf{q}, \mathsf{p} \in [0,1]^m \\ \mathsf{q}' \in [0,1]^n}} \sum_{i=1}^{m} (\mathsf{q}_i + \mathsf{p}_i)\, \ell(h(x_i), y_i) + \|\mathsf{q}\|_1 \overline{d} + \overline{\mathrm{dis}}(\mathsf{q}', \mathsf{p}) + \overline{\mathrm{dis}}(\mathsf{p}^0, (\mathsf{q}, \mathsf{q}')) \quad (10)$$
$$+ \lambda_\infty \|(\mathsf{q}, \mathsf{q}')\|_\infty \|h\|^2 + \lambda_1 \|(\mathsf{q}, \mathsf{q}') - \mathsf{p}^0\|_1 + \lambda_2 (\|\mathsf{q}\|_2^2 + \|\mathsf{q}'\|_2^2),$$

where λ_1, λ_2 and λ_∞ are non-negative hyperparameters and where we used the shorthand $\overline{d} = \overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q})$. We are omitting subscripts to simplify the presentation but, as discussed in the previous section, the unlabeled discrepancies in the optimization problem may be local unlabeled discrepancies, which are finer quantities. As in the best-effort adaptation, a natural choice for p^0 in the domain adaptation scenario is the uniform distribution over the input points of S', or an appropriate mixture of the empirical distribution over S and the empirical distribution over S'. In practice, specific applications may motivate various choices.

We will refer by BEST-DA to the algorithm seeking to minimize this objective. Our comments and analysis of the BEST optimization (Section 4) apply similarly here. In particular, the problem can be similarly cast as a DC-programming problem

or a convex optimization problem. The unlabeled discrepancy term $\overline{d} = \overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q})$ can be accurately estimated from $\overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q})$. In Appendix B.3, we show in detail how to compute $\overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q})$ and how to evaluate the sub-gradients of the weighted discrepancy terms.

Discussion of new BEST-DA algorithm

Our BEST-DA algorithm benefits from more favorable guarantees than previous discrepancy-based algorithms (Mansour et al., 2009a; Cortes and Mohri, 2014; Cortes et al., 2019) and algorithms seeking to minimize the learning bound (8), with the unlabeled discrepancy upper bounded by the label discrepancy. This is because, as already pointed out, BEST-DA is based on a learning guarantee that admits as a special case (8). Thus, the best choice of the weights and predictor sought by the algorithm include those corresponding to previous algorithms as a special case.

Moreover, as discussed in Section 3, our upper bounds in terms of local discrepancy are finer than those used in previous work. In particular, BEST-DA improves upon the DM algorithm (*discrepancy minimization*) of Cortes and Mohri (2014), which has been shown empirically by the authors to outperform other domain adaptation baselines in regression tasks. DM seeks to minimize (8) via a two-stage method, by first seeking weights that minimize the unlabeled weighted-discrepancy (second term) and subsequently seeking $h \in \mathcal{H}$ to minimize the empirical loss for that fixed choice of q. This two-stage method may be suboptimal, compared to an algorithm seeking to directly minimize the bound to find (h, q). The solution q found to minimize the discrepancy term in the first stage may, for example, assign significantly larger weights to some sample points, which could lead to a poor choice of the predictor in the second stage.

An alternative sophisticated technique based on the so-called *generalized discrepancy* is advocated by Cortes et al. (2019). The main benefit of this technique is to allow for the weights to be chosen as a function of the hypotheses, unlike the two-stage DM solution of Cortes and Mohri (2014). Our BEST-DA algorithm, however, already offers that advantage since the hypothesis h and the weights q, q' and p are sought simultaneously as a solution of the optimization problem. Note, however that the choice of the weights in the generalized discrepancy method does not take into consideration the empirical losses, unlike our algorithm. Furthermore, BEST-DA minimizes a learning bound admitting as a special case (8), the best learning guarantee presented by the authors in support of their algorithm. Let us add that authors state that their guarantee for the generalized discrepancy method is not comparable to that of DM algorithm.

5.3 Labeled discrepancy upper bounds

The analysis of Section 3 is based on the labeled discrepancy measure $\mathrm{dis}(\mathcal{P}, \mathcal{Q})$ or its estimate from finite samples $\mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$, which assumes access to labeled data from the target distribution $\mathcal{P}$. In typical domain adaptation problems, however, there is little labeled data or none from the target domain $\mathcal{P}$. Thus, instead we need to resort to an upper-bound on $\mathrm{dis}(\mathcal{P}, \mathcal{Q})$ in terms of the unlabeled discrepancy, which only uses unlabeled data from $\mathcal{P}$.

We will discuss two types of upper bounds, first in the special case of the squared loss, next in the case of an arbitrary μ-Lipschitz loss. Our analysis benefits from that of previous work (Cortes and Mohri, 2014; Cortes et al., 2019) but improves upon that, as discussed later.

Squared loss. Here, we give an upper bound on the labeled discrepancy in the case of the squared loss. For any hypothesis $h_0 \in \mathcal{H}$, we denote by $\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$ the *squared-loss label discrepancy* of $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{Q}}$:

$$\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) = \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[h(x)(y - h_0(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[h(x)(y - h_0(x))] \right|. \quad (11)$$

Lemma 8 *Let ℓ be the squared loss. Then, for any hypothesis h_0 in $\mathcal{H}$, the following upper bound holds for the labeled discrepancy:*

$$\mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) \leq \overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) + 2\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}).$$

The proof is given below in Appendix B.1. The local unlabeled discrepancy $\overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$ captures the closeness of the input distributions $\widehat{\mathcal{P}}_X$ and $\widehat{\mathcal{Q}}_X$. It is a significantly more favorable term that the standard unlabeled discrepancy since it admits only a maximum over $h \in \mathcal{H}$ and not over both h and h' in $\mathcal{H}$.

For a suitable choice of $h_0 \in \mathcal{H}$, the term $\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$ captures the closeness of the empirical output labels on $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{Q}}$. Note that for $\widehat{\mathcal{P}} = \widehat{\mathcal{Q}}$, we have $\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) = 0$ for any $h_0 \in \mathcal{H}$. When the covariate-shift assumption holds and the problem is separable, h_0 can be chosen so that $\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) = 0$. More generally, when h_0 can be chosen so that $|y - h_0(x)|$ is relatively small on both samples corresponding to $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{Q}}$ and the hypotheses $h \in \mathcal{H}$ are bounded by some $M > 0$, then $\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$ is relatively small. Note that adaptation is in general not possible if the learner receives vastly different labels on the source domain $\mathcal{Q}$ than those corresponding to the target $\mathcal{P}$.

μ-Lipschitz loss. Here, we give an upper bound on the labeled discrepancy for any μ-Lipschitz loss. For any hypothesis $h_0 \in \mathcal{H}$, we denote by $\eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$ the *Lipschitz loss labeled discrepancy* defined by

$$\eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) = \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[|y - h_o(x)|] + \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[|y - h_o(x)|]. \quad (12)$$

Lemma 9 *Let ℓ be a loss function that is μ-Lipschitz with respect to its second argument. Then, for any hypothesis h_0 in $\mathcal{H}$, the following upper bound holds for the labeled discrepancy:*

$$\mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) \leq \overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) + \mu\, \eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}).$$

The proof is given below in Appendix B.2.

The Lipschitz loss labeled discrepancy $\eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})$ is a coarser quantity than $\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})$. In particular, even when $\widehat{\mathcal{P}}=\widehat{\mathcal{Q}}$, $\eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})$ is not zero. However, as with $\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})$ it captures the closeness of the output labels on $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{Q}}$. When h_0 can be chosen so that the sum of expected values $|y-h_0(x)|$ is relatively small on both samples corresponding to $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{Q}}$ then, $\eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})$ is relatively small. As already pointed out, adaptation is not possible when the learner received very different labels on the two domains.

The upper bounds of Lemmas 8 and 9 hold in the stochastic setting and are thus more general than those derived for the deterministic label setting in previous work (Cortes and Mohri, 2014; Cortes et al., 2019). They are also finer bounds expressed in terms of the more favorable local discrepancy and somewhat more favorable label discrepancy terms defined in terms of expectation over the empirical distributions as opposed to a supremum.

In both the squared loss and Lipschitz cases, when a relatively small labeled sample S' drawn i.i.d. from $\mathcal{P}$ is available, we can use it to select h_0 via

$$h_0 = \operatorname*{argmin}_{h_0\in\mathcal{H}} \delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}_{S'},\widehat{\mathcal{Q}}) \text{ or } h_0 = \operatorname*{argmin}_{h_0\in\mathcal{H}} \eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}_{S'},\widehat{\mathcal{Q}}).$$

When no labeled data from the target domain is at our disposal, we cannot choose h_0 by leveraging any existing information. We can then assume that $\min_{h_0\in\mathcal{H}} \delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) \ll 1$ in the squared loss case or $\min_{h_0\in\mathcal{H}} \eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) \ll 1$ in the Lipschitz case, that is that the source labels are relatively close to the target ones based on these measures and use the standard unlabeled discrepancy:

$$\mathrm{dis}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) \le \overline{\mathrm{dis}}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) + 2\min_{h_0\in\mathcal{H}} \delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})$$
$$\mathrm{dis}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) \le \overline{\mathrm{dis}}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) + \mu\min_{h_0\in\mathcal{H}} \eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}).$$

6 Experimental evaluation

We evaluated our algorithms in best-effort adaptation, fine-tuning, and (unsupervised) domain adaptation. We performed cross-validation using labeled data from the target to pick the hyperparameters for our algorithms and the baselines. See Appendix C for details on data and experimental procedures. For all the experiments we use the SBEST algorithm.

6.1 Best-Effort adaptation

Here, we have labeled data both from the source and the target domains. Two natural baselines are to train solely on $\mathcal{P}$, or solely $\mathcal{Q}$. A third baseline is the α-reweighted q as described in Appendix 3.2.

We consider a linear binary classification task with the labels for $\mathcal{P}$ generated as $\mathrm{sgn}(w_p\cdot x)$ for a randomly chosen unit vector w_p. The distribution $\mathcal{Q}$ admits two parts. For $\eta\in(0.5,1)$, $(1-\eta)m$ examples are labeled according to $\mathrm{sgn}(w_q\cdot x)$ where $\|w_p-w_q\|\le\epsilon$, while the remaining examples are set to a fixed vector u and labeled +1.

Table 1 Performance of SBEST, compared to baseline approaches on UCI/Newsgroups classification tasks. Best results are marked in boldface, ties in italics.

Dataset	Train source $\mathcal{Q}$	Train target $\mathcal{P}$	KMM	gapBoost	SBEST
Adult	82.72 ± 0.10	81.61 ± 0.42	81.24 ± 0.01	*83.1 ± 0.02*	*83.30 ± 0.28*
German	68.24 ± 0.21	69.87 ± 0.27	65.7 ± 0.01	69.8 ± 0.03	$\mathbf{71.26 \pm 0.11}$
Accent	27.20 ± 0.26	81.64 ± 0.22	53.1 ± 0.03	81.2 ± 0.04	$\mathbf{84.15 \pm 0.30}$
comp vs sci	83.2 ± 0.004	89.4 ± 0.03	83.1 ± 0.004	92.08 ± 0.01	$\mathbf{94.4 \pm 0.01}$
rec vs sci	79.2 ± 0.007	91.3 ± 0.02	79.7 ± 0.004	92.2 ± 0.01	$\mathbf{92.4 \pm 0.004}$
comp vs talk	71.4 ± 0.002	89.9 ± 0.02	71 ± 0.006	90.6 ± 0.01	$\mathbf{91 \pm 0.02}$
comp vs rec	65.4 ± 0.007	85.2 ± 0.01	67.7 ± 0.007	85.9 ± 0.01	$\mathbf{88 \pm 0.01}$
rec vs talk	81.3 ± 0.004	88 ± 0.02	81.2 ± 0.005	89.2 ± 0.01	$\mathbf{92.3 \pm 0.03}$
sci vs talk	88.2 ± 0.005	93.3 ± 0.008	88.5 ± 0.003	*94.6 ± 0.01*	*94.6 ± 0.02*

These ηm examples represent the noise in $\mathcal{Q}$ and, as η increases, $\mathrm{dis}(\mathcal{P}, \mathcal{Q})$ gets larger. For this setting, we evaluated the baselines and SBEST with the logistic loss and linear hypotheses. In line with the previously mentioned setting, all experiments involving linear hypothesis sets in this work are conducted without incorporating a bias term. For further elaboration and illustrative examples, please consult Appendix C.

Simulated data. The goal of this experiment was to demonstrate that SBEST outperforms the simple baselines just mentioned and to compare the performance of the Alternate Minimization (SBEST-AM) and the DC-programming (SBEST-DC) optimization solutions.

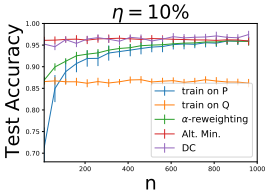

Fig. 1 Simulated data.

Figure 1 shows the performance for $\eta = 10\%$ as n increases. For small sizes, n, of the target data $\mathcal{P}$, both α-reweighting and the baseline that trains solely on $\mathcal{Q}$ are significantly impacted. This is because these methods cannot distinguish between non-noisy and noisy data points. On the other hand, both SBEST-AM and SBEST-DC can counter the effect of the noise by generating q-weights that are predominantly supported on the non-noisy samples. The performance of these algorithms is fairly independent of the size of n as, for $\eta = 10\%$, they can still make an effective use of 90% of the $m = 1000$ examples. As n increases, α-reweighting and the baseline that trains solely on $\mathcal{P}$ reach the performance of SBEST. We also note that SBEST-AM and SBEST-DC perform equivalently and in all the following experiments, we use SBEST-AM. For experiments with other values of η and further discussion of this experiment, see Appendix C.

Real-world data. Baselines. We compare SBEST with the popular Kernel Mean Matching (KMM) algorithm (Huang et al., 2006) and also the recently proposed gapBoost algorithm (Wang et al., 2019). The gapBoost algorithm constructs an ensemble of classifiers. In round t, the algorithm maintains a distribution q over the entire data. It then trains a classifier h_t using q-weighted loss minimization as well as classifiers $h_{t,\mathcal{Q}}$ and $h_{t,\mathcal{P}}$ trained on the source and the target data respectively, again using weighted loss minimization. It then uses the disagreement among the three classifiers to update the weights for the next round. Finally, it outputs a weighted

Table 2 Performance of SBEST, compared to baseline approaches in CIFAR-10.

Fine-tuning	Train on $\mathcal{P}$	gapBoost	SBEST
Last layer (CIFAR-10)	$88.61 \pm .43$	$87.1 \pm .01$	$\mathbf{89.62 \pm .32}$
Full model (CIFAR-10)	$90.18 \pm .31$	$90.8 \pm .02$	$\mathbf{92.30 \pm .24}$
Last layer (Civil)	$63.1 \pm .12$	$64.7 \pm .11$	$\mathbf{65.8 \pm .12}$
Full model (Civil)	$65.8 \pm .01$	$67.2 \pm .01$	$\mathbf{68.3 \pm .14}$

combination of the classifiers h_t.

Real-world data. Classification. We used three datasets from the UCI machine learning repository (Dua and Graff, 2017): the Adult-Income, German-Credit, and Speaker Accent Recognition. In addition, we used six adaptation tasks derived from the Newsgroups dataset, as considered in prior work (Wang et al., 2019). For the definition of $\mathcal{Q}$ and $\mathcal{P}$, and other experimental parameters, see Appendix C. The results are reported in Table 1. The KMM algorithm does not make use of labels for matching distributions, and is naturally outperformed by SBEST, and so is gapBoost.

Real-world data. Regression. We also carried out experiments on five regression datasets from the UCI repository (Dua and Graff, 2017) and compared against baselines KMM and the DM algorithm (Cortes and Mohri, 2014). We did not compare with gapBoost, since the algorithm was designed only for classification Wang et al. (2019). See Appendix C for similarly strong results in this setting.

6.2 Fine-tuning tasks

Here, we applied our algorithms to fine-tuning pre-trained models in classification. In the pre-training/fine-tuning paradigm (Raffel et al., 2020), a model is pre-trained on a generalist dataset (coming from $\mathcal{Q}$). The model is then fine-tuned on a task-specific dataset (generated from $\mathcal{P}$). Two predominantly used fine-tuning approaches are *last-layer fine-tuning* (Subramanian et al., 2018; Kiros et al., 2015) and *full-model fine-tuning* (Howard and Ruder, 2018). In the former, the representations obtained from the last layer of the pre-trained model are used to train a simple model (often a linear hypothesis) on the data from $\mathcal{P}$. We chose the simple model to be a multi-class logistic regression model. In the latter approach, the model is initialized from the pre-trained model and all the parameters are fine-tuned (often via gradient descent) on $\mathcal{P}$. We explored the additional advantages of combining data from both $\mathcal{P}$ and $\mathcal{Q}$ during fine-tuning. There has been recent interest in carefully combining various tasks/data for the purpose of fine-tuning and avoid the phenomenon of "negative transfer" (Aribandi et al., 2021). Our proposed theory presents a principled approach.

We used the CIFAR-10 vision dataset (Krizhevsky et al., 2009) and formed a pre-training task (source) by combining data from classes: {'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog'}. For this task we use a standard ResNet-18 architecture (He et al., 2016). The fine-tuning task (target) consists of data from classes: {'frog', 'horse', 'ship', 'truck'}. In addition, we also used the Civil Comments dataset. For

Table 3 Relative MSE achieved by GDM, DM and KMM against a normalized MSE of 1.0 obtained via BEST-DA on various adaptation tasks. For reference we also report the relative MSE achieved by training only on the source $\mathcal{Q}$.

$\mathcal{Q}$	$\mathcal{P}$	GDM	DM	KMM	TRAIN ON $\mathcal{Q}$
BOOKS	**DVD**	1.25 ± 0.01	1.26 ± 0.11	1.43 ± 0.08	2.34 ± 0.19
	ELEC	0.88 ± 0.01	0.89 ± 0.03	1.50 ± 0.05	2.13 ± 0.13
	KTCHN	1.06 ± 0.03	1.08 ± 0.04	1.47 ± 0.01	1.55 ± 0.01
DVD	**BOOKS**	1.14 ± 0.02	1.17 ± 0.10	1.64 ± 0.14	2.18 ± 0.18
	ELEC	1.08 ± 0.01	1.10 ± 0.12	2.40 ± 0.05	3.26 ± 0.07
	KTCHN	1.1 ± 0.03	1.12 ± 0.02	1.10 ± 0.02	2.34 ± 0.05
ELEC	BOOKS	0.98 ± 0.01	1.00 ± 0.01	1.33 ± 0.06	1.34 ± 0.04
	DVD	0.98 ± 0.02	1.00 ± 0.06	1.00 ± 0.06	1.04 ± 0.08
	KTCHN	0.96 ± 0.01	0.98 ± 0.06	1.04 ± 0.01	1.14 ± 0.01
KTCHN	**BOOKS**	1.00 ± 0.03	1.04 ± 0.07	1.27 ± 0.09	1.12 ± 0.08
	DVD	1.2 ± 0.002	1.33 ± 0.03	1.32 ± 0.03	1.42 ± 0.04
	ELEC	1.64 ± 0.02	1.67 ± 0.54	1.87 ± 0.56	1.89 ± 0.56

this we used a BERT-small model (Devlin et al., 2019) for pre-training. For more detail on the dataset and experimental procedure, see Appendix C. As can be seen from Table 2, SBEST comfortably outperforms both the standard approach of training just on $\mathcal{P}$, as well as gapBoost.

6.3 Domain adaptation

We next evaluated our proposed BEST-DA algorithm in the domain adaptation setting. No labeled target data was used by our algorithm or other baselines for training. However, we used a small labeled validation set of size 50 to determine the parameters for all the algorithms. This is consistent with experimental results reported in prior work (e.g., (Cortes and Mohri, 2014)).

We used the multi-domain sentiment analysis dataset of (Blitzer et al., 2007) that has been used in prior work on domain adaptation (Cortes and Mohri, 2014; Cortes et al., 2019) for the regression setting. The dataset consists of text reviews associated with a star rating from 1 to 5 for different categories. We considered four categories namely BOOKS, DVD, ELECTRONICS, and KITCHEN. Our methodology is inspired by prior work (Mohri and Muñoz Medina, 2012; Cortes and Mohri, 2014) with certain simplifications, see Appendix C for details.

For each category, we formed a regression task by converting the review text to a 128-dimensional vector and fitting a linear regression model to predict the rating. The predictions of the model are then defined as the ground truth regression labels. We then formed adaptation problems for each pair of distinct tasks: (TaskA, TaskB) where TaskA, TaskB are in {BOOKS, DVD, ELECTRONICS, KITCHEN}. In each case, we formed the source domain ($\mathcal{Q}$) by taking 500 labeled samples from TaskA and 200 labeled examples from TaskB. The target ($\mathcal{P}$) was formed by taking 300 unlabeled examples from TaskB. This led to 12 adaptation problems with varying levels of difficulty.

We compared with KMM and the discrepancy minimization algorithms (GDM) (Cortes et al., 2019) and DM (Cortes and Mohri, 2014). We report in Table 3 the results averaged over 10 independent source/target splits, where we normalized the error (MSE) of BEST-DA to be 1.0 and presented the relative MSE achieved by the other methods. In all but one adaptation category (`elec`), BEST-DA outperforms or ties with existing methods (boldface). GDM is considered the state-of-the-art and does indeed outperform DM in our experiments. Appendix C contains additional experimental details as well as experiments for domain adaptation in the covariate-shift setting.

7 Related work

7.1 Adaptation and transfer learning

Discrepancy-based adaptation theory. The work we present includes a significant theoretical component and benefits from prior theoretical analyses of domain adaptation. The theoretical analysis of domain adaptation was initiated by Kifer et al. (2004) and Ben-David et al. (2006) with the introduction of a d_A*-distance* between distributions. They used this notion to derive VC-dimension learning bounds for the zero-one loss, which was elaborated on in follow-up publications like (Blitzer et al., 2008; Ben-David et al., 2010). Later, Mansour et al. (2009a) and Cortes and Mohri (2011, 2014) presented a general analysis of single-source adaptation for arbitrary loss functions, where they introduced the notion of *discrepancy*, which they argued is a divergence measure tailored to domain adaptation. The notion of discrepancy coincides with the d_A-distance in the special case of the zero-one loss. It takes into account the loss function and the hypothesis set and, importantly, can be estimated from finite samples. The authors further gave Rademacher complexity learning bounds in terms of the discrepancy for arbitrary hypothesis sets and loss functions, as well as pointwise learning bounds for kernel-based hypothesis sets. They also gave a discrepancy minimization algorithm based on a reweighting of the losses of sample points. We use their notion of discrepancy in our new analysis. Cortes et al. (2019) presented an extension of the discrepancy minimization algorithm based on the so-called *generalized discrepancy*, which allows for the weights to be hypothesis-dependent and which works with a less conservative notion of *local discrepancy* defined by a supremum over a subset of the hypothesis set. The notion of local discrepancy has been since adopted in several recent publications, in the study of active learning or adaptation (de Mathelin et al., 2022; Zhang et al., 2019c, 2020) and is also used in part of our analysis. Finally, a PAC-Bayesian analysis of adaptation has also been given by Germain et al. (2013), using a related notion of discrepancy. Note also that, as argued in Appendix A.3, for our analysis of best-effort adaptation and algorithms, we can restrict ourselves to a small ball $\mathsf{B}(h_{\mathcal{P}}, r)$ around the best hypothesis found by training on $\mathcal{P}$, with r in the order of $1/\sqrt{n}$. This leads to a more favorable discrepancy term, which is similar to the *super transfer* or *localization* benefits mentioned by Hanneke and Kpotufe (2019). This advantage can be leveraged when there is a sufficient amount of labeled data from the target distribution, as in the scenario of

best-effort adaptation. In standard domain adaptation, however, it would not be possible to estimate such local discrepancy quantities, which are also used in the analysis of Zhang et al. (2020), and thus the corresponding learning bounds or notions would be not be algorithmically useful.

A theoretical analysis and algorithm for driting distributions are given by Mohri and Muñoz Medina (2012). The assumptions made in the analysis of adaptation were discussed by Ben-David et al. (2010) who presented several negative results for the zero-one loss.

Many of the theoretical guarantees for domain adaptation (Ben-David et al., 2006; Ben-David et al., 2010; Zhang et al., 2019a) have upper bounds that include the term $\lambda_{\mathcal{H}} = \min_{h \in \mathcal{H}} \{\mathcal{L}(\mathcal{P}, h) + \mathcal{L}(\mathcal{Q}, h)\}$, which, as pointed out by Mansour et al. (2009a), roughly doubles the representation error one incurs for $\mathcal{H}$ and results overall in learning bounds with a factor of 3 of the error with the respect to an ideal target. This can make these bounds vacuous in some natural scenarios. Moreover, the $\lambda_{\mathcal{H}}$ terms cannot be estimated from observations. The learning bounds of Mansour et al. (2009a) do not admit the factor of 3 of the error drawback, but they also contain terms depending on the best-in-class predictors with respect to both distributions that cannot be estimated. In general, they are not comparable with the bounds of Ben-David et al. (2006). Our learning bounds differ from these analyses since we compare the target loss of a predictor with an empirical q-weighted empirical loss on a sample from $\mathcal{Q}$ or both $\mathcal{Q}$ and $\mathcal{P}$ and not just with an unweighted loss for a sample drawn from $\mathcal{Q}$. Furthermore, our learning guarantees are high-probability bounds, while those of these previous work hold with probability one. The latter can be derived from straightforward applications of triangle inequality. Crucially, our learning bounds can be leveraged by algorithms, while previous bounds do not include any non-trivial term that can be optimized.

Multiple-source adaptation theory. Mansour et al. (2021) presented a theory of multiple-source adaptation with limited target labeled data using the notion of discrepancy. A series of publications by Mansour et al. (2009a,b), Hoffman et al. (2018, 2021, 2022) and Cortes et al. (2021) give an extensive theoretical and algorithmic analysis of the problem of *multiple-source adaptation* (MSA) scenario where the learner has access to unlabeled samples and a trained predictor for each source domain, with no access to source labeled data. This approach has been further used in many applications such as object recognition (Hoffman et al., 2012; Gong et al., 2013a,b). Zhao et al. (2018) and Wen et al. (2020) considered MSA with only unlabeled target data available and provided generalization bounds for classification and regression.

Other adaptation analyses. There are alternative analyses of the adaptation problem based on divergences between distributions that do not take into account the specific loss function or hypothesis set used. These include methods based on importance weighting (Sugiyama et al., 2007; Zhang et al., 2020; Lu et al., 2021; Sugiyama et al., 2007). Cortes et al. (2010) gave a theoretical analysis of importance weighting, including learning bounds based on the analysis of unbounded loss functions (see also (Cortes et al., 2019)), showing both theoretically and empirically that importance

weighting can fail in a number of cases, depending on the magnitude of the second-moment of the weights, including in simple cases of the two domain being Gaussian distributions. This holds even for perfectly estimated importance weights. The publications in this category also include those using the Wasserstein distance (Courty et al., 2017; Redko et al., 2017), which in some sense is closer to the notion of discrepancy but yet does not capture the hypothesis set used. An alternative distance used is that of Kernel Mean Matching (KMM), which is the difference between the expectation of the feature vector in the source domain and the target domain (Huang et al., 2006). Several other publications have also adopted also that distance (Long et al., 2015; Redko and Bennani, 2016). The KMM algorithm seeks to reweight the source sample to make this difference as small as possible. This, however, ignores other moments of the distributions, as well as the loss function and the hypothesis sets. Nevertheless, in some instances, the distance is close to and somewhat related to discrepancy. The experiments reported by Cortes and Mohri (2014) suggest that, while in some instances KMM performs well, in some others it does not. This variance might be due to the fact that the distance does not always capture key aspects related to the loss function and the hypothesis set. In other experiments reported by Cortes et al. (2019), the performance of KMM is sometimes worse than training on the sample S drawn from $\mathcal{Q}$ (without reweighting). This problem was already reported for another algorithm, KLIEP, by Sugiyama et al. (2007). Variants of boosting designed for transfer also tacitly reweight examples (Huang et al., 2017; Zheng et al., 2020).

Note that the algorithms suggested for KMM, importance-weighting, KLIEP and other similar methods can all be viewed as specific methods for reweighting the sample losses. In that sense, they are all covered by our general analysis, when the weights are bounded. However, note also that they are all two-stage algorithms: the weights are first chosen to reduce or minimize some distance, irrespective of their effect on the weighted empirical loss, and next the weights are fixed and used to minimize the empirical weighted loss.

An interesting non-parametric analysis of adaptation is presented in (Kpotufe and Martinet, 2018; Hanneke and Kpotufe, 2019). Hanneke and Kpotufe (2019) do not give an adaptation algorithm, however. A causal view of adaptation is also analyzed in (Zhang et al., 2013; Gong et al., 2016).

Transfer learning analyses. Other scenarios of transfer learning have been studied by Kuzborskij and Orabona (2013); Perrot and Habrard (2015); Du et al. (2017) including by leveraging smaller target labeled data and auxiliary hypotheses (see also (Hanneke and Kpotufe, 2019) already mentioned). The problem of active adaptation or transfer learning has been investigated by several publications Yang et al. (2013); Chattopadhyay et al. (2013); Berlind and Urner (2015). Another somewhat related problem is that of multi-task learning studied by Maurer (2006); Maurer et al. (2016); Pentina and Lampert (2017); Pentina and Ben-David (2018). The scenario of lifelong learning is also somewhat related (Pentina and Lampert, 2014, 2015; Pentina and Urner, 2016; Balcan et al., 2019).

Other adaptation or transfer learning publications. The space of transfer learning and domain adaptation approaches is massive (Chen et al., 2011; Zhang

et al., 2019b; Wang and Mahadevan, 2011; Sener et al., 2016; Hoffman et al., 2012; Ghifary et al., 2016; Zhao et al., 2019, 2018; Li et al., 2018; Bousmalis et al., 2017; Sun et al., 2016; Kundu et al., 2020; Sun and Saenko, 2016; Ghifary et al., 2016; Long et al., 2016; Courty et al., 2016; Saito et al., 2018; Wang et al., 2018; Motiian et al., 2017; Sun and Saenko, 2016) and includes interesting analyses and observations such as that of III (2007) about a surprisingly good baseline and follow-up by Sun et al. (2016). We recommend readers to surveys such as Pan and Yang (2009); Wang and Deng (2018); Li (2012) for a comprehensive overview. We briefly outline the most relevant approaches here.

There is a very large recent literature dealing with experimental studies of domain adaptation in various tasks. Ganin et al. (2016) proposed to learn features that cannot discriminate between source and target domains. Tzeng et al. (2015) proposed a CNN architecture to exploit unlabeled and sparsely labeled target domain data. Motiian et al. (2017), Motiian et al. (2017) and Wang et al. (2019) proposed to train maximally separated features via adversarial learning. Saito et al. (2019) proposed to use a minmax entropy method for domain adaptation.

Several algorithms have been proposed for multiple-source adaptation. Khosla et al. (2012); Blanchard et al. (2011) proposed to combine all the source data and train a single model. Duan et al. (2009, 2012) used unlabeled target data to obtain a regularizer. Domain adaptation via adversarial learning was studied by Pei et al. (2018); Zhao et al. (2018). Crammer et al. (2008) considered learning models for each source domain, using close-by data of other domains. Gong et al. (2012) ranked multiple source domains by how well they can adapt to a target domain. Other solutions to multiple-source domain adaptation include, clustering (Liu et al., 2016), learning domain-invariant features (Gong et al., 2013a), learning intermediate representations (Jhuo et al., 2012), subspace alignment techniques (Fernando et al., 2013), attributes detection (Gan et al., 2016), using a linear combination of pre-trained classifiers (Yang et al., 2007), using multitask auto-encoders (Ghifary et al., 2015), causal approaches (Sun et al., 2011), two-state weighting approaches (Sun et al., 2011), moments alignment techniques (Peng et al., 2019) and domain-invariant component analysis (Muandet et al., 2013).

When some labeled data from both source and target are available, a variety of practical methods have been studied. III (2007) performs an empirical comparison amongst a collection of basic models when some labeled data is available from both source and target: source-only, target-only, training on all data together, uniformly α-weighting the source data and $(1-\alpha)$-weighting the target data, using the prediction of a model on the source as a feature for training on the target, linearly interpolating between source-only and target-only models, and a "lifted" approach where each sample is projected into $\mathcal{X}^3$, corresponding to source/target/general information copies of the feature space, and show empirically that each of these benchmarks performs fairly well, with the latter outperforming the others most of the time.

Some recent work focuses on adversarial adaptation (Motiian et al., 2017; Pei et al., 2018; Ganin et al., 2016). The problem of *domain generalization*, that is generalization to an arbitrary target distribution within some set has been studied by

(Mohri et al., 2019) and is also related to that of robust learning (Chen et al., 2017; Konstantinov and Lampert, 2019; Jhuo et al., 2012).

We discuss separately, in the following section, the relationship of our work with fine-tuning methods.

7.2 Relationship with fine-tuning methods

Here, we discuss the connection of our work with fine-tuning (Howard and Ruder, 2018; Peters et al., 2018; Houlsby et al., 2019) of pre-trained models. A comprehensive description of fine-tuning methods is beyond the scope of this work, but see (Guo et al., 2019; You et al., 2020; Aribandi et al., 2021; Aghajanyan et al., 2021; Wei et al., 2021) for some recent results. A related area is few shot-learning algorithms and related meta-learning algorithms such as MAML (Finn et al., 2017) include (Wang et al., 2019; Motiian et al., 2017), and Reptile (Nichol et al., 2018).

In general, consider a scenario where there exists good common feature mapping $\Phi: \mathcal{X} \to \mathbb{R}^d$ for both the $\mathcal{Q}$ and $\mathcal{P}$. Let f be the result of pre-training a neural network on $\mathcal{Q}$ data. The mapping in f corresponding to some depth of the hidden layers can then be viewed as a good approximation of Φ. Alternatively, Φ may be the output of a representation learning algorithm.

There are several fine-tuning methods introduced in the literature (Subramanian et al., 2018; Kiros et al., 2015; Howard and Ruder, 2018; Raffel et al., 2020) that consists of adapting f to domain $\mathcal{P}$. This may be by using f as an initialization point and applying SGD with sample S' drawn from $\mathcal{P}$, while fixing the hidden layer parameters to a given depth. It may be by *forgetting* the weights at the top layer(s) and retraining them by using S' alone. Or, it may be done by continuing training with a mixture of S' and a new sample from S. Training on such a mixture avoids 'catastrophic forgetting'. In all cases, the problem can be cast as that of learning a hypothesis with feature vector Φ by using sample S and S', or sample S' alone, which is a special case of the scenario we analyzed in Section 3. The algorithms presented in Section 4 provide a principled solution to this problem by taking into consideration the discrepancy between $\mathcal{Q}$ and $\mathcal{P}$ and by selecting suitable q-weights to guarantee a better generalization.

8 Conclusion

We presented a comprehensive study of best-effort adaptation (or supervised adaptation), including a new discrepancy-based theoretical analysis, algorithms benefiting from the corresponding learning guarantees, as well as a series of empirical results showcasing their performance in several tasks. We further demonstrated how our analysis can be leveraged to derive learning guarantees in domain adaptation, as well as new enhanced adaptation algorithms. Our analysis and algorithms are likely to be useful in the study of other adaptation scenarios and admit a variety of other applications. In fact, our analysis applies to any sample reweighting method.

Data availability statement

The datasets analyzed in this study are all public datasets and are available from the URLs referenced. Our artificial dataset used for a simulation is described in detail and the code generating it can be provided upon request.

Declarations

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Aghajanyan, A., A. Gupta, A. Shrivastava, X. Chen, L. Zettlemoyer, and S. Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning.

Aribandi, V., Y. Tay, T. Schuster, J. Rao, H.S. Zheng, S.V. Mehta, H. Zhuang, V.Q. Tran, D. Bahri, J. Ni, J. Gupta, K. Hui, S. Ruder, and D. Metzler. 2021. Ext5: Towards extreme multi-task scaling for transfer learning.

Balcan, M., M. Khodak, and A. Talwalkar 2019. Provable guarantees for gradient-based meta-learning. In *Proceedings of ICML*, Volume 97, pp. 424–433. PMLR.

Bartlett, P.L. and S. Mendelson. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3: 463–482 .

Beck, A. 2015. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM J. Optim. 25*(1): 185–209 .

Ben-David, S., J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. 2010. A theory of learning from different domains. *Machine learning 79*(1-2): 151–175 .

Ben-David, S., J. Blitzer, K. Crammer, and F. Pereira 2006. Analysis of representations for domain adaptation. In *Proceedings of NIPS*, pp. 137–144. MIT Press.

Ben-David, S., T. Lu, T. Luu, and D. Pál. 2010. Impossibility theorems for domain adaptation. *Journal of Machine Learning Research - Proceedings Track* 9: 129–136 .

Berlind, C. and R. Urner 2015. Active nearest neighbors in changing environments. In *Proceedings of ICML*, Volume 37, pp. 1870–1879. JMLR.org.

Blanchard, G., G. Lee, and C. Scott 2011. Generalizing from several related classification tasks to a new unlabeled sample. In *NIPS*, pp. 2178–2186.

Blitzer, J., K. Crammer, A. Kulesza, F. Pereira, and J. Wortman 2008. Learning bounds for domain adaptation. In *Proceedings of NIPS*, pp. 129–136.

Blitzer, J., M. Dredze, and F. Pereira 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, pp. 440–447.

Bousmalis, K., N. Silberman, D. Dohan, D. Erhan, and D. Krishnan 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731.

Boyd, S.P. and L. Vandenberghe. 2014. *Convex Optimization*. Cambridge University Press.

Chattopadhyay, R., W. Fan, I. Davidson, S. Panchanathan, and J. Ye 2013. Joint transfer and batch-mode active learning. In *Proceedings of ICML*, Volume 28, pp. 253–261. JMLR.org.

Chen, M., K.Q. Weinberger, and J. Blitzer 2011. Co-training for domain adaptation. In *Nips*, Volume 24, pp. 2456–2464. Citeseer.

Chen, R.S., B. Lucier, Y. Singer, and V. Syrgkanis 2017. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, pp. 4705–4714.

Cortes, C., S. Greenberg, and M. Mohri. 2019. Relative deviation learning bounds and generalization with unbounded loss functions. *Ann. Math. Artif. Intell. 85*(1): 45–70 .

Cortes, C., Y. Mansour, and M. Mohri 2010. Learning bounds for importance weighting. In *Proceedings of NIPS*, pp. 442–450. Curran Associates, Inc.

Cortes, C. and M. Mohri 2011. Domain adaptation in regression. In *Proceedings of ALT*, pp. 308–323.

Cortes, C. and M. Mohri. 2014. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.* 519: 103–126 .

Cortes, C., M. Mohri, and A. Muñoz Medina. 2019. Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.* 20: 1:1–1:30 .

Cortes, C., M. Mohri, A. Theertha Suresh, and N. Zhang 2021. A discriminative technique for multiple-source adaptation. In M. Meila and T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Volume 139 of *Proceedings of Machine Learning Research*, pp. 2132–2143. PMLR.

Courty, N., R. Flamary, D. Tuia, and A. Rakotomamonjy. 2016. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence 39*(9): 1853–1865 .

Courty, N., R. Flamary, D. Tuia, and A. Rakotomamonjy. 2017. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell. 39*(9): 1853–1865 .

Crammer, K., M.J. Kearns, and J. Wortman. 2008. Learning from multiple sources. *Journal of Machine Learning Research 9*(Aug): 1757–1774 .

de Mathelin, A., F. Deheeger, M. Mougeot, and N. Vayatis 2022. Discrepancy-based active learning for domain adaptation. In *The Tenth International Conference*

on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Devlin, J., M. Chang, K. Lee, and K. Toutanova 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics.

Du, S.S., J. Koushik, A. Singh, and B. Póczos 2017. Hypothesis transfer learning via transformation functions. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 574–584.

Dua, D. and C. Graff. 2017. UCI machine learning repository.

Duan, L., I.W. Tsang, D. Xu, and T. Chua 2009. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, Volume 382, pp. 289–296.

Duan, L., D. Xu, and I.W. Tsang. 2012. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems 23*(3): 504–518 .

Fernandes, K. 2015, 08. A proactive intelligent decision support system for predicting the popularity of online news. In *Springer Science and Business Media LLC‘¡*.

Fernando, B., A. Habrard, M. Sebban, and T. Tuytelaars 2013. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pp. 2960–2967.

Finn, C., P. Abbeel, and S. Levine 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR.

Gan, C., T. Yang, and B. Gong 2016. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 87–97.

Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. 2016. Domain-adversarial training of neural networks.

The Journal of Machine Learning Research 17(1): 2096–2030 .

Garcke, J. and T. Vanck 2014. Importance weighted inductive transfer learning for regression. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo (Eds.), *Proceedings of ECML*, Volume 8724 of *Lecture Notes in Computer Science*, pp. 466–481. Springer.

Germain, P., A. Habrard, F. Laviolette, and E. Morvant 2013. A PAC-bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, Volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 738–746. JMLR.org.

Ghifary, M., D. Balduzzi, W.B. Kleijn, and M. Zhang. 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence 39*(7): 1414–1430 .

Ghifary, M., W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559.

Ghifary, M., W.B. Kleijn, M. Zhang, D. Balduzzi, and W. Li 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pp. 597–613. Springer.

Gong, B., K. Grauman, and F. Sha 2013a. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, Volume 28, pp. 222–230.

Gong, B., K. Grauman, and F. Sha 2013b. Reshaping visual datasets for domain adaptation. In *NIPS*, pp. 1286–1294.

Gong, B., Y. Shi, F. Sha, and K. Grauman 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pp. 2066–2073.

Gong, M., K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf 2016. Domain adaptation with conditional transferable components. In M. Balcan and K. Q. Weinberger (Eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, Volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 2839–2848. JMLR.org.

Grippo, L. and M. Sciandrone. 2000. On the convergence of the block nonlinear gauss-seidel method under convex constraints. *Oper. Res. Lett. 26*(3): 127–136 .

Guo, Y., H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris 2019, June. Spottune: Transfer learning through adaptive fine-tuning. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Hanneke, S. and S. Kpotufe 2019. On the value of target data in transfer learning. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9867–9877.

Haslett, J. and A.E. Raftery. 1989. Space-time modeling with long-memory dependence: assessing ireland's wind-power resource. technical report. *Journal of the Royal Statistical Society 38*(1): 1–50 .

He, K., X. Zhang, S. Ren, and J. Sun 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hedegaard, L., O.A. Sheikh-Omar, and A. Iosifidis. 2021. Supervised domain adaptation: A graph embedding perspective and a rectified experimental protocol. *IEEE Trans. Image Process.* 30: 8619–8631 .

Hoffman, J., B. Kulis, T. Darrell, and K. Saenko 2012. Discovering latent domains for multisource domain adaptation. In *ECCV*, Volume 7573, pp. 702–715.

Hoffman, J., M. Mohri, and N. Zhang 2018. Algorithms and theory for multiple-source adaptation. In *Proceedings of NeurIPS*, pp. 8256–8266.

Hoffman, J., M. Mohri, and N. Zhang. 2021. Multiple-source adaptation theory and algorithms. *Annals of Mathematics and Artificial Intelligence 89*(3-4): 237–270 .

Hoffman, J., M. Mohri, and N. Zhang. 2022. Multiple-source adaptation theory and algorithms - addendum. *Annals of Mathematics and Artificial Intelligence 90*(6): 569–572 .

Horst, R. and N.V. Thoai. 1999. DC programming: overview. *Journal of Optimization Theory and Applications 103*(1): 1–43 .

Houlsby, N., A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. 2019. Parameter-efficient transfer learning for NLP. *CoRR* abs/1902.00751: 1–12 .

Howard, J. and S. Ruder 2018, July. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 328–339. Association for Computational Linguistics.

Huang, J., A.J. Smola, A. Gretton, K.M. Borgwardt, and B. Schölkopf 2006. Correcting sample selection bias by unlabeled data. In *NIPS 2006*, Volume 19, pp. 601–608.

Huang, X., Y. Rao, H. Xie, T.L. Wong, and F.L. Wang 2017. Cross-domain sentiment classification via topic-related tradaboost. In *Thirty-First AAAI Conference on Artificial Intelligence*.

III, H.D. 2007. Frustratingly easy domain adaptation. In J. Carroll, A. van den Bosch, and A. Zaenen (Eds.), *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Ikonomovska, E. 2009. Airline dataset. Online.

Jhuo, I.H., D. Liu, D. Lee, and S.F. Chang 2012. Robust visual domain adaptation with low-rank reconstruction. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2168–2175. IEEE.

Khosla, A., T. Zhou, T. Malisiewicz, A.A. Efros, and A. Torralba 2012. Undoing the damage of dataset bias. In *ECCV*, Volume 7572, pp. 158–171.

Kifer, D., S. Ben-David, and J. Gehrke 2004. Detecting change in data streams. In M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer (Eds.), *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB 2004, Toronto, Canada, August 31 - September 3 2004*, pp. 180–191. Morgan Kaufmann.

Kiros, R., Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pp. 3294–3302.

Koenecke, A., A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, and S. Goel. 2020. Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. USA 117*(14): 7684–7689 .

Koltchinskii, V. and D. Panchenko. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics* 30: 1–42 .

Konstantinov, N. and C. Lampert 2019. Robust learning from untrusted sources. In *International Conference on Machine Learning*, pp. 3488–3498.

Kpotufe, S. and G. Martinet 2018. Marginal singularity, and the benefits of labels in covariate-shift. In S. Bubeck, V. Perchet, and P. Rigollet (Eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, Volume 75 of *Proceedings of Machine Learning Research*, pp. 1882–1886. PMLR.

Krizhevsky, A., G. Hinton, et al. 2009. Learning multiple layers of features from tiny images. Technical report, Toronto University.

Kundu, J.N., N. Venkat, R.V. Babu, et al. 2020. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4544–4553.

Kuzborskij, I. and F. Orabona 2013. Stability and hypothesis transfer learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, Volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 942–950. JMLR.org.

Kwon, T.M. 2004. TMC traffic data automation for Mn/DOT's traffic monitoring program. *Univ. of Minnesota* Report no. Mn/DOT 2004-29: 1–51 .

Li, J., K. Lu, Z. Huang, L. Zhu, and H.T. Shen. 2018. Transfer independently together: A generalized framework for domain adaptation. *IEEE transactions on cybernetics 49*(6): 2144–2155 .

Li, Q. 2012. Literature survey: domain adaptation algorithms for natural language processing.

Li, Q., Z. Zhu, and G. Tang 2019. Alternating minimizations converge to second-order optimal solutions. In K. Chaudhuri and R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Volume 97 of *Proceedings of Machine Learning Research*, pp. 3935–3943. PMLR.

Liu, H., M. Shao, and Y. Fu 2016. Structure-preserved multi-source domain adaptation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1059–1064. IEEE.

Long, M., Y. Cao, J. Wang, and M.I. Jordan 2015. Learning transferable features with deep adaptation networks. In F. R. Bach and D. M. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, Volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 97–105. JMLR.org.

Long, M., H. Zhu, J. Wang, and M.I. Jordan 2016. Unsupervised domain adaptation with residual transfer networks. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 136–144.

Lu, N., T. Zhang, T. Fang, T. Teshima, and M. Sugiyama. 2021. Rethinking importance weighting for transfer learning.

Mansour, Y., M. Mohri, J. Ro, A. Theertha Suresh, and K. Wu 2021. A theory of multiple-source adaptation with limited target labeled data. In A. Banerjee and K. Fukumizu (Eds.), *The 24th International Conference on Artificial Intelligence*

and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event, Volume 130 of *Proceedings of Machine Learning Research*, pp. 2332–2340. PMLR.

Mansour, Y., M. Mohri, and A. Rostamizadeh 2009a. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*.

Mansour, Y., M. Mohri, and A. Rostamizadeh 2009b. Domain adaptation with multiple sources. In *NIPS*, pp. 1041–1048.

Maurer, A. 2006. Bounds for linear multi-task learning. *J. Mach. Learn. Res.* 7: 117–139 .

Maurer, A., M. Pontil, and B. Romera-Paredes. 2016. The benefit of multitask representation learning. *J. Mach. Learn. Res.* 17: 81:1–81:32 .

Meir, R. and T. Zhang. 2003. Generalization error bounds for Bayesian mixture algorithms. *J. Mach. Learn. Res.* 4: 839–860 .

Mohri, M. and A. Muñoz Medina 2012. New analysis and algorithm for learning with drifting distributions. In N. H. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann (Eds.), *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, Volume 7568 of *Lecture Notes in Computer Science*, pp. 124–138. Springer.

Mohri, M., A. Rostamizadeh, and A. Talwalkar. 2018. *Foundations of Machine Learning* (Second ed.). MIT Press.

Mohri, M., G. Sivek, and A.T. Suresh 2019. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR.

Motiian, S., Q. Jones, S. Iranmanesh, and G. Doretto 2017. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 6670–6680.

Motiian, S., M. Piccirilli, D.A. Adjeroh, and G. Doretto 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5715–5725.

Muandet, K., D. Balduzzi, and B. Schölkopf 2013. Domain generalization via invariant feature representation. In *ICML*, Volume 28, pp. 10–18.

Nichol, A., J. Achiam, and J. Schulman. 2018. On first-order meta-learning algorithms.

Pan, S.J. and Q. Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*(10): 1345–1359 .

Pavlopoulos, J., J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos. 2020. Toxicity detection: Does context really matter?

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830 .

Pei, Z., Z. Cao, M. Long, and J. Wang 2018. Multi-adversarial domain adaptation. In *AAAI*, pp. 3934–3941.

Peng, X., Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415.

Pentina, A. and S. Ben-David 2018. Multi-task Kernel Learning based on Probabilistic Lipschitzness. In F. Janoos, M. Mohri, and K. Sridharan (Eds.), *Algorithmic Learning Theory, ALT 2018, 7-9 April 2018, Lanzarote, Canary Islands, Spain*, Volume 83 of *Proceedings of Machine Learning Research*, pp. 682–701. PMLR.

Pentina, A. and C.H. Lampert 2014. A PAC-bayesian bound for lifelong learning. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, Volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 991–999. JMLR.org.

Pentina, A. and C.H. Lampert 2015. Lifelong learning with non-i.i.d. tasks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 1540–1548.

Pentina, A. and C.H. Lampert 2017. Multi-task learning with labeled and unlabeled tasks. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Volume 70 of *Proceedings of Machine Learning Research*, pp. 2807–2816. PMLR.

Pentina, A. and R. Urner 2016. Lifelong learning with weighted majority votes. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3612–3620.

Perrot, M. and A. Habrard 2015. A theoretical analysis of metric hypothesis transfer learning. In F. R. Bach and D. M. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, Volume 37 of *JMLR Workshop and Conference Proceedings*, pp.

1708–1717. JMLR.org.

Peters, M.E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer 2018, June. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp. 2227–2237. Association for Computational Linguistics.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P.J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21: 140:1–140:67 .

Redko, I. and Y. Bennani. 2016. Non-negative embedding for fully unsupervised domain adaptation. *Pattern Recognit. Lett.* 77: 35–41 .

Redko, I., A. Habrard, and M. Sebban 2017. Theoretical analysis of domain adaptation with optimal transport. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Dzeroski (Eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part II*, Volume 10535 of *Lecture Notes in Computer Science*, pp. 737–753. Springer.

Rodriguez-Lujan, I., J. Fonollosa, A. Vergara, M. Homer, and R. Huerta. 2014. On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. *Chemometrics and Intelligent Laboratory Systems* 130: 123–134. https://doi.org/https://doi.org/10.1016/j.chemolab.2013.10.012 .

Saito, K., D. Kim, S. Sclaroff, T. Darrell, and K. Saenko 2019. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8050–8058.

Saito, K., K. Watanabe, Y. Ushiku, and T. Harada 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732.

Sener, O., H.O. Song, A. Saxena, and S. Savarese 2016. Learning transferrable representations for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 2110–2118.

Sriperumbudur, B.K., D.A. Torres, and G.R.G. Lanckriet 2007. Sparse eigen methods by D.C. programming. In *ICML*, pp. 831–838.

Subramanian, S., A. Trischler, Y. Bengio, and C.J. Pal 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

OpenReview.net.

Sugiyama, M., M. Krauledat, and K. Müller. 2007. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* 8: 985–1005 .

Sugiyama, M., S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 1433–1440. Curran Associates, Inc.

Sun, B., J. Feng, and K. Saenko 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 30.

Sun, B. and K. Saenko 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer.

Sun, Q., R. Chattopadhyay, S. Panchanathan, and J. Ye 2011. A two-stage weighting framework for multi-source domain adaptation. In *Advances in neural information processing systems*, pp. 505–513.

Tao, P.D. and L.T.H. An. 1997. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica 22*(1): 289–355 .

Tao, P.D. and L.T.H. An. 1998. A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization 8*(2): 476–505 .

Tuy, H. 1964. Concave programming under linear constraints. *Translated Soviet Mathematics* 5: 1437–1440 .

Tzeng, E., J. Hoffman, T. Darrell, and K. Saenko 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4068–4076.

Vergara, A., S. Vembu, T. Ayhan, M.A. Ryan, M.L. Homer, and R. Huerta. 2012. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical* 166-167: 320–329. https://doi.org/https://doi.org/10.1016/j.snb.2012.01.074 .

Wang, B., J.A. Mendez, M. Cai, and E. Eaton 2019. Transfer learning via minimizing the performance gap between domains. In *Proceedingz of NeurIPS*, pp. 10644–10654.

Wang, C. and S. Mahadevan 2011. Heterogeneous domain adaptation using manifold alignment. In *Twenty-second international joint conference on artificial intelligence*.

Wang, J., W. Feng, Y. Chen, H. Yu, M. Huang, and P.S. Yu 2018. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 402–410.

Wang, M. and W. Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312: 135–153 .

Wang, T., X. Zhang, L. Yuan, and J. Feng 2019. Few-shot adaptive faster r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7173–7182.

Wei, J., M. Bosma, V.Y. Zhao, K. Guu, A.W. Yu, B. Lester, N. Du, A.M. Dai, and Q.V. Le. 2021. Finetuned language models are zero-shot learners.

Wen, J., R. Greiner, and D. Schuurmans 2020. Domain aggregation networks for multi-source domain adaptation. In *International Conference on Machine Learning*, pp. 10214–10224. PMLR.

Yang, J., R. Yan, and A.G. Hauptmann 2007. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, pp. 188–197.

Yang, L., S. Hanneke, and J.G. Carbonell. 2013. A theory of transfer learning with applications to active learning. *Mach. Learn. 90*(2): 161–189 .

You, K., Z. Kou, M. Long, and J. Wang 2020. Co-tuning for transfer learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yuille, A.L. and A. Rangarajan. 2003. The concave-convex procedure. *Neural Computation 15*(4): 915–936 .

Zhang, K., B. Schölkopf, K. Muandet, and Z. Wang 2013. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, Volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 819–827. JMLR.org.

Zhang, T., I. Yamane, N. Lu, and M. Sugiyama 2020. A one-step approach to covariate shift adaptation. In *Proceedings of ACML*, Volume 129 of *Proceedings of Machine Learning Research*, pp. 65–80. PMLR.

Zhang, Y., T. Liu, M. Long, and M. Jordan 2019a, 09–15 Jun. Bridging theory and algorithm for domain adaptation. In K. Chaudhuri and R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, Volume 97 of *Proceedings of Machine Learning Research*, pp. 7404–7413. PMLR.

Zhang, Y., T. Liu, M. Long, and M. Jordan 2019b. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pp. 7404–7413. PMLR.

Zhang, Y., T. Liu, M. Long, and M.I. Jordan 2019c. Bridging theory and algorithm for domain adaptation. In K. Chaudhuri and R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Volume 97 of *Proceedings of Machine Learning Research*, pp. 7404–7413. PMLR.

Zhang, Y., M. Long, J. Wang, and M.I. Jordan. 2020. On localized discrepancy for domain adaptation.

Zhao, H., R.T. Des Combes, K. Zhang, and G. Gordon 2019. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR.

Zhao, H., S. Zhang, G. Wu, J.M. Moura, J.P. Costeira, and G.J. Gordon. 2018. Adversarial multiple source domain adaptation. *Advances in neural information processing systems* 31: 8559–8570 .

Zheng, L., G. Liu, C. Yan, C. Jiang, M. Zhou, and M. Li. 2020. Improved tradaboost and its application to transaction fraud detection. *IEEE Transactions on Computational Social Systems 7*(5): 1304–1316 .

Contents of Appendix

A Best-effort adaptation **40**
A.1 Theorems and proofs 40
A.2 Convex optimization solution 44
A.3 Discrepancy estimation 45
A.4 Pseudocode of alternate minimization procedure 46
A.5 α-reweighting method 46

B Domain adaptation **49**
B.1 Proof of Lemma 8 49
B.2 Proof of Lemma 9 49
B.3 Sub-Gradients and estimation of unlabeled discrepancy terms 50
B.3.1 Sub-Gradients of unlabeled weighted discrepancy terms . . . 50
B.3.2 Estimation of unlabeled discrepancy terms 51

C Further details about experimental settings **53**
C.1 Best-Effort adaptation 53
C.1.1 Simulated data 53
C.1.2 Real-world data: classification and regression 54
C.2 Fine-tuning tasks 56
C.3 Domain adaptation 57
C.3.1 Domain adaptation – covariate-shift 58

Appendix A Best-effort adaptation

A.1 Theorems and proofs

Below we will work with a notion of discrepancy extended to finite signed measures, as defined in (4).

Theorem 1 *Fix a vector* q *in* $[0,1]^{[m+n]}$*. Then, for any* $\delta > 0$*, with probability at least* $1-\delta$ *over the draw of an i.i.d. sample* S *of size* m *from* $\mathcal{Q}$ *and an i.i.d. sample* S' *of size* n *from* $\mathcal{P}$*, the following holds for all* $h \in \mathcal{H}$*:*

$$\mathcal{L}(\mathcal{P},h) \leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i),y_i) + \mathrm{dis}\Big(\big[(1-\|\mathsf{q}\|_1)+\overline{\mathsf{q}}\big]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}\Big) + 2\mathfrak{R}_{\mathsf{q}}(\ell\circ\mathcal{H}) + \|\mathsf{q}\|_2\sqrt{\frac{\log\frac{1}{\delta}}{2}}.$$

Proof Let $S = ((x_1,y_1),\ldots,(x_m,y_m))$ be a sample of size m drawn i.i.d. from $\mathcal{Q}$ and similarly $S' = ((x_{m+1},y_{m+1}),\ldots,(x_{m+n},y_{m+n}))$ a sample of size n drawn i.i.d. from $\mathcal{P}$. Let T denote the sample formed by S and S', $T = (S,S')$. For any such sample T, define $\Phi(T)$ as follows:

$$\Phi(T) = \sup_{h\in\mathcal{H}}\Big\{\overline{\mathsf{q}}\mathcal{L}(\mathcal{Q},h) + (\|\mathsf{q}\|_1 - \overline{\mathsf{q}})\mathcal{L}(\mathcal{P},h) - \mathcal{L}_T(\mathsf{q},h)\Big\},$$

with $\mathcal{L}_T(\mathsf{q},h) = \sum_{i=1}^{m+n}\mathsf{q}_i\ell(h(x_i),y_i)$. Changing point x_i to some other point x_i' affects $\Phi(T)$ by at most q_i. Thus, by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:

$$\overline{\mathsf{q}}\mathcal{L}(\mathcal{Q},h) + (\|\mathsf{q}\|_1 - \overline{\mathsf{q}})\mathcal{L}(\mathcal{P},h) \leq \mathcal{L}_T(\mathsf{q},h) + \mathbb{E}[\Phi(T)] + \|\mathsf{q}\|_2\sqrt{\frac{\log\frac{1}{\delta}}{2}}. \tag{A1}$$

Now, let $T' = ((x_1',y_1'),\ldots,(x_m',y_m'),(x_{m+1}',y_{m+1}'),\ldots,(x_{m+n}',y_{m+n}')))$ be a sample drawn according to the same distribution as T, then we can write:

$$\begin{aligned}
\mathop{\mathbb{E}}_{T'}[\mathcal{L}_{T'}(q,h)] &= \sum_{i=1}^{m} q_i\,\mathbb{E}[\ell(h(x_i'),y_i')] + \sum_{i=m+1}^{m+n} q_i\,\mathbb{E}[\ell(h(x_i'),y_i')] \\
&\qquad \text{(linearity of expectation and weights } \mathsf{q}_i \text{ independent of } T\text{)} \\
&= \sum_{i=1}^{m} q_i\mathcal{L}(\mathcal{Q},h) + \sum_{i=m+1}^{m+n} q_i\mathcal{L}(\mathcal{P},h) \qquad \text{(i.i.d. sample)} \\
&= \overline{\mathsf{q}}\mathcal{L}(\mathcal{Q},h) + (\|\mathsf{q}\|_1 - \overline{\mathsf{q}})\mathcal{L}(\mathcal{P},h). \qquad \text{(A2)}
\end{aligned}$$

In light of that equality, we can analyze the expectation term as follows:

$$\begin{aligned}
\mathbb{E}[\Phi(T)] &= \mathop{\mathbb{E}}_{T}\left[\sup_{h\in\mathcal{H}} \overline{\mathsf{q}}\mathcal{L}(\mathcal{Q},h) + (\|\mathsf{q}\|_1 - \overline{\mathsf{q}})\mathcal{L}(\mathcal{P},h) - \mathcal{L}_T(\mathsf{q},h)\right] \\
&= \mathop{\mathbb{E}}_{T}\left[\sup_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{T'}[\mathcal{L}_{T'}(\mathsf{q},h)] - \mathcal{L}_T(\mathsf{q},h)\right] \\
&= \mathop{\mathbb{E}}_{T}\left[\sup_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{T'}[\mathcal{L}_{T'}(\mathsf{q},h) - \mathcal{L}_T(\mathsf{q},h)]\right] \qquad (\mathcal{L}_{T'}(\mathsf{q},h) \text{ independent of } T') \\
&\leq \mathop{\mathbb{E}}_{T,T'}\left[\sup_{h\in\mathcal{H}} \mathcal{L}_{T'}(\mathsf{q},h) - \mathcal{L}_T(\mathsf{q},h)\right] \qquad \text{(sub-additivity of supremum)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{T,T'}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i'), y_i') - \mathsf{q}_i \ell(h(x_i), y_i)\right] \\
&= \mathop{\mathbb{E}}_{T,T',\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m+n} \sigma_i\big(\mathsf{q}_i \ell(h(x_i'), y_i') - \mathsf{q}_i \ell(h(x_i), y_i)\big)\right] \\
&\qquad\qquad \text{(introducing Rademacher variables } \sigma_i\text{)} \\
&\leq \mathop{\mathbb{E}}_{T',\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m+n} \sigma_i \mathsf{q}_i \ell(h(x_i'), y_i')\right] + \mathop{\mathbb{E}}_{T,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m+n} -\sigma_i \mathsf{q}_i \ell(h(x_i), y_i)\right] \\
&\qquad\qquad \text{(sub-addivity of supremum and linearity of expectation)} \\
&= 2 \mathop{\mathbb{E}}_{T,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m+n} \sigma_i \mathsf{q}_i \ell(h(x_i), y_i)\right] \quad (-\sigma_i \text{ and } \sigma_i \text{ follow the same distribution}) \\
&= 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}).
\end{aligned}
$$

Finally, using the upper bound

$$
\begin{aligned}
\mathcal{L}(\mathcal{P}, h) - [\overline{\mathsf{q}}\mathcal{L}(\mathcal{Q}, h) + (\|\mathsf{q}\|_1 - \overline{\mathsf{q}})\mathcal{L}(\mathcal{P}, h)] &= [(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{L}(\mathcal{P}, h) - \overline{\mathsf{q}}\mathcal{L}(\mathcal{Q}, h) \\
&\leq \mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}),
\end{aligned}
$$

inequality (A1), and the upper bound on $\mathbb{E}[\Phi(T)]$, we obtain:

$$
\begin{aligned}
\mathcal{L}(\mathcal{P}, h) &\leq [\overline{\mathsf{q}}\mathcal{L}(\mathcal{Q}, h) + (\|\mathsf{q}\|_1 - \overline{\mathsf{q}})\mathcal{L}(\mathcal{P}, h)] + \mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) \\
&\leq \mathcal{L}_T(\mathsf{q}, h) + \mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) + 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) + \|\mathsf{q}\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{2}},
\end{aligned}
$$

which completes the proof. □

Next, we show that the learning bound just proven is tight in terms of the weighted-discrepancy term.

Theorem 2 *Fix a distribution* q *in the simplex* Δ_{m+n}*. Then, for any* $\epsilon > 0$*, there exists* $h \in \mathcal{H}$ *such that, for any* $\delta > 0$*, the following lower bound holds with probability at least* $1 - \delta$ *over the draw of an i.i.d. sample* S *of size* m *from* $\mathcal{Q}$ *and an i.i.d. sample* S' *of size* n *from* $\mathcal{P}$*:*

$$
\mathcal{L}(\mathcal{P}, h) \geq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) - 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) - \|\mathsf{q}\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{2}} - \epsilon.
$$

In particular, for $\|\mathsf{q}\|_2, \mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) \in O\Big(\frac{1}{\sqrt{m+n}}\Big)$*, we have:*

$$
\mathcal{L}(\mathcal{P}, h) \geq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \Omega\left(\frac{1}{\sqrt{m+n}}\right).
$$

Proof Let $\mathcal{L}(\mathsf{q}, h)$ denote $\sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i)$. By definition of discrepancy as a supremum, for any $\epsilon > 0$, there exists $h \in \mathcal{H}$ such that $\mathcal{L}(\mathcal{P}, h) - \mathcal{L}(\mathcal{Q}, h) \geq \mathrm{dis}(\mathcal{P}, \mathcal{Q}) - \epsilon$. For that h, we have

$$
\begin{aligned}
\mathcal{L}(\mathcal{P}, h) - \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) - \mathcal{L}(\mathsf{q}, h) &\geq \mathcal{L}(\mathcal{P}, h) - \overline{\mathsf{q}}(\mathcal{L}(\mathcal{P}, h) - \mathcal{L}(\mathcal{Q}, h)) - \mathcal{L}(\mathsf{q}, h) - \epsilon \\
&= (1 - \overline{\mathsf{q}})\mathcal{L}(\mathcal{P}, h) + \overline{\mathsf{q}}\mathcal{L}(\mathcal{Q}, h) - \mathcal{L}(\mathsf{q}, h) - \epsilon \\
&= \mathbb{E}[\mathcal{L}(\mathsf{q}, h)] - \mathcal{L}(\mathsf{q}, h) - \epsilon.
\end{aligned}
$$

By McDiarmid's inequality, with probability at least $1-\delta$, we have $\mathbb{E}[\mathcal{L}(\mathsf{q},h)] - \mathcal{L}(\mathsf{q},h) \geq -2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) - \|\mathsf{q}\|_2\sqrt{\frac{\log\frac{1}{\delta}}{2}}$. Thus, we have:

$$\mathcal{L}(\mathcal{P},h) - \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P},\mathcal{Q}) - \mathcal{L}(\mathsf{q},h) \geq -2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) - \|\mathsf{q}\|_2\sqrt{\frac{\log\frac{1}{\delta}}{2}} - \epsilon.$$

The last inequality follows directly by using the assumptions and Lemma 10. □

Theorem 3 *For any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{Q}$ and an i.i.d. S' of size n from $\mathcal{P}$, the following holds for all $h \in \mathcal{H}$ and $\mathsf{q} \in \left\{\mathsf{q}\colon \|\mathsf{q}-\mathsf{p}^0\|_1 < 1\right\}$:*

$$\mathcal{L}(\mathcal{P},h) \leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i),y_i) + \mathrm{dis}\Big(\Big[(1-\|\mathsf{q}\|_1)+\overline{\mathsf{q}}\Big]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}\Big) + \mathrm{dis}(\mathsf{p}^0,\mathsf{q})$$
$$+ 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) + 7\|\mathsf{q}-\mathsf{p}^0\|_1 + \Big[\|\mathsf{q}\|_2 + 2\|\mathsf{q}-\mathsf{p}^0\|_1\Big]\left[\sqrt{\log\log_2 \tfrac{2}{1-\|\mathsf{q}-\mathsf{p}^0\|_1}} + \sqrt{\tfrac{\log\frac{2}{\delta}}{2}}\right].$$

Proof Consider two sequences $(\epsilon_k)_{k\geq 0}$ and $(\mathsf{q}^k)_{k\geq 0}$. By Theorem 1, for any fixed $k \geq 0$, we have:

$$\mathbb{P}\left[\mathcal{L}(\mathcal{P},h) > \sum_{i=1}^{m+n} \mathsf{q}_i^k \ell(h(x_i),y_i) + \mathrm{dis}\Big(\Big[(1-\|\mathsf{q}^k\|_1)+\overline{\mathsf{q}}^k\Big]\mathcal{P}, \overline{\mathsf{q}}^k\mathcal{Q}\Big) \right.$$
$$\left. + 2\mathfrak{R}_{\mathsf{q}^k}(\ell \circ \mathcal{H}) + \frac{\|\mathsf{q}^k\|_2}{\sqrt{2}}\epsilon_k\right] \leq e^{-\epsilon_k^2}.$$

Choose $\epsilon_k = \epsilon + \sqrt{2\log(k+1)}$. Then, by the union bound, we can write:

$$\mathbb{P}\left[\exists k \geq 1\colon \mathcal{L}(\mathcal{P},h) > \sum_{i=1}^{m+n} \mathsf{q}_i^k \ell(h(x_i),y_i) + \mathrm{dis}\Big(\Big[(1-\|\mathsf{q}^k\|_1)+\overline{\mathsf{q}}^k\Big]\mathcal{P}, \overline{\mathsf{q}}^k\mathcal{Q}\Big)\right. \tag{A3}$$
$$\left. + 2\mathfrak{R}_{\mathsf{q}^k}(\ell \circ \mathcal{H}) + \frac{\|\mathsf{q}^k\|_2}{\sqrt{2}}\epsilon_k\right]$$
$$\leq \sum_{k=0}^{+\infty} e^{-\epsilon_k^2} \leq \sum_{k=0}^{+\infty} e^{-\epsilon^2-\log((k+1)^2)} = e^{-\epsilon^2}\sum_{k=1}^{+\infty}\frac{1}{k^2} = \frac{\pi^2}{6}e^{-\epsilon^2} \leq 2e^{-\epsilon^2}.$$

We can choose q^k such that $\|\mathsf{q}^k-\mathsf{p}^0\|_1 = 1-\frac{1}{2^k}$. Then, for any $\mathsf{q} \in \left\{\mathsf{q}\colon \|\mathsf{q}-\mathsf{p}^0\|_1 < 1\right\}$, there exists $k \geq 0$ such that $\|\mathsf{q}^k-\mathsf{p}^0\|_1 \leq \|\mathsf{q}-\mathsf{p}^0\|_1 < \|\mathsf{q}^{k+1}-\mathsf{p}^0\|_1$ and thus such that

$$\sqrt{2\log(k+1)} = \sqrt{2\log\log_2\frac{1}{1-\|\mathsf{q}^{k+1}-\mathsf{p}^0\|_1}} = \sqrt{2\log\log_2\frac{2}{1-\|\mathsf{q}^{k}-\mathsf{p}^0\|_1}}$$
$$\leq \sqrt{2\log\log_2\frac{2}{1-\|\mathsf{q}-\mathsf{p}^0\|_1}}.$$

Furthermore, for that k, the following inequalities hold:

$$\sum_{i=1}^{m+n} \mathsf{q}_i^k \ell(h(x_i),y_i) \leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i),y_i) + \mathrm{dis}(\mathsf{q}^k,\mathsf{q})$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \mathrm{dis}(\mathsf{q}^k, \mathsf{p}^0) + \mathrm{dis}(\mathsf{p}^0, \mathsf{q}) \\
&\leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \|\mathsf{q}^k - \mathsf{p}^0\|_1 + \mathrm{dis}(\mathsf{p}^0, \mathsf{q}) \\
&\leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \|\mathsf{q} - \mathsf{p}^0\|_1 + \mathrm{dis}(\mathsf{p}^0, \mathsf{q}),
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{dis}\Big(\Big[(1 - \|\mathsf{q}^k\|_1) + \overline{\mathsf{q}}^k\Big]\mathcal{P}, \overline{\mathsf{q}}^k\mathcal{Q}\Big) &\leq \mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) \\
&\quad + \Big\|\Big[(\|\mathsf{q}\|_1 - \overline{\mathsf{q}}) - (\|\mathsf{q}^k\|_1 - \overline{\mathsf{q}}^k)\Big]\mathcal{P} + \Big[\overline{\mathsf{q}} - \overline{\mathsf{q}}^k\Big]\mathcal{Q}\Big\|_1 \\
&\leq \mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) + \|\mathsf{q}^k - \mathsf{q}\|_1 \\
&\leq \mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) + 2\|\mathsf{q} - \mathsf{p}^0\|_1,
\end{aligned}
$$

$$
\mathfrak{R}_{\mathsf{q}^k}(\ell \circ \mathcal{H}) \leq \mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) + \|\mathsf{q}^k - \mathsf{q}\|_1 \leq \mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) + 2\|\mathsf{q} - \mathsf{p}^0\|_1,
$$

$$
\begin{aligned}
\text{and} \qquad \|\mathsf{q}^k\|_2 &\leq \|\mathsf{q}\|_2 + \|\mathsf{q}^k - \mathsf{q}\|_2 \\
&\leq \|\mathsf{q}\|_2 + \|\mathsf{q}^k - \mathsf{q}\|_1 \leq \|\mathsf{q}\|_2 + 2\|\mathsf{q} - \mathsf{p}^0\|_1.
\end{aligned}
$$

Plugging in these inequalities in (A3) concludes the proof. □

Corollary 4 *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{Q}$ and an i.i.d. sample S' of size n from $\mathcal{P}$, the following holds for all $h \in \mathcal{H}$ and $\mathsf{q} \in \left\{\mathsf{q} : \|\mathsf{q} - \mathsf{p}^0\|_1 < 1\right\}$:*

$$
\begin{aligned}
\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m+n} \mathsf{q}_i \ell(h(x_i), y_i) + \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \mathrm{dis}(\mathsf{p}^0, \mathsf{q}) + 2\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) \\
+ 8\|\mathsf{q} - \mathsf{p}^0\|_1 + \Big[\|\mathsf{q}\|_2 + 2\|\mathsf{q} - \mathsf{p}^0\|_1\Big]\left[\sqrt{\log\log_2 \tfrac{2}{1 - \|\mathsf{q} - \mathsf{p}^0\|_1}} + \sqrt{\frac{\log \frac{2}{\delta}}{2}}\right].
\end{aligned}
$$

Proof Note that the discrepancy term of the bound of Theorem 3 can be further upper bounded as follows:

$$
\begin{aligned}
&\mathrm{dis}([(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}]\mathcal{P}, \overline{\mathsf{q}}\mathcal{Q}) \\
&= \sup_{h \in \mathcal{H}} \left\{[(1 - \|\mathsf{q}\|_1) + \overline{\mathsf{q}}] \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}}[\ell(h(x), y)] - \overline{\mathsf{q}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{Q}}[\ell(h(x), y)]\right\} \\
&\leq \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \left|1 - \|\mathsf{q}\|_1\right| \sup_{h \in \mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}}[\ell(h(x), y)] \\
&\leq \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \left|1 - \|\mathsf{q}\|_1\right| \\
&= \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \left|\|\mathsf{p}^0\|_1 - \|\mathsf{q}\|_1\right| \\
&\leq \overline{\mathsf{q}}\mathrm{dis}(\mathcal{P}, \mathcal{Q}) + \|\mathsf{p}^0 - \mathsf{q}\|_1.
\end{aligned}
$$

Plugging this in the right-hand side in the bound of Theorem 3 completes the proof. □

Lemma 10 *Fix a distribution* q *over* $[m+n]$. *Then, the following holds for the* q*-weighted Rademacher complexity:*

$$\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) \leq \|\mathsf{q}\|_\infty (m+n)\, \mathfrak{R}_{m+n}(\ell \circ \mathcal{H}).$$

Proof Since for any $i \in [m+n]$, the function $\varphi_i : x \mapsto \mathsf{q}_i x$ is q_i-Lipschitz and thus $\|\mathsf{q}\|_\infty$-Lipschitz, the result is an application of the result of Meir and Zhang (Meir and Zhang, 2003, Theorem 7). □

Note that the bound of the lemma is tight: equality holds when q is chosen to be the uniform distribution. By McDiarmid's inequality, the q-weighted Rademacher complexity can be estimated from the empirical quantity

$$\widehat{\mathfrak{R}}_{\mathsf{q},S,S'}(\ell \circ \mathcal{H}) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m+n} \sigma_i \mathsf{q}_i \ell(h(x_i), y_i)\right],$$

modulo a term in $O(\|\mathsf{q}\|_2)$.

A.2 Convex optimization solution

In the case of the squared loss with the hypothesis set of linear functions or kernel-based functions, the optimization algorithm for BEST can be formulated as a convex optimization problem.

We can proceed as follows when ℓ is the squared loss. We introduce new variables $\mathsf{u}_i = 1/\mathsf{q}_i$, $\mathsf{v}_i = 1/\mathsf{p}_i^0$ and define the convex set $\mathcal{U} = \{\mathsf{u} : \mathsf{u}_i \geq 1\}$. Using the following four expressions:

$$\mathsf{q}_i (h(x_i) - y_i)^2 = \frac{(h(x_i) - y_i)^2}{\mathsf{u}_i}, \qquad \|\mathsf{q}\|_2^2 = \sum_i \frac{1}{\mathsf{u}_i^2},$$

$$\|\mathsf{q}\|_\infty \|h\|^2 = \max_i \frac{\|h\|^2}{\mathsf{u}_i} = \frac{\|h\|^2}{\mathsf{u}_{\min}}, \qquad \|\mathsf{q} - \mathsf{p}^0\|_1 \leq \sum_i |\mathsf{v}_i - \mathsf{u}_i| = \|\mathsf{u} - \mathsf{v}\|_1,$$

leads to the following convex optimization problem with new hyperparameters $\gamma_\infty, \gamma_1, \gamma_2$:

$$\min_{h \in \mathcal{H}, \mathsf{u} \in \mathcal{U}} \sum_{i=1}^{m+n} \frac{(h(x_i) - y_i)^2 + d_i}{\mathsf{u}_i} + \mathrm{dis}\left(\left(\frac{1}{\mathsf{u}_i}\right)_i, \left(\frac{1}{\mathsf{v}_i}\right)_i\right) + \gamma_\infty \frac{\|h\|^2}{\mathsf{u}_{\min}} + \gamma_1 \|\mathsf{u} - \mathsf{v}\|_1 + \gamma_2 \sum_{i=1}^{m+n} \frac{1}{\mathsf{u}_i^2}.$$

Note that the first term is jointly convex as a sum of quadratic-over-linear or matrix fractional functions (Boyd and Vandenberghe, 2014). When $\mathcal{H}$ is a subset of the reproducing kernel Hilbert space associated to a positive definite kernel K, for a fixed u, the problem coincides with a standard kernel ridge regression problem. Thus, we

can rewrite it in terms of dual variables $\boldsymbol{\alpha}$, the kernel matrix K, $Y = (y_1, \ldots, y_{m+n})^\top$ and $U = (u_1, \ldots, u_{m+n})^\top$ as follows:

$$\min_{\mathsf{u}\in\mathcal{U}} \max_{\boldsymbol{\alpha}} -\boldsymbol{\alpha}^\top \left(K + \frac{\gamma_\infty}{\mathsf{u}_{\min}} U\right)\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top Y + \sum_{i=1}^{m+n} \frac{d_i}{\mathsf{u}_i}$$
$$\operatorname{dis}\left(\left(\tfrac{1}{\mathsf{u}_i}\right)_i, \left(\tfrac{1}{\mathsf{v}_i}\right)_i\right) + \gamma_1 \|\mathsf{u} - \mathsf{v}\|_1 + \gamma_2 \sum_{i=1}^{m+n} \frac{1}{\mathsf{u}_i^2}.$$

Solving for $\boldsymbol{\alpha}$ yields the following convex optimization problem:

$$\min_{\mathsf{u}\in\mathcal{U}} Y^\top \left(K + \frac{\gamma_\infty}{\mathsf{u}_{\min}} U\right)^{-1} Y + \sum_{i=1}^{m+n} \frac{d_i}{\mathsf{u}_i} + \gamma_1 \|\mathsf{u} - \mathsf{v}\|_1 + \gamma_2 \sum_{i=1}^{m+n} \frac{1}{\mathsf{u}_i^2}.$$

Standard descent methods such as SGD can be used to solve this problem. Note that the above can be further simplified using the upper bound $1/\mathsf{u}_{\min} \leq \sum_{i=1}^{m+n} 1/u_i$.

A.3 Discrepancy estimation

First, note that if the $\mathcal{P}$-drawn labeled sample at our disposal is sufficiently large, we can reserve a sub-sample of size n_1 to train a relatively accurate model $h_{\mathcal{P}}$. Thus, we can subsequently reduce $\mathcal{H}$ to a ball $\mathsf{B}(h_{\mathcal{P}}, r)$ of radius $r \sim \frac{1}{\sqrt{n_1}}$. This helps us work with a finer local labeled discrepancy since the maximum in the definition is then taken over a smaller set.

We do not have access to the discrepancy value $\operatorname{dis}(\mathcal{P}, \mathcal{Q})$, which defines d_is. Instead, we can use the labeled samples from $\mathcal{Q}$ and $\mathcal{P}$ to estimate it. Our estimate $\widehat{d}$ of the discrepancy is given by

$$\widehat{d} = \max_{h\in\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=m+1}^{m+n} \ell(h(x_i), y_i) - \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) \right\}. \tag{A4}$$

Thus, for a convex loss ℓ, the optimization problems for computing $\widehat{d}$ can be naturally cast as DC-programming problem, which can be tackled using the DCA algorithm (Tao and An, 1998) and related methods already discussed for SBEST. For the squared loss, the DCA algorithms is guaranteed to converge to a global optimum (Tao and An, 1998).

By McDiarmid's inequality, with high probability, $|\operatorname{dis}(\mathcal{P}, \mathcal{Q}) - \widehat{d}|$ can be bounded by $O(\sqrt{\frac{m+n}{mn}})$. More refined bounds such as relative deviation bounds or Bernstein-type bounds provide more favorable guarantee when the discrepancy is relatively small. When $\mathcal{H}$ is chosen to be a small ball $\mathsf{B}(h_{\mathcal{P}}, r)$, our estimate of the discrepancy is further refined.

The optimization problem (A4) can be equivalently solved via the following minimization:

$$\widehat{d} = \min_{h\in\mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) - \frac{1}{n} \sum_{i=m+1}^{m+n} \ell(h(x_i), y_i) \right\}.$$

The DCA solution for this problem then consists of solving a sequence of T convex optimization problems where $h_1 \in \mathcal{H}$ is chosen arbitrarity and where h_{t+1}, $t \in [T]$ is obtained as follows

$$h_{t+1} \in \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) - \frac{1}{n} \sum_{i=m+1}^{m+n} \nabla \ell(h_t(x_i), y_i) \cdot (h - h_t) \right\}.$$

The second term of the objective is obtained by a linearization of the loss.

A.4 Pseudocode of alternate minimization procedure

Input: Samples $\{(x_1, y_1), \ldots (x_{m+n}, y_{m+n})\}$, tolerance τ, distribution p_0, max iterations T, hyperparameters $\lambda_\infty, \lambda_1, \lambda_2$, discrepancy estimate $\hat{d}$.

1. Initialize q_0 to be the uniform distribution over $[m+n]$.
2. Initialize $h_0 = \operatorname{argmin}_{h \in H} \sum_{i=1}^{m+n} q_{0,i} \ell(h(x_i), y_i) + \lambda_\infty \|q_0\|_\infty \|h\|^2$.
3. For $t = 1, \ldots T$,
   - Set curr_obj_val $= \sum_{i=1}^{m} q_{t-1,i} \big(\ell(h_{t-1}(x_i), y_i) + \hat{d} \big) + \sum_{i=m+1}^{m+n} q_{t-1,i} \ell(h_{t-1}(x_i), y_i) + \lambda_\infty \|q_{t-1}\|_\infty \|h_{t-1}\|^2 + \lambda_1 \|q_{t-1} - p_0\|_1 + \lambda_2 \|q_{t-1}\|^2$.
   - Compute $q_t = \operatorname{argmin}_{q \in \Delta_{m+n}} \sum_{i=1}^{m} q_i \big(\ell(h_{t-1}(x_i), y_i) + \hat{d} \big) + \sum_{i=m+1}^{m+n} q_i \ell(h_{t-1}(x_i), y_i) + \lambda_\infty \|q\|_\infty \|h_{t-1}\|^2 + \lambda_1 \|q - p_0\|_1 + \lambda_2 \|q\|^2$.
   - Compute $h_t = \operatorname{argmin}_{h \in H} \sum_{i=1}^{m} q_{t,i} \big(\ell(h_{t-1}(x_i), y_i) + \hat{d} \big) + \sum_{i=m+1}^{m+n} q_{t,i} \ell(h_{t-1}(x_i), y_i) + \lambda_\infty \|q_t\|_\infty \|h\|^2$.
   - Set new_obj_val $= \sum_{i=1}^{m} q_{t,i} \big(\ell(h_t(x_i), y_i) + \hat{d} \big) + \sum_{i=m+1}^{m+n} q_{t,i} \ell(h_t(x_i), y_i) + \lambda_\infty \|q_t\|_\infty \|h_t\|^2 + \lambda_1 \|q_t - p_0\|_1 + \lambda_2 \|q_t\|^2$.
   - If $|$curr_obj_val $-$ new_obj_val$| \leq \tau$, return q_t, h_t
4. Print: *AM did not converge in T iterations*. Return q_T, h_T.

Fig. A1 Alternate minimization procedure for best effort adaptation.

A.5 α-reweighting method

Let $d = \operatorname{dis}(\mathcal{P}, \mathcal{Q})$, $\widehat{d}$ and $\widehat{d} = \operatorname{dis}(\widehat{\mathcal{Q}}, \widehat{\mathcal{P}})$. Consider the following simple, and in general suboptimal, choice of q as a distribution defined by:

$$\overline{\mathsf{q}} = \frac{\alpha m}{m+n} \qquad \mathsf{q}_i = \begin{cases} \frac{\overline{\mathsf{q}}}{m} = \frac{\alpha}{m+n} & \text{if } i \in [m]; \\ \frac{1-\overline{\mathsf{q}}}{n} = \frac{m(1-\alpha)+n}{(m+n)n} & \text{otherwise,} \end{cases}$$

where $\alpha = \Psi(1-d)$ for some non-decreasing function Ψ with $\Psi(0) = 0$ and $\Psi(1) = 1$. We will compare the right-hand side of the bound of Theorem 1, which we denote by B, with its right-hand side B_0 for q chosen to be uniform over S' corresponding to

supervised learning on just S':

$$B_0 = \mathcal{L}(\widehat{\mathcal{P}}, h) + 2\mathfrak{R}_n(\ell \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

We now show that under some assumptions, we have $B - B_0 \leq 0$. Thus, even for this sub-optimal choice of $\overline{\mathsf{q}}$, under those assumptions, the guarantee of the theorem is then strictly more favorable than the one for training on S' only, uniformly over $h \in \mathcal{H}$.

By definition of $\widehat{d}$, we can write:

$$\mathcal{L}(\mathsf{q}, h) = \overline{\mathsf{q}}\mathcal{L}(\widehat{\mathcal{Q}}, h) + (1 - \overline{\mathsf{q}})\mathcal{L}(\widehat{\mathcal{P}}, h) \leq \overline{\mathsf{q}}\widehat{d} + \mathcal{L}(\widehat{\mathcal{P}}, h).$$

By definition of the q-Rademacher complexity and the sub-additivity of the supremum, the following inequality holds:

$$\mathfrak{R}_{\mathsf{q}}(\ell \circ \mathcal{H}) \leq \overline{\mathsf{q}}\mathfrak{R}_m(\ell \circ \mathcal{H}) + (1 - \overline{\mathsf{q}})\mathfrak{R}_n(\ell \circ \mathcal{H}).$$

By definition of q, we can write:

$$\begin{aligned}\|\mathsf{q}\|_2^2 n = n\left[m\left(\frac{\overline{\mathsf{q}}}{m}\right)^2 + n\left(\frac{1-\overline{\mathsf{q}}}{n}\right)^2\right] &= \frac{n}{m}\overline{\mathsf{q}}^2 + (1-\overline{\mathsf{q}})^2 \\ &= 1 - 2\overline{\mathsf{q}} + \frac{m+n}{m}\overline{\mathsf{q}}^2 \\ &= 1 - (2-\alpha)\overline{\mathsf{q}} \leq 1 - \overline{\mathsf{q}}.\end{aligned}$$

Thus, using the inequality $\sqrt{1-x} \leq 1 - \frac{x}{2}$, $x \leq 1$, we have:

$$\begin{aligned}B - B_0 &\leq 2\overline{\mathsf{q}}[\mathfrak{R}_m(\ell \circ \mathcal{H}) - \mathfrak{R}_n(\ell \circ \mathcal{H})] + \overline{\mathsf{q}}(d + \widehat{d}) + \left[\sqrt{1-\overline{\mathsf{q}}} - 1\right]\left[\frac{\log \frac{1}{\delta}}{2n}\right]^{\frac{1}{2}} \\ &\leq 2\overline{\mathsf{q}}[\mathfrak{R}_m(\ell \circ \mathcal{H}) - \mathfrak{R}_n(\ell \circ \mathcal{H})] + \overline{\mathsf{q}}(d + \widehat{d}) - \overline{\mathsf{q}}\left[\frac{\log \frac{1}{\delta}}{8n}\right]^{\frac{1}{2}}.\end{aligned}$$

Suppose we are in the regime of relatively small discrepancies and that, given n, both the discrepancy and the empirical discrepancies are upper bounded as follows: $\max\{d, \overline{d}\} < \sqrt{\frac{\log 1/\delta}{32n}}$. Assume also that for $m \gg n$ (which is the setting we are interested in), we have $\mathfrak{R}_m(\ell \circ \mathcal{H}) - \mathfrak{R}_n(\ell \circ \mathcal{H}) \leq 0$. Then, the first term is non-positive and, regardless of the choice of $\alpha < 1$, we have $B - B_0 \leq 0$. Thus, even for this sub-optimal choice of $\overline{\mathsf{q}}$, under some assumptions, the guarantee of the theorem is then strictly more favorable than the one for training on S' only, uniformly over $h \in \mathcal{H}$.

Note that the assumption about the difference of Rademacher complexities is natural. For example, for a kernel-based hypothesis set $\mathcal{H}$ with a normalized kernel such as the Gaussian kernel and the norm of the weight vectors in the reproducing kernel Hilbert space (RKHS) bounded by Λ, it is known that the following inequalities

hold: $\frac{1}{\sqrt{2}}\frac{\Lambda}{\sqrt{m}} \leq \mathfrak{R}_m(\mathcal{H}) \leq \frac{\Lambda}{\sqrt{m}}$ (Mohri et al., 2018). Thus, for $m > 2n$, we have $\mathfrak{R}_m(\mathcal{H}) - \mathfrak{R}_n(\mathcal{H}) \leq \frac{\Lambda}{\sqrt{m}} - \frac{\Lambda}{\sqrt{2n}} < 0$.

Appendix B Domain adaptation

B.1 Proof of Lemma 8

Lemma 8 *Let ℓ be the squared loss. Then, for any hypothesis h_0 in $\mathcal{H}$, the following upper bound holds for the labeled discrepancy:*

$$\mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) \leq \overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) + 2\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}).$$

Proof For any h_0, using the definition of the squared loss, the following inequalities hold:

$$\begin{aligned}
\mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) &= \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),y)] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),y)] \right| \\
&\leq \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),h_0(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),h_0(x))] \right| \\
&\quad + \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),y)] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),h_0(x))] \right. \\
&\qquad \left. + \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),h_0(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),y)] \right| \\
&= \overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) \\
&\quad + 2\sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[h(x)(y-h_0(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[h(x)(y-h_0(x))] \right| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(def. of squared loss)} \\
&= \overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) + 2\delta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}). \qquad \text{(def. of local discrepancy)}
\end{aligned}$$

This completes the proof. □

B.2 Proof of Lemma 9

Lemma 9 *Let ℓ be a loss function that is μ-Lipschitz with respect to its second argument. Then, for any hypothesis h_0 in $\mathcal{H}$, the following upper bound holds for the labeled discrepancy:*

$$\mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) \leq \overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) + \mu\,\eta_{\mathcal{H},h_0}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}).$$

Proof When the loss function ℓ is μ-Lipschitz with respect to its second argument, we can use the following upper bound:

$$\begin{aligned}
\mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}) &= \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),y)] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),y)] \right| \\
&\leq \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),h_0(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),h_0(x))] \right| \\
&\quad + \sup_{h\in\mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),y)] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x),h_0(x))] \right. \\
&\qquad \left. + \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),h_0(x))] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[\ell(h(x),y)] \right|
\end{aligned}$$

$$\leq \overline{\mathrm{dis}}_{\mathcal{H}\times\{h_0\}}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) + \mu \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[|y - h_o(x)|] + \mu \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{Q}}}[|y - h_o(x)|].$$

(ℓ assumed μ-Lipschitz)

This completes the proof. □

B.3 Sub-Gradients and estimation of unlabeled discrepancy terms

Here, we first describe how to compute the sub-gradients of the unlabeled weighted discrepancy term $\mathrm{dis}(\mathsf{q}', \mathsf{p})$ that appears in the optimization problem for domain adaptation (10), and similarly $\overline{\mathrm{dis}}(\mathsf{p}^0, (\mathsf{q}, \mathsf{q}'))$, in the case of the squared loss with linear functions. Next, we show how the same analysis can be used to compute the empirical discrepancy term $\overline{\mathrm{dis}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$, which provides an accurate estimate of $\overline{d} = \overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q})$.

B.3.1 Sub-Gradients of unlabeled weighted discrepancy terms

Let ℓ be the squared loss and let $\mathcal{H}$ be the family of linear functions defined by $\mathcal{H} = \{x \mapsto \mathbf{w} \cdot \mathbf{\Phi}(x) : \|\mathbf{w}\|_2 \leq \Lambda\}$, where $\mathbf{\Phi}$ is a feature mapping from $\mathcal{X}$ to $\mathbb{R}^k$. We can analyze the unlabeled discrepancy term $\overline{\mathrm{dis}}(\mathsf{q}', \mathsf{p})$ using an analysis similar to that of Cortes and Mohri (2014). By definition of the unlabeled discrepancy, we can write:

$$\begin{aligned}
\overline{\mathrm{dis}}(\mathsf{q}', \mathsf{p}) &= \sup_{h,h'\in\mathcal{H}} \left\{ \sum_{i=1}^{n} \mathsf{q}'_i \ell(h(x_{m+i}), h'(x_{m+i})) - \sum_{i=1}^{m} \mathsf{p}_i \ell(h(x_i), h'(x_i)) \right\} \\
&= \sup_{\|\mathbf{w}\|_2, \|\mathbf{w}'\|_2 \leq \Lambda} \left\{ \sum_{i=1}^{n} \mathsf{q}'_i [(\mathbf{w} - \mathbf{w}') \cdot \mathbf{\Phi}(x_{m+i})]^2 - \sum_{i=1}^{m} \mathsf{p}_i [(\mathbf{w} - \mathbf{w}') \cdot \mathbf{\Phi}(x_i)]^2 \right\} \\
&= \sup_{\|\mathbf{u}\|_2 \leq 2\Lambda} \left\{ \sum_{i=1}^{n} \mathsf{q}'_i [\mathbf{u} \cdot \mathbf{\Phi}(x_{m+i})]^2 - \sum_{i=1}^{m} \mathsf{p}_i [\mathbf{u} \cdot \mathbf{\Phi}(x_i)]^2 \right\} \\
&= \sup_{\|\mathbf{u}\|_2 \leq 2\Lambda} \left\{ \sum_{i=1}^{n} \mathsf{q}'_i \mathbf{u}^\top \mathbf{\Phi}(x_{m+i}) \mathbf{\Phi}(x_{m+i})^\top \mathbf{u} - \sum_{i=1}^{m} \mathsf{p}_i \mathbf{u}^\top \mathbf{\Phi}(x_i) \mathbf{\Phi}(x_i)^\top \mathbf{u} \right\} \\
&= \sup_{\|\mathbf{u}\|_2 \leq 2\Lambda} \left\{ \mathbf{u}^\top \left[\sum_{i=1}^{n} \mathsf{q}'_i \mathbf{\Phi}(x_{m+i}) \mathbf{\Phi}(x_{m+i})^\top - \sum_{i=1}^{m} \mathsf{p}_i \mathbf{\Phi}(x_i) \mathbf{\Phi}(x_i)^\top \right] \mathbf{u} \right\} \\
&= 4\Lambda^2 \sup_{\|\mathbf{u}\|_2 \leq 1} \mathbf{u}^\top \mathbf{M}(\mathsf{q}', \mathsf{p}) \mathbf{u} \\
&= 4\Lambda^2 \max\left\{ 0, \sup_{\|\mathbf{u}\|_2 = 1} \mathbf{u}^\top \mathbf{M}(\mathsf{q}', \mathsf{p}) \mathbf{u} \right\} \\
&= 4\Lambda^2 \max\{0, \lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p}))\},
\end{aligned}$$

where $\mathbf{M}(\mathsf{q}', \mathsf{p}) = \sum_{i=1}^{n} \mathsf{q}'_i \mathbf{\Phi}(x_{m+i}) \mathbf{\Phi}(x_{m+i})^\top - \sum_{i=1}^{m} \mathsf{p}_i \mathbf{\Phi}(x_i) \mathbf{\Phi}(x_i)^\top$ and where $\lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p}))$ denotes the maximum eigenvalue of the symmetric matrix $\mathbf{M}(\mathsf{q}', \mathsf{p})$. Thus, the unlabeled discrepancy $\overline{\mathrm{dis}}(\mathsf{q}', \mathsf{p})$ can be obtained from the maximum eigenvalue of a symmetric matrix that is an affine function of q' and p. Since $\lambda_{\max}$ is a convex function and since composition with an affine function preserves

convexity, $\lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p}))$ is a convex function of q' and p. Since the maximum of two convex function is convex, $\max\{0, \lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p}))\}$ is also convex.

Rewriting $\lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p}))$ as $\max_{\|\mathbf{u}\|_2=1} \mathbf{u}^\top \mathbf{M}(\mathsf{q}', \mathsf{p})\mathbf{u}$ helps derive the sub-gradient of $\lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p}))$ using the sub-gradient calculation of the maximum of a set of functions:

$$\nabla_{(\mathsf{q}', \mathsf{p})} \lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p})) = \begin{bmatrix} \mathbf{u}^\top \mathbf{\Phi}(x_{m+1}) \mathbf{\Phi}(x_{m+1})^\top \mathbf{u} \\ \vdots \\ \mathbf{u}^\top \mathbf{\Phi}(x_{m+n}) \mathbf{\Phi}(x_{m+n})^\top \mathbf{u} \\ -\mathbf{u}^\top \mathbf{\Phi}(x_1) \mathbf{\Phi}(x_1)^\top \mathbf{u} \\ \vdots \\ -\mathbf{u}^\top \mathbf{\Phi}(x_m) \mathbf{\Phi}(x_m)^\top \mathbf{u} \end{bmatrix} = \begin{bmatrix} (\mathbf{\Phi}(x_{m+1}) \cdot \mathbf{u})^2 \\ \vdots \\ (\mathbf{\Phi}(x_{m+n}) \cdot \mathbf{u})^2 \\ -(\mathbf{\Phi}(x_1) \cdot \mathbf{u})^2 \\ \vdots \\ -(\mathbf{\Phi}(x_m) \cdot \mathbf{u})^2 \end{bmatrix},$$

where $\mathbf{u}$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{M}(\mathsf{q}', \mathsf{p})$. Alternatively, we can approximate the maximum eigenvalue via the softmax expression

$$f(\mathsf{q}', \mathsf{p}) = \frac{1}{\mu} \log\left[\sum_{j=1}^{k} e^{\mu \lambda_j(\mathbf{M}(\mathsf{q}', \mathsf{p}))}\right] = \frac{1}{\mu} \log\left[\mathrm{Tr}\left(e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}\right)\right],$$

where $e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}$ denotes the matrix exponential of $\mu \mathbf{M}(\mathsf{q}', \mathsf{p})$ and $\lambda_j(\mathbf{M}(\mathsf{q}', \mathsf{p}))$ the jth eigenvalue of $\mathbf{M}(\mathsf{q}', \mathsf{p})$. The matrix exponential can be computed in $O(k^3)$ time by computing the singular value decomposition (SVD) of the matrix. We have:

$$\lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p})) \leq f(\mathsf{q}', \mathsf{p}) \leq \lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p})) + \frac{\log k}{\mu}.$$

Thus, for $\mu = \frac{\log k}{\epsilon}$, $f(\mathsf{q}', \mathsf{p})$ provides a uniform ϵ-approximation of $\lambda_{\max}(\mathbf{M}(\mathsf{q}', \mathsf{p}))$. The gradient of $f(\mathsf{q}', \mathsf{p})$ is given for all $j \in [n]$ and $i \in [m]$ by

$$\nabla_{\mathsf{q}'_j} f(\mathsf{q}', \mathsf{p}) = \frac{\left\langle e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}, \mathbf{\Phi}(x_{m+j}) \mathbf{\Phi}(x_{m+j})^\top \right\rangle}{\mathrm{Tr}\left(e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}\right)} = \frac{\mathbf{\Phi}(x_{m+j})^\top e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})} \mathbf{\Phi}(x_{m+j})}{\mathrm{Tr}\left(e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}\right)}$$

$$\nabla_{\mathsf{p}_i} f(\mathsf{q}', \mathsf{p}) = -\frac{\left\langle e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}, \mathbf{\Phi}(x_i) \mathbf{\Phi}(x_i)^\top \right\rangle}{\mathrm{Tr}\left(e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}\right)} = \frac{\mathbf{\Phi}(x_i)^\top e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})} \mathbf{\Phi}(x_i)}{\mathrm{Tr}\left(e^{\mu \mathbf{M}(\mathsf{q}', \mathsf{p})}\right)}.$$

The sub-gradient of the unlabeled discrepancy term $\overline{\mathrm{dis}}(\mathsf{p}^0, (\mathsf{q}, \mathsf{q}'))$ or a smooth approximation can be derived in a similar, using the same analysis as above.

B.3.2 Estimation of unlabeled discrepancy terms

The unlabeled discrepancy $\overline{d} = \overline{\mathrm{dis}}(\mathcal{P}, \mathcal{Q})$ can be accurately estimated from its empirical version $\overline{\mathrm{dis}}(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$ (Mansour et al., 2009a). In view of the analysis of the previous

section, we have

$$\begin{aligned}\overline{\mathrm{dis}}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}}) &= 4\Lambda^2\lambda_{\max}\big(\mathbf{M}(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})\big)\\ &= 4\Lambda^2\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Phi}(x_{m+i})\mathbf{\Phi}(x_{m+i})^\top - \frac{1}{m}\sum_{i=1}^{m}\mathbf{\Phi}(x_i)\mathbf{\Phi}(x_i)^\top\right).\end{aligned}$$

Thus, this last expression can be used in place of $\overline{d}$ in the optimization problem for domain adaptation.

Appendix C Further details about experimental settings

In this section we provide further details on our experimental setup starting with best effort adaptation.

C.1 Best-Effort adaptation

Recall that in this setting we have labeled data from both source and target, however the amount of labeled data from the source is much larger. We start by describing the baselines that we compare our algorithms with. For the best-effort adaptation problem two natural baselines are to learn a hypothesis solely on the target $\mathcal{P}$, or train solely on the source $\mathcal{Q}$. A third baseline that we consider is the α-reweighted q as discussed in Section 3.2. Note, $\alpha = 1$ corresponds to training on all the available data with a uniform weighting.

C.1.1 Simulated data

We first consider a simulated scenario where n samples from the target distribution $\mathcal{P}$ are generated by first drawing the feature vector x i.i.d. from a normal distribution with zero mean and spherical covariance matrix, i.e, $N(0, I_{d\times d})$. Given x, a binary label $y \in \{-1, +1\}$ is generated as $\text{sgn}(w_p \cdot x)$ for a randomly chosen unit vector $w_p \in \mathbb{R}^d$. For a fixed $\eta \in (0.5, 1)$, $m = 1{,}000$ i.i.d. samples from the source distribution $\mathcal{Q}$ are generated by first drawing $(1-\eta)m$ examples from $N(0, I_{d\times d})$ and labeled according to $\text{sgn}(w_q \cdot x)$ where $\|w_p - w_q\| \leq \epsilon$, for a small value of ϵ. Notice that when ϵ is small, the $(1-\eta)m$ samples are highly relevant for learning the target $\mathcal{P}$. The remaining ηm examples from $\mathcal{Q}$ are all set to a fixed vector u and are labeled as $+1$. These examples represent the noise in $\mathcal{Q}$ and as η increases the presence of such examples makes $\text{dis}(\mathcal{P}, \mathcal{Q})$ larger. In our experiments we set $d = 20, \epsilon = 0.01$, and vary $\eta \in \{0.05, 0.1, 0.15, 0.2\}$.

On the above adaptation problem we evaluate the performance of the previously discussed baselines with our proposed SBEST algorithm implemented via the alternate minimization, SBEST-AM, and the DC-programming algorithms, SBEST-DC, where the loss function considered is the logistic loss and the hypothesis set is the set of linear models with zero bias. For each value of η, the results are averaged over 50 independent runs using the data generation process described above.

Figure C2 shows the performance of the different algorithms for various values of the noise level η and as the number of examples n from the target increases. As can be seen from the figure, both α-reweighting and the baseline that trains solely on $\mathcal{Q}$ degrade significantly in performance as η increases. This is due to the fact the α-reweighting procedure cannot distinguish between non-noisy and noisy data points within the m samples generated from $\mathcal{Q}$.

In Figure C3(Left) we plot the best α chosen by the α-reweighting procedure as a function of n. For reference we also plot the amount of mass on the non-noisy points from $\mathcal{Q}$, i.e., $(1-\eta) \cdot m/(m+n)$. As can be seen from the figure, as n increases the amount of mass selected over the source $\mathcal{Q}$ decreases. Furthermore, as

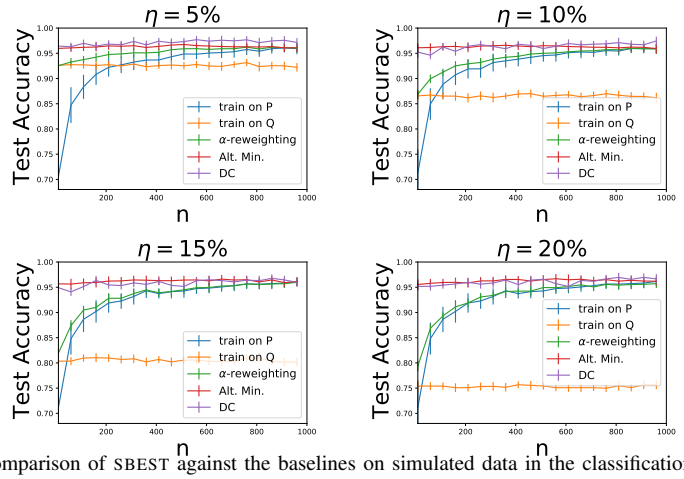

Fig. C2 Comparison of SBEST against the baselines on simulated data in the classification setting. As the noise rate and therefore the discrepancy between $\mathcal{P}$ and $\mathcal{Q}$ increases the performance of the baselines degrades. In contrast, both the alternate minimization and the DC-programming algorithms effectively find a good q-weighting and can adapt to the target.

expected this decrease is sharper as the amount of noise level increases. In particular, α-reweighting is not able to effectively use the non-noisy samples from $\mathcal{Q}$.

On the other hand, both SBEST-AM and SBEST-DC are able to counter the effect of the noise by generating q-weightings that are predominantly supported on the non-noisy samples. In Figure C3(Right) we plot the amount of probability mass that the alternate minimization and the DC-programming implementations of SBEST assign to the noisy data points.

As can be seen from the figure, the total probability mass decreases with n and is also decreasing with the noise levels. These results also demonstrate that our algorithms that compute a good q-weighting can do effective outlier detection since they lead to solutions that assign much smaller mass to the noisy points.

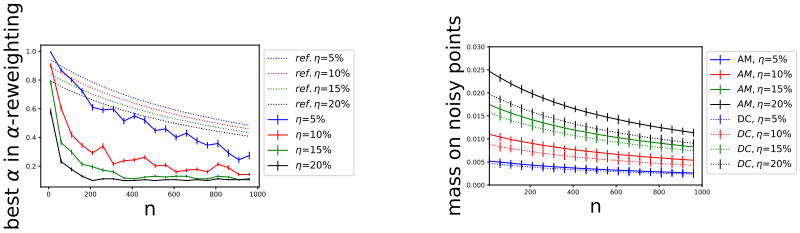

Fig. C3 (Left) Best α chosen by α-reweighting as a function of n. (Right) Total probability mass assigned by SBEST to the noisy points.

C.1.2 Real-world data: classification and regression

Classification Next we evaluate our proposed algorithms and baselines for three real-world datasets obtained from the UCI machine learning repository (Dua and Graff,

2017). We first describe the datasets and our choices of the source and target domains in each case. The first dataset we consider is the `Adult-Income` dataset. This is a classification task where the goal is to predict whether the income of a given individual is greater than or equal to \$50K. The dataset has $32,561$ examples. We form the source domain $\mathcal{Q}$ by taking examples where the attribute *gender* equals 'Male' and the target domain $\mathcal{P}$ corresponds to examples where the *gender* is 'Female'. This leads to $21{,}790$ examples from $\mathcal{Q}$ and $10{,}771$ examples from $\mathcal{P}$.

The second dataset we consider is the `South-German-Credit` dataset. This dataset consists of $1{,}000$ examples and the goal is to predict whether a given individual has good credit or bad credit. We form the source domain $\mathcal{Q}$ by condition on the *residence* attribute and taking all examples where the attribute value is in $\{3, 4\}$ (indicating that the individual has lived at the current residence for 3 or more than 4 years.) The target domain is formed by taking examples where the residence attribute value is in $\{1, 2\}$. This split leads to 620 examples from $\mathcal{Q}$ and 380 training examples from $\mathcal{P}$.

The third dataset we consider is the `Speaker-Accent-Recognition` dataset. In this dataset the goal is to predict the accent of a speaker given the speech signal. We consider the source $\mathcal{Q}$ to be examples where the *accent* is 'US' or 'UK' and the target to be examples where the *accent* is in {'ES', 'FR', 'GE', 'IT'}. This split leads to 150 training examples from $\mathcal{Q}$ and 120 training examples from $\mathcal{P}$.

In each case we randomly split the examples from $\mathcal{P}$ into a training set of 70% examples and a test set of 20% examples. The remaining 10% of the data is used for cross validation. We provide results averaged over 10 such random splits. For the six tasks from the Newsgroups dataset we follow the same methodology as in Wang et al. (2019) to create the tasks.

In each of the above three cases we consider training a logistic regression classifier and compare the performance of SBEST with the baselines that we previously discussed. The results are shown in Table 1 in the main paper.

Regression Next we consider the following regression datasets from the UCI repository.

The `wind` dataset (Haslett and Raftery, 1989) where the task is to predict wind speed from the given features. The source consists of data from months January to November and the target is the data from December. This leads to a total of $5,500$ examples from $\mathcal{Q}$, 350 examples from $\mathcal{P}$ used for training and validation and 200 examples from $\mathcal{P}$ for testing. We create 10 random splits by dividing the 300 examples from $\mathcal{P}$ into a train set of size 150 and a validation set of size 200.

The `airline` dataset is derived from (Ikonomovska, 2009). We create the task of predicting the amount of time the flight is delayed from various features such as the arrival time, distance, whether or not the flight was diverted, and the day of the week. We take a subset of the data for the Chicago O'Haire International Airport (ORD) in 2008. The source and target consists of datat from different hours of the day. This leads to $16,000$ examples from $\mathcal{Q}$ and 500 examples from $\mathcal{P}$ (used as 200 for training and 300 for validation) and 300 examples for testing.

The `gas` dataset (Rodriguez-Lujan et al., 2014; Vergara et al., 2012; Dua and Graff, 2017) where the task is to predict the concentration level from various sensor

Table C1 MSE of the SBEST algorithm against baselines. We report relative errors normalized so that training on target has an MSE of 1.0. Best results or ties in boldface.

Dataset	KMM	DM	SBEST
Wind	1.2 ± 0.04	1.14 ± 0.03	$\mathbf{0.97 \pm 0.02}$
Airline	2.4 ± 0.09	1.72 ± 0.1	$\mathbf{0.952 \pm 0.03}$
Gas	0.41 ± 0.01	0.39 ± 0.01	$\mathbf{0.38 \pm 0.02}$
News	1.08 ± 0.01	1.1 ± 0.01	$\mathbf{0.99 \pm 0.01}$
Traffic	2.1 ± 0.1	2.08 ± 0.08	$\mathbf{0.99 \pm 0.002}$

measurements. The dataset consists of pre-determined batches and we take the first six to be the source and the last batch of size $360,000$ as the target (600 for training and 1000 for validation and 1000 for testing).

The news dataset (Fernandes, 2015; Dua and Graff, 2017) where the goal is to predict the popularity of an article. Our source data consists of articles from Monday to Saturday and the target consists of articles from Sunday. This leads to 32500 examples from the source and 2737 examples for the target (737 for training, 1000 for validation and 1000 for testing).

The traffic dataset from the Minnesota Department of Transportation (Kwon, 2004; Dua and Graff, 2017) where the goal is to predict the traffic volume. We create source and target by splitting based on the time of the day. This leads to 2200 examples from the source and 1000 examples from the test set (200 for training, 400 for validation and 400 for testing).

In each of the datasets above we create 10 random splits based on the shuffling of the training and validation set and report mean and average values over the splits. We compare as baselines the KMM (Huang et al., 2006) algorithm and the DM algorithm (Cortes and Mohri, 2014). Since both the algorithms were originally designed for the setting when the target has no labels we modify them in the following way. We run KMM (DM) on the source vs. target data to get a weight distribution q over the source data. Finally, we perform weighted loss minimization by using the weights in q for the source and uniform $1/n$ weights on the target of size n. The results are shown in Table C1. As can be seen SBEST consistently outperforms the baselines.

C.2 Fine-tuning tasks

In this section we demonstrate the effectiveness of our proposed algorithms for the purpose of fine-tuning pre-trained representations. In the standard pre-training/fine-tuning paradigm (Raffel et al., 2020) a model is first pre-trained on a generalist dataset (which is identified as coming from distribution $\mathcal{Q}$). Once a good representation is learned, the model is then fine-tuned on a task specific dataset (generated from target $\mathcal{P}$). Two of the predominantly used fine-tuning approaches in the literature are *last layer fine-tuning* (Subramanian et al., 2018; Kiros et al., 2015) and *full model fine-tuning* (Howard and Ruder, 2018). In the former approach the representations obtained from the last layer of the pre-trained model are used to train a simple model (often a linear hypothesis) on the data coming from $\mathcal{P}$. In our experiments we fix the choice of the simple model to be a multi-class logistic regression model. In the latter

approach, the model when train on $\mathcal{P}$, is initialized from the pre-trained model and all the parameters of the model are fine-tuned (via gradient descent) on the target distribution $\mathcal{P}$. In this section we explore the additional advantages of combining data from both $\mathcal{P}$ and $\mathcal{Q}$ during the fine-tuning stage via our proposed algorithms. There has been recent interest in carefully combining various tasks/data for the purpose of fine-tuning and avoid the phenomenon of "negative transfer" (Aribandi et al., 2021). Our proposed theoretical results present a principled approach towards this.

To evaluate the effectiveness of our theory for this purpose, we consider the CIFAR-10 vision dataset (Krizhevsky et al., 2009). The dataset consists of 50000 training and 10000 testing examples belonging to 10 classes. We form a pre-training task on data from $\mathcal{Q}$, by combining all the data belonging to classes: {'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog'}. The fine-tuning task consists of data belonging to classes: {'frog', 'horse', 'ship', 'truck'}. We consider both the approaches of last layer fine-tuning and full-model fine-tuning and compare the standard approach of fine-tuning only using data from $\mathcal{P}$ with our proposed algorithms. We use 60% of the data from the source for pre-training, and the remaining 40% is used in fine-tuning.

We split the fine-tuning data from $\mathcal{P}$ randomly into a 70% training set to be used in fine-tuning, 10% for cross validation and and the remaining 20% to be used as a test set. The results are reported over 5 such random splits. We perform pre-training on a standard ResNet-18 architecture (He et al., 2016) by optimizing the cross-entropy loss via the Adam optimizer. As can be seen in Table 2 both gapBoost and SBEST that combine data from $\mathcal{P}$ and $\mathcal{Q}$ lead to a classifier with better performance for the downstream task, however, SBEST clearly outperforms gapBoost.

The second dataset we consider is the `Civil Comments` dataset Pavlopoulos et al. (2020). This dataset consists of text comments in online forums and the goal is to predict whether a given comment is toxic or not. Each data point is also labeled with *identity terms* that describes which subgroup the text in the comment is related to. We create a subsample of the dataset where the target consists of examples from the data points where the identity terms is "asian" and the source is the remaining set of points. This leads to $394,000$ points from the source and $20,000$ points from the target. We create 5 random splits of the data by randomly partitioning the target data into $10,000$ examples for finetuning, 2000 for validation and 8000 for testing. We perform pre-training on a BERT-small model (Devlin et al., 2019) starting from the default checkpoint as obtained from the standard tensorflow implementation of the model.

C.3 Domain adaptation

In this section we evaluate the effectiveness of our proposed BEST-DA objective for adaptation in settings where the target has very little to no labeled data. In order to do this we consider multi-domain sentiment analysis dataset of (Blitzer et al., 2007) that has been used in prior works on domain adaptation. The dataset consists of text reviews associated with a star rating from 1 to 5 for various different categories such as BOOKS, DVD, etc. We specifically consider four categories namely BOOKS, DVD, ELECTRONICS, and KITCHEN. Inspired form the methodology adapted in prior works

(Mohri and Muñoz Medina, 2012; Cortes and Mohri, 2014), for each category, we form a regression task by converting the review text to a 128 dimensional vector and fitting a linear regression model to predict the rating. In order to get the features we first combine all the data from the four tasks and convert the raw text to a TF-IDF representation using scikit-learn's feature extraction library (Pedregosa et al., 2011). Following this, we compute the top 5000 most important features by using scikit-learn's feature selection library, that in turn uses a chi-squared test to perform feature selection. Finally, we project the obtained onto a 128 dimensional space via performing principal component analysis.

After feature extraction, for each task we fit a ridge regression model in the 128 dimensional space to predict the ratings. The predictions of the model are then defined as the ground truth regression labels. Following the above pre-processing we form 12 adaptation problems for each pair of distinct tasks: (TaskA, TaskB) where TaskA, TaskB are in {BOOKS, DVD, ELECTRONICS, KITCHEN}. In each case we form the source domain ($\mathcal{Q}$) by taking 500 labeled samples from TaskA and 200 labeled examples from TaskB. The target ($\mathcal{P}$) is formed by taking 300 unlabeled examples from TaskB. To our knowledge, there exists no principled method for cross-validation in fully unsupervised domain adaptation. Thus, in our adaptation experiments, we used a small labeled validation set of size 50 to determine the parameters for all the algorithms. This is consistent with experimental results reported in prior work (e.g., (Cortes and Mohri, 2014)).

We compare our BEST-DA algorithm with the discrepancy minimization (DM) algorithm of Cortes and Mohri (2014), and the (GDM) algorithm, (Cortes et al., 2019), which is a state of the art adaptation algorithm for regression problems. We also compare with the popular Kernel Mean Matching (KMM) algorithm, (Huang et al., 2006), for domain adaptation. the results averaged over 10 independent source and target splits, where we normalize the mean squared error (MSE) of BEST-DA to be 1.0 and present the relative MSE achieved by the other methods. The results show that in most adaptation problems, BEST-DA outperforms (boldface) or ties with (italics) existing methods.

C.3.1 Domain adaptation – covariate-shift

Here we perform experiments for domain adaptation only under covariate shift and compare the performance of our proposed BEST-DA objective with previous state of the art algorithms. We again consider the multi-domain sentiment analysis dataset (Blitzer et al., 2007) from the previous section and in particular focus on the *books* category. We use the same feature representation as before and define the ground truth as $y = w^* \cdot x + \sigma^2$ where w^* is obtained by fitting a ridge regression classifier. We let the target be the uniform distribution over the entire dataset. We define the source as follows: for a fixed value of ϵ, we pick a random hyperplane w and consider a mixture distribution with mixture weight 0.99 on the set $w \cdot x \geq \epsilon$ and the mixture weight of 0.01 on the set $w \cdot x < \epsilon$. The performance of BEST-DA as compared to DM and KMM is shown in Table C2. As can be seen our proposed algorithm either matches or outperforms current algorithms.

Hyperparameters for the algorithms.

Table C2 MSE achieved by BEST-DA as compared to DM and KMM on the covariate shift task for various values of ϵ.

METHOD	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$	$\epsilon = 1.0$
TRAIN ON Q	0.051 ± 0.001	0.06 ± 0.001	0.06 ± 0.004	0.07 ± 0.006	0.073 ± 0.002	0.073 ± 0.005
KMM	$0.05 \pm 1e-4$	$0.05 \pm 1e-4$	$0.05 \pm 3e-4$	$0.06 \pm 1e-4$	$0.06 \pm 1e-4$	$0.07 \pm 2e-4$
DM	0.02 ± 0.005	0.06 ± 0.003	0.05 ± 0.003	0.05 ± 0.001	0.06 ± 0.005	0.06 ± 0.003
BEST-DA	0.01 ± 0.006	0.02 ± 0.006	0.027 ± 0.005	0.04 ± 0.004	0.04 ± 0.007	0.04 ± 0.004

For our proposed SBEST and SBEST-DA algorithms the hyperparameters $\lambda_\infty, \lambda_1, \lambda_2$ were chosen via cross validation in the range $\{1e-3, 1e-2, 1e-1\} \cup \{0, 1, 2, \ldots, 10\} \cup \{0, 1000, 2000, 10000, 50000, 100000\}$. The h optimization step of alternate minimization was performed using sklearn's linear regression/logistic regression methods (Pedregosa et al., 2011). During full layer fine-tuning on ResNet/BERT models we use the Adam optimizer for the h optimization step with a learning rate of $1e-3$ used for the CIFAR-10 dataset and a learning rate of $1e-5$ for the BERT-small models.

For the q optimization we used projected gradient descent and the step size was chosen via cross validation in the range $\{1e-3, 1e-2, 1e-1\}$.

We re-implemented the gapBoost algorithm (Wang et al., 2019) in Python. Following the prescription by the authors of gapBoost we set the parameter $\gamma = 1/n$ where n is the size of the target. We tune parameters ρ_S, ρ_T in the range $\{0.1, 0.2, \ldots, 1\}$ and the number of rounds of boosting in the range $\{5, 10, 15, 20\}$. We also re-implemented baselines DM (Cortes and Mohri, 2014) and the GDM algorithm (Cortes et al., 2019). These DM algorithm was implemented via gradient descent and the second stage of the GDM algorithm was implemented via alternate minimization. The learning rates in each case searched in the range $\{1e-3, 1e-2, 1e-1\}$ and the regularization parameters were searched in the range $\{1e-3, 1e-2, 1e-1, 0, 10, 100\}$. The radius parameter for GDM was searched in the range $[0.01, 1]$ in steps of 0.01. In line with our proposed algorithms, all baselines were implemented without incorporating a bias term.

To our knowledge, there exists no principled method for cross-validation in fully unsupervised domain adaptation. Thus, in our unsupervised adaptation experiments, we used a small labeled validation set of size 50 to determine the parameters for all the algorithms. This is consistent with experimental results reported in prior work (Cortes and Mohri, 2014; Cortes et al., 2019).