ODD: Overlap-aware Estimation of Model Performance under Distribution Shift

Aayush Mishra[1] **Anqi Liu**[1]

[1]Department of Computer Science,
Johns Hopkins University, Baltimore, Maryland, USA
{amishr24, aliu.cs}@jhu.edu

Abstract

Reliable and accurate estimation of the error of an ML model in unseen test domains is an important problem for safe intelligent systems. Prior work uses *disagreement discrepancy* (Dis^2) to derive practical error bounds under distribution shifts. It optimizes for a maximally disagreeing classifier on the target domain to bound the error of a given source classifier. Although this approach offers a reliable and competitively accurate estimate of the target error, we identify a problem in this approach which causes the disagreement discrepancy objective to compete in the overlapping region between source and target domains. With an intuitive assumption that the target disagreement should be no more than the source disagreement in the overlapping region due to high enough support, we devise Overlap-aware Disagreement Discrepancy (ODD). Our ODD-based bound uses domain-classifiers to estimate domain-overlap and better predicts target performance than Dis^2. We conduct experiments on a wide array of benchmarks to show that our method improves the overall performance-estimation error while remaining valid and reliable. Our code and results are available on GitHub.

1 INTRODUCTION

The ability of machine learning models to know when they do not know, is an important characteristic to make them safe for deployment, especially in high-stakes applications like the medical domain. Modern neural networks have an incredible capacity to learn complex functions but they are prone to catastrophic errors on inputs outside of their training distributions. Hence, it is crucial to be able to predict the performance of these models on distributionally shifted test domains.

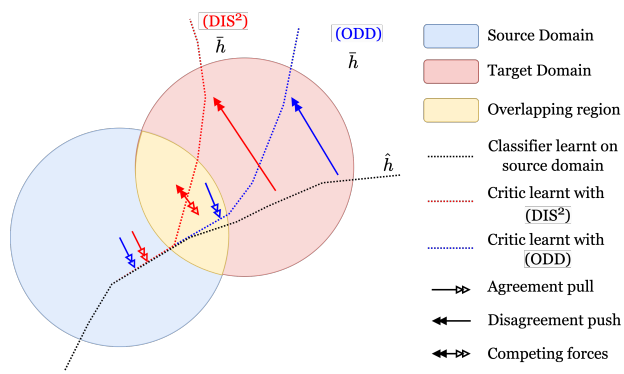

Figure 1: Given a source domain $\mathcal{S}$, and a classifier $\hat{h}$ trained in this domain, we aim to predict the performance of $\hat{h}$ in a target domain $\mathcal{T}$ *without labels*. Rosenfeld and Garg [2023] learn a worst-case *critic* $\bar{h}$, which maximally disagrees with $\hat{h}$ in $\mathcal{T}$ while agreeing with $\hat{h}$ in $\mathcal{S}$ (Dis^2), to bound this performance. But this optimization leads to a competition in the overlapping region. ODD discounts disagreement in the overlapping region making $\bar{h}$ agree with $\hat{h}$ in this region. We show that ODD tightens the gap between true and predicted performance compared to Dis^2, while remaining reliable.

Many works have attempted to address this problem of test performance prediction [Ben-David et al., 2006, 2010a, Mansour et al., 2009, Ben-David et al., 2010b]. These methods provide uniform convergence bounds which are often vacuous in practice because we are interested in bounding the error of a single hypothesis. Other recent works attempt to use unlabeled test samples to make point-wise (per hypothesis) prediction of performance bounds in neural networks [Lu et al., 2023, Baek et al., 2022, Garg et al., 2022, Guillory et al., 2021]. Although these methods achieve low performance prediction error on average, they lack reliability and often overestimate the performance, especially under large distribution shift (when reliability is most important). To address these challenges, Rosenfeld and Garg [2023] proposed Disagreement Discrepancy (Dis^2), which provides almost provable performance bounds using unlabeled test samples.

It trains a *critic* that minimizes its disagreement with a given classifier on the source distribution while maximizing its disagreement on the target domain under a sufficiently expressive hypothesis class. This allows DIS^2 to assume a worst-case shift in the target domain while remaining faithful to the trained classifier in the source domain. This approach gives non-vacuous bounds that almost always remain valid.

In this work, we find that DIS^2 can be further improved using domain-overlap awareness. The agreement with source and disagreement with target creates a tension in the region of domain overlap, resulting in unstable optimization, which leads to more pessimistic critics than necessary. We are motivated by the intuition that if a classifier was trained on labeled samples from the target domain, its support in the overlapping region would be similar to what the source trained classifier had. Hence, we propose to discount the disagreement of the critic with target overlapping samples through a new training objective. Leveraging domain classifiers for estimate domain-overlap, our Overlap-aware Disagreement Discrepancy (ODD) removes the instability of optimization in the overlapping region and results in practically tighter performance bounds than DIS^2. We visually illustrate our method in Figure 1. In summary:

- We derive a general version of the DIS^2 bound using the notion of ideal joint hypothesis, and show that theoretical performance bounds obtained using ODD are as tight as those obtained using DIS^2.
- We design an ODD-based disagreement loss to find more *optimistic critics* for practically estimating the performance bounds.
- On several real datasets and training method combinations, ODD-based bounds are tighter than DIS^2-based bounds, and still maintain reliable coverage.

2 THE GENERAL DIS^2 BOUND

Setup: We follow the notation setup used by Rosenfeld and Garg [2023]. Let $\mathcal{S}, \mathcal{T}$ denote the source and target distributions, respectively, over labeled inputs $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and let $\hat{\mathcal{S}}, \hat{\mathcal{T}}$ denote the empirically observed sets (with cardinality n_S and n_T) of samples from these distributions. *In the target domain $\mathcal{T}$, we observe only the covariates and not the labels.* We let $p_{\mathcal{S}}$ and $p_{\mathcal{T}}$ denote the marginal distributions of covariates in the source and target domains. We consider classifiers $h : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ which output a vector of logits ($h(x)_y$ denotes the logits of class y in this vector), and let $\hat{h}$ denote the particular classifier (trained on $\hat{\mathcal{S}}$) under study for which we want to bound the error in the target domain. We use $\mathcal{H}$ to denote a hypothesis class of such classifiers. For a domain $\mathcal{D}$ on $\mathcal{X}$, let $\epsilon_{\mathcal{D}}(h, h') := \mathbb{E}_{x \sim p_{\mathcal{D}}}[\mathbf{1}\{\arg\max_y h(x)_y \neq \arg\max_y h'(x)_y\}]$ denote the one-hot disagreement between any classifiers h and h' on $\mathcal{D}$.

The labeling function: Let $y^*_{\mathcal{S}}$ and $y^*_{\mathcal{T}}$ denote the true labeling functions in the source and target domains, respectively. Rosenfeld and Garg [2023] assume a fixed y^* to represent the true labeling function for all domains. In contrast, we use the notion of ideal joint hypothesis [Ben-David et al., 2010a] to analyze *adaptability* in the general case.

Definition 2.1. The ideal joint hypothesis is defined as:

$$y^* := \arg\min_{h \in \mathcal{H}} \epsilon_{\mathcal{S}}(h, y^*_{\mathcal{S}}) + \epsilon_{\mathcal{T}}(h, y^*_{\mathcal{T}}) \tag{1}$$

with the corresponding joint risk defined as $\lambda := \epsilon_{\mathcal{S}}(y^*) + \epsilon_{\mathcal{T}}(y^*)$. For brevity, we overload $\epsilon_{\mathcal{D}}(h)$ to mean $\epsilon_{\mathcal{D}}(h, y^*_{\mathcal{D}})$, i.e. the 0-1 error of classifier h on distribution $\mathcal{D}$. If λ is high, no h trained on the source domain can be expected to perform well on the target domain (low adaptability).

Now, Rosenfeld and Garg [2023] define *disagreement discrepancy* (DIS^2) as follows.

Definition 2.2. DIS^2 $\Delta(h, h')$ is the disagreement between any h and h' on $\mathcal{T}$ minus their disagreement on $\mathcal{S}$:

$$\Delta(h, h') := \epsilon_{\mathcal{T}}(h, h') - \epsilon_{\mathcal{S}}(h, h'). \tag{2}$$

This immediately implies the following lemma:

Lemma 2.3. *For any classifier h,*
$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \Delta(h, y^*) + \lambda$.

The proof can be found in Appendix A. This gives us a method to bound the target risk in terms of source risk and the discrepancy term Δ. However, as y^* is unknown, we can optimize for an alternate critic $\bar{h}$ which would act as a proxy for the worst-case y^*. For this, Rosenfeld and Garg [2023] assume the following:

Assumption 2.4. For a critic $\bar{h} \in \mathcal{H}$ which maximizes a concave surrogate to the empirical DIS^2, $\Delta(\hat{h}, y^*) \leq \Delta(\hat{h}, \bar{h})$.

This assumption is based on the practical observation that y^* is not chosen adversarially with respect to $\hat{h}$. So it is reasonable that there exists another function $h^* \in \mathcal{H}$ having higher disagreement discrepancy than y^*. $\bar{h}$ is a concave surrogate to h^* that we can find empirically (which approaches h^* with increasing sample size). Finally, we have the error bound:

Theorem 2.5 (DIS^2 Bound)**.** *Under Assumption 2.4, with probability $\geq 1 - \delta$,*

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\hat{\mathcal{S}}}(\hat{h}) + \hat{\Delta}(\hat{h}, \bar{h}) + \sqrt{\frac{(n_S + 4n_T)\log\frac{1}{\delta}}{2n_S n_T}} + \lambda.$$

Here, $\epsilon_{\hat{\mathcal{S}}}$ and $\hat{\Delta}$ denote the empirical versions of their corresponding population terms. The proof can be found in Appendix A. Even with a suitable $\bar{h}$, calculating λ still requires

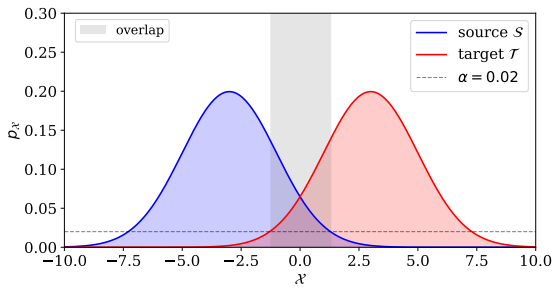

Figure 2: For source ($\mathcal{S}$) and target ($\mathcal{T}$) domains with $p_{\mathcal{S}} = \mathcal{N}(-3, 2)$ and $p_{\mathcal{T}} = \mathcal{N}(3, 2)$ respectively, the domain overlapping region using $\alpha = 0.02$ is shown.

access to target labels, which we do not have. Hence, we follow Rosenfeld and Garg [2023] and fix a common y^* across domains (as it happens in most practical cases). When $y^* = y^*_{\mathcal{S}} = y^*_{\mathcal{T}}$, $\lambda = 0$ and the above bound collapses exactly to their bound. In the next section, we make DIS^2 more effective with overlap awareness.

3 ODD: OVERLAP-AWARE DIS^2

The problem with disagreement discrepancy: Maximizing DIS^2 with respect to a given classifier $\hat{h}$ loosely translates to finding a *critic* $\bar{h}$ which maximally disagrees with $\hat{h}$ in the target domain (first term) while maximally agreeing with $\hat{h}$ in the source domain (second term). This works as expected in the case when the source and target domains are disjoint. However, there is often an overlap between source and target domains in practice. Maximizing DIS^2 creates a competition between the two terms in the region of *domain overlap*. In this region, the maximizing hypothesis has to both agree and disagree with $\hat{h}$. Hence, by definition, the DIS^2 objective is unstable in the overlapping region, where both agreement and disagreement can result in the same discrepancy value. This presents a problem during practical optimization and can result in over-conservative critics which disagree with $\hat{h}$ more than necessary.

Domain overlap: To tackle this problem, we first define domain overlap using a parameter α, which quantifies the minimum density requirement in both domains of interest for an element to be considered inside the overlapping region.

Definition 3.1. The *domain overlap* set between $\mathcal{S}$ and $\mathcal{T}$ is defined as $\mathcal{D}_\alpha = \{x \in \mathcal{X} : p_{\mathcal{S}}(x) > \alpha, p_{\mathcal{T}}(x) > \alpha\}$ for some $\alpha > 0$.

Probability distributions (like the normal distribution), may have non-zero support everywhere, and $\alpha = 0$ makes the entire domain space part of the overlapping region. Practically, an overlap is meaningful only in a region with reasonably high support in domains of interest (see Figure 2).

Now, let $\mathcal{K}$ denote $[\mathbf{1}\{\arg\max_y h(x)_y \neq \arg\max_y h'(x)_y\}]$. Then, $\epsilon_{\mathcal{D}}(h, h') := \mathbb{E}_{x\sim p_{\mathcal{D}}}\mathcal{K} = \int_{x\in\mathcal{D}} p_{\mathcal{D}}(x)\mathcal{K}dx$. This explicit definition of expectation as an integral allows us to think of the density $p_{\mathcal{D}}$ and domain of application $\mathcal{D}$ independently. If the domain $\mathcal{D}$ is split into two subsets, say $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{A} \cap \mathcal{B} = \phi$ and $\mathcal{A} \cup \mathcal{B} = \mathcal{D}$, then $\epsilon_{\mathcal{D}}(h, h') = \int_{x\in\mathcal{A}} p_{\mathcal{D}}(x)\mathcal{K}dx + \int_{x\in\mathcal{B}} p_{\mathcal{D}}(x)\mathcal{K}dx$. We use this separability under the same density as a property to define $\epsilon^{\mathcal{D}}_{\mathcal{A}}(h, h') = \int_{x\in\mathcal{A}} p_{\mathcal{D}}(x)\mathcal{K}dx$. Using this definition, we have:

$$\epsilon_{\mathcal{T}}(h, h') := \underbrace{\epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(h, h')}_{\text{Overlap Disagreement}} + \underbrace{\epsilon^{\mathcal{T}}_{\mathcal{T}\setminus\mathcal{D}_\alpha}(h, h')}_{\text{Non-Overlap Disagreement}}$$

This breakdown of disagreement in the overlapping and non-overlapping portions allows us rethink the DIS^2 objective.

Overlap-aware Disagreement Discrepancy (ODD): As $\hat{h}$ is trained to minimize the risk in $\mathcal{S}$, its disagreement with y^* ($=y^*_{\mathcal{S}}$) in regions with high support is expected to be low. This is why the DIS^2 objective aims to minimize the disagreement between the critic $\bar{h}$ and $\hat{h}$ in the source domain. However, DIS^2 also tries to maximize the disagreement in $\mathcal{D}_\alpha$, which is a part of the same high source support region. We address this issue by making DIS^2 overlap aware.

Definition 3.2. The ***O**verlap-aware **D**isagreement **D**iscrepancy (ODD)* is defined as the disagreement between any h and h' in the *non-overlapping* region of $\mathcal{T}$ minus their disagreement in the *non-overlapping* region of $\mathcal{S}$:

$$\Delta(h, h', \alpha) := \epsilon^{\mathcal{T}}_{\mathcal{T}\setminus\mathcal{D}_\alpha}(h, h') - \epsilon^{\mathcal{S}}_{\mathcal{S}\setminus\mathcal{D}_\alpha}(h, h'). \tag{3}$$

We also have the corresponding *overlap discrepancy*:

Definition 3.3. $\underline{\Delta}(h, h', \alpha) = \epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(h, h') - \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(h, h')$.

Note that, $\Delta(h, h')(\text{DIS}^2) = \Delta(h, h', \alpha)(\text{ODD}) + \underline{\Delta}(h, h', \alpha)$. With our separation of overlapping and non-overlapping discrepancies, we can rewrite Theorem 2.5 as:

Theorem 3.4. *Under Assumption 2.4, with probability* $\geq 1 - \delta$,

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\hat{\mathcal{S}}}(\hat{h}) + \hat{\Delta}(\hat{h}, \bar{h}, \alpha) + \hat{\underline{\Delta}}(\hat{h}, \bar{h}, \alpha) + \sqrt{\frac{(n_S + 4n_T)\log\frac{1}{\delta}}{2n_S n_T}} + \lambda.$$

Remember, $\bar{h}$ is the critic from Assumption 2.4. This is the most general form of our analysis, where terms with ($\hat{}$) denote the empirical counterparts of their corresponding population terms. As we increase α, we decrease the size of $\mathcal{D}_\alpha$. Starting from $\alpha = 0$ when $\mathcal{D}_\alpha = \mathcal{S} \cup \mathcal{T}$, until we reach some high $\alpha = \alpha_0$ for which $\mathcal{D}_\alpha = \phi$. The notion of overlap becomes interesting for some intermediate α values, for which samples from either domain behave similarly on

learning algorithms. Next, we discuss why the $\underline{\Delta}$ term is expected to be small due to domain overlap.

Overlap Discrepancy between $\hat{h}$ and y^*: Of all $h \in \mathcal{H}$, we usually only care to examine ones which are trained to achieve low source risk, i.e. $\hat{h}$ typically has low disagreement with y^* in regions of high support. For a reasonably chosen $\mathcal{D}_\alpha$, we have similarly high support from the target domain. A classifier trained with ground truth target labels should be expected to have similar disagreement in this region. In fact, there is no reason to expect $\epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(\hat{h}, y^*)$ to be any higher than $\epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, y^*)$. Hence, we make the following assumption:

Assumption 3.5. Target disagreement between $\hat{h}$ and y^* is bounded by the source disagreement in the region of domain overlap, i.e., $\epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(\hat{h}, y^*) \leq \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, y^*)$.

In our experiments with randomly generated datasets(Figure 4), we show how the overlap discrepancy $(\epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(\hat{h}, y^*) - \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, y^*))$ consistently remains close to zero, supporting our intuition. With Assumption 3.5 and a common y^* across domains we can derive a practically tighter bound by maximizing ODD. We change Assumption 2.4 to apply only in the non-overlapping region:

Assumption 3.6. For an $\bar{h} \in \mathcal{H}$ which maximizes a concave surrogate to the empirical ODD, $\Delta(\hat{h}, y^*, \alpha) \leq \Delta(\hat{h}, \bar{h}, \alpha)$.

Hence, Theorem 3.4 is simplified to:

Theorem 3.7 (ODD Bound)**.** *Under Assumption 3.6 and Assumption 3.5, with probability $\geq 1 - \delta$,*

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\hat{\mathcal{S}}}(\hat{h}) + \hat{\Delta}(\hat{h}, \bar{h}, \alpha) + \sqrt{\frac{(n_S + 4n_T)\log\frac{1}{\delta}}{2n_S n_T}}$$

We discuss this derivation in Appendix A. Note that this bound does not theoretically improve upon the DIS^2 based bound from [Rosenfeld and Garg, 2023] for a given critic $\bar{h}$. However, optimizing for ODD allows for the selection of a better critic, as we will see in the next section, which agrees more with $\hat{h}$ in the overlapping source domain. This better critic results in the ODD term being lower than the original DIS^2 term, making the bound practically tighter.

4 MAXIMIZING ODD

Competition in the overlapping region: The *disagreement logistic loss* of a classifier h on a labelled sample (x, y) is described in [Rosenfeld and Garg, 2023] as:

$$\ell_{\text{dis}}(h, x, y) := \frac{1}{\log 2}\log\left(1 + \exp\left(h(x)_y - \frac{1}{|\mathcal{Y}| - 1}\sum_{\hat{y} \neq y} h(x)_{\hat{y}}\right)\right).$$

ℓ_{dis} is shown to be convex in $h(x)$ and to upper bound the 0-1 disagreement loss. To maximize the empirical disagreement discrepancy, a combined loss on source and target samples is used:

$$\hat{\mathcal{L}}_\Delta(\bar{h}) := \frac{1}{|\hat{\mathcal{S}}|}\sum_{x \in \hat{\mathcal{S}}} \ell_{\text{logistic}}(\bar{h}, x, \hat{h}(x)) + \frac{1}{|\hat{\mathcal{T}}|}\sum_{x \in \hat{\mathcal{T}}} \ell_{\text{dis}}(\bar{h}, x, \hat{h}(x)).$$

Here, $\ell_{\text{logistic}} := -\frac{1}{\log|\mathcal{Y}|}\log\text{softmax}(h(x))_y$ is the standard log-loss. This is where we see the competition: for points that lie in the overlapping region, one would contribute by reducing ℓ_{logistic}, making $\bar{h}$ agree with $\hat{h}$ while the other would contribute by reducing ℓ_{dis}, making $\bar{h}$ disagree with $\hat{h}$. In practice, this typically results in an $\bar{h}$ that agrees with $\hat{h}$ a little less than it could have in the overlapping region. We illustrate this phenomena in Figure 4 and show how our proposed method improves on it, which we describe next.

Mitigating the competition: We propose to improve this loss by discounting the disagreement of target samples in the overlapping region, i.e.,

$$\hat{\mathcal{L}}_\Delta(\bar{h}, \alpha) := \frac{1}{|\hat{\mathcal{S}}|}\sum_{x \in \hat{\mathcal{S}}} \ell_{\text{logistic}}(\bar{h}, x, \hat{h}(x)) + \frac{1}{|\hat{\mathcal{T}}|}\sum_{x \in \hat{\mathcal{T}}} \mathbf{1}\left\{x \notin (\hat{\mathcal{D}}_\alpha)\right\} * \ell_{\text{dis}}(\bar{h}, x, \hat{h}(x)).$$

Here, $\mathbf{1}\left\{x \notin (\hat{\mathcal{D}}_\alpha)\right\}$ is an indicator function which only selects samples that are outside the overlapping set. For samples in the overlapping set, the ℓ_{dis} term is discounted to zero, encouraging the maximizing $\bar{h}$ to not disagree with $\hat{h}$ in this region. Note that this loss still maximizes the empirical ODD, as the disagreement in the non-overlapping target region is still maximized and the disagreement in the non-overlapping source region is still minimized.

How to select $\mathcal{D}_\alpha$? The ideal way to find the overlap set $\mathcal{D}_\alpha$ is to find the densities $p_{\mathcal{S}}$ and $p_{\mathcal{T}}$, and tune for the smallest α on a held-out validation set which satisfies Assumption 3.5. However, density estimation is typically difficult with high-dimensional data and accessing a held-out set may be challenging in many domains where data is scarce (where reliable performance estimation is the most essential). We propose to do it with a simple method of domain classification. If there is overlap between the source and target domains, a domain classifier should have difficulty in classifying samples from this overlapping region. As we already know domain labels, we use the domain classifier's prediction as a proxy to quantify the extent of overlap.

Suppose a domain classifier $d : \mathcal{X} \to \{0, 1\}$ is trained to classify all source samples with the label 0 and all target samples with the label 1. It produces logits $\{d(x)_0, d(x)_1\}$

Algorithm 1 Finding the critic $\bar{h}$ for a given $\hat{h}$

Input: $\hat{h}, \hat{\mathcal{S}} = \{(x_i^{\mathcal{S}}, y_i^*)\}_{i=1}^{n_{\mathcal{S}}}, \hat{\mathcal{T}} = \{(x_i^{\mathcal{T}})\}_{i=1}^{n_{\mathcal{T}}}$
$d \leftarrow \texttt{train}(\{(x_i^{\mathcal{S}}, 0)\}_{i=1}^{n_{\mathcal{S}}} \cup \{(x_i^{\mathcal{T}}, 1)\}_{i=1}^{n_{\mathcal{T}}})$ ▹ Train domain classifier.
$s \leftarrow \texttt{softmax}(d)$
$p^{\hat{h}} \leftarrow \arg\max \hat{h}$ ▹ Returns the $\hat{h}$ predicted label.
$\bar{h} \leftarrow \hat{h}$ ▹ Initialize critic with $\hat{h}$ parameters.
$\Theta \in \bar{h} \leftarrow$ require grad ▹ Choose parameters to learn.
while not converged **do**
  $\mathcal{S}_{\text{loss}} \leftarrow \texttt{mean}(\{\ell_{\text{logistic}}((x_i^{\mathcal{S}}, p^{\hat{h}}(x_i^{\mathcal{S}})))\}_{i=1}^{n_{\mathcal{S}}})$
  $\mathcal{T}_{\text{loss}} \leftarrow \texttt{mean}(\{s(x_i^{\mathcal{T}})_1 * \ell_{\text{dis}}((x_i^{\mathcal{T}}, p^{\hat{h}}(x_i^{\mathcal{T}})))\}_{i=1}^{n_{\mathcal{T}}})$
  $\Theta'_t \leftarrow \Theta'_{t-1} - \frac{\partial(\mathcal{S}_{\text{loss}} + \mathcal{T}_{\text{loss}})}{\partial \Theta}$ ▹ Update $\bar{h}$ parameters.
end while
return $\bar{h}$

for all samples and taking the arg max gives us the classified domain. We can replace the indicator function $\mathbf{1}\left\{x \notin (\hat{\mathcal{D}}_\alpha)\right\}$ with $\mathbf{1}\left\{\arg\max(d(x)) \neq 1)\right\}$ to find target domain samples which are being classified as source samples, and hence have an estimate of the set $\mathcal{D}_\alpha$. Note that this assumes our domain classifier model is accurate. In practice, we can convert this problem of hard selection of the set $\mathcal{D}_\alpha$ into a soft version, by simply weighting each sample in the target domain by its probability of being classified in the target domain. Let $\{s(x)_0, s(x)_1\}$ denote the soft-max of $\{d(x)_0, d(x)_1\}$. Then, the indicator function $\mathbf{1}\left\{x \notin (\hat{\mathcal{D}}_\alpha)\right\}$ can be replaced with $s(x)_1$. Although the soft-max probabilities are not guaranteed to be calibrated without the use of a held-out calibration set, this soft version makes the discounting of overlapping target samples smooth and it works well in practice. Hence, we use the following loss to maximize ODD while making $\bar{h}$ agree with $\hat{h}$ in the overlapping region:

$$\hat{\mathcal{L}}_\Delta(\bar{h}, \alpha) := \frac{1}{|\hat{\mathcal{S}}|} \sum_{x \in \hat{\mathcal{S}}} \ell_{\text{logistic}}(\bar{h}, x, \hat{h}(x)) + \frac{1}{|\hat{\mathcal{T}}|} \sum_{x \in \hat{\mathcal{T}}} s(x)_1 * \ell_{\text{dis}}(\bar{h}, x, \hat{h}(x)).$$

We describe our method step by step in Algorithm 1.

Why not also discount source overlapping samples? We want $\bar{h}$ to agree with $\hat{h}$ in this region, which includes both source and target samples. Discounting loss from both overlapping source samples and overlapping target samples would lead to a problem similar to that of $\hat{\mathcal{L}}_\Delta(\bar{h})$, where instead of competition, the optimization would be indifferent to all the samples in the overlapping region. In Appendix B, we conduct experiments to validate this and discuss how this may lead to worse critics.

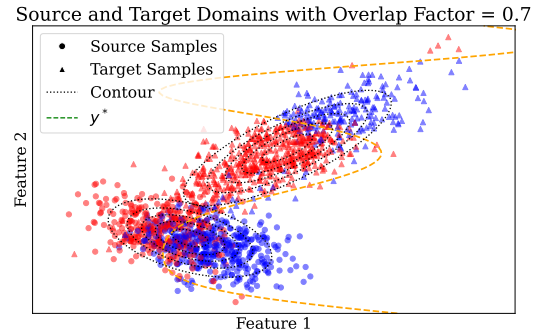

Figure 3: Sample 2-D synthetic dataset used for analysis. With a complex y^*, the model learned on source data can make errors in the target domain, but the overlapping region has high agreement.

5 EXPERIMENTAL RESULTS

5.1 SYNTHETIC DATA EXPERIMENTS

Dataset: To visually illustrate the benefit of our overlap-aware disagreement discrepancy, we conduct experiments on 2-D synthetic data. First, we sample source and target domains from 2 randomly initialized Gaussian distributions with a randomly sampled overlap factor, which determines how close their means are. Then, we initialize a complex decision surface y^* which determines the labels for each sample (with some noise). One sample dataset can be seen in Figure 3. Then, we train a model $\hat{h}$ on the source data. Finally, we use Algorithm 1 to find $\bar{h}$ to obtain a bound on the target performance (accuracy). For comparison and analysis, we also use DIS^2 to obtain the same bound. The details of the dataset generation can be found in Appendix B.

Methodology: We randomly sample an overlap rate 100 times between $[0, 1]$ and generate a new dataset for each (around 2000 training samples and 1250 validation samples in each domain). We repeat the experiment 40 times, effectively creating 4000 unique datasets, to smoothen any finite sampling issues. We use a small 3−layer, 16−neurons wide MLP to train the $\hat{h}$, $\bar{h}$ and d. Finally, we distribute the 4000 overlap rates into 20 equal width bins to average our findings across multiple datasets. Instead of predicting the error bound (as in Theorem 3.7), we predict the accuracy bound to follow the convention in Rosenfeld and Garg [2023]. Hence, a tighter bound implies a larger lower bound in accuracy.

Findings—why ODD is better than DIS^2: In Figure 4, we show the relationship between various metrics (discrepancy, target accuracy, predicted target accuracy (the bound) and individual domain agreements) with the overlap rate (for both ODD and DIS^2). The target performance prediction is calculated as $\epsilon_{\mathcal{S}}(\hat{h}, y^*) - \Delta(\hat{h}, \bar{h}) - \varepsilon$ for DIS^2, and $\epsilon_{\mathcal{S}}(\hat{h}, y^*) - \Delta(\hat{h}, \bar{h}, \alpha) - \varepsilon$ for ODD (ε is the concentration term in Theorem 3.7). We find that:

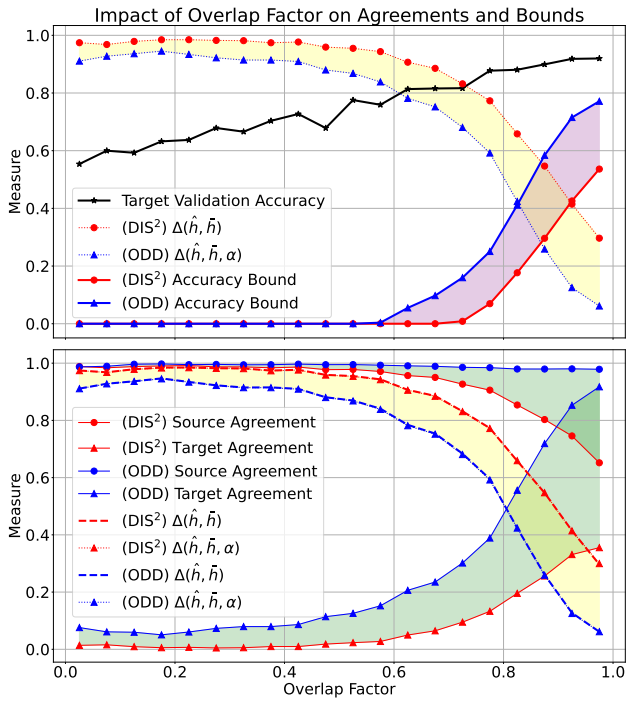

Figure 4: **TOP:** As the overlap factor increases, ODD becomes increasingly smaller than DIS^2 (yellow shaded region), resulting in a tighter bound (purple shaded region) that is closer to the target accuracy. **BOTTOM:** We decompose the discrepancy into individual source and target agreements. Due to the competition in the overlapping region, maximizing DIS^2 makes $\bar{h}$ start disagreeing with $\hat{h}$ in the source domain, while maximizing ODD maintains agreement. ODD also allows for much higher target agreement in high overlap cases. Both shown with green shaded region. This results in a less conservative $\bar{h}$. Moreover, we find that $\Delta(\hat{h}, \bar{h}) \approx \Delta(\hat{h}, \bar{h}, \alpha)$ for both ODD and DIS^2 based critics (difference $< 0.01\%$ making plot-lines coincide in this figure), as predicted by our theory.

- As the overlap increases, DIS^2-based critics start disagreeing with $\hat{h}$ in the source domain, while ODD-based critics maintain their source agreement.
- Even with high overlap, DIS^2-based critics disagree substantially with $\hat{h}$ in the target domain, making them overly pessimistic. ODD fixes this by increasing target agreement in high overlap regions.
- For both DIS^2 and ODD based critics, overall discrepancy $\Delta(\hat{h}, \bar{h})$ is approximately equal to non-overlapping discrepancy $\Delta(\hat{h}, \bar{h}, \alpha)$, validating our assumption and theory in section 3. This shows that *ODD improves the bound by selecting a better critic*, not by reducing the disagreement for a given critic.

ODD consistently improves over DIS^2 by choosing a more optimistic $\bar{h}$ that agrees more with $\hat{h}$ in the overlapping region, while remaining valid and reliable. It makes sure that the critic is chosen as optimistically as possible based on the overlap, without compromising disagreement in the unseen target region.

5.2 REAL DATA EXPERIMENTS

Datasets: As we aim to achieve an improvement of DIS^2 [Rosenfeld and Garg, 2023], we follow their setup and conduct experiments across all vision benchmark datasets used in their study for a fair comparison between DIS^2 and ODD. These include four BREEDs datasets [Santurkar et al., 2021]: Entity13, Entity30, Nonliving26, and Living17; FMoW [Christie et al., 2018] from WILDS [Koh et al., 2021]; Officehome [Venkateswara et al., 2017]; Visda [Peng et al., 2018, 2017]; CIFAR10, CIFAR100 [Krizhevsky and Hinton, 2009]; and Domainet [Peng et al., 2019]. In addition, we conduct additional experiments on a language dataset: CivilComments [Borkan et al., 2019] for breadth of applicable domains. These datasets contain multiple domains and include a wide variety of subpopulation and natural distribution shifts. We present some details about these datasets and their shift variations, and complete details about the new CivilComments dataset in Appendix B. Comprehensive details can be found in the DIS^2 paper.

Methods and baselines: Models are trained with ERM and Unsupervised Domain Adaptation methods (which help improve target performance with unlabeled target data) like FixMatch [Sohn et al., 2020], DANN [Ganin et al., 2016], CDAN [Long et al., 2018], and BN-adapt [Li et al., 2017]). Although our paper is positioned as an improvement to DIS^2, we also present comparisons with other baselines like Average Confidence (AC) [Guo et al., 2017], Difference of Confidences (DoC) [Guillory et al., 2021], Average Thresholded Confidence (ATC) [Garg et al., 2022], and Confidence Optimal Transport (COT) [Lu et al., 2023], for completeness. All methods are calibrated with temperature scaling [Guo et al., 2017] using source validation data. As our general setup is identical, we follow the implementation details in the DIS^2 paper. We list experimental details on training the domain classifiers in Appendix B. For the correction term in Theorem 3.7, we use a small $\delta = 0.01$ (same as DIS^2) in all our experiments.

Bound calculation strategies: As pointed out by Rosenfeld and Garg [2023], the value of the error bound is expected to decrease if the critic is searched for in a restricted hypothesis class. Their real-data experiments are performed on the deep features of the inputs, as classifiers trained on these features have been shown to have the capacity to generalize under distribution shifts [Rosenfeld et al., 2022, Kirichenko et al., 2023]. This makes $\hat{h}$ belong to the linear hypothesis class. Even logits of the source classifier may contain information necessary for this task as prior work suggests that deep network representations have small effective ranks [Arora et al., 2018, Huh et al., 2024, Pezeshki et al., 2021]. Therefore, we

	MAE (↓)		Coverage (↑)		Overest. (↓)	
DA?	✗	✓	✗	✓	✗	✓
Prediction Method						
AC [Guo et al., 2017]	0.1086	0.1091	0.0989	0.0333	0.1187	0.1123
DoC [Guillory et al., 2021]	0.1052	0.1083	0.1648	0.0167	0.1230	0.1095
ATC NE [Garg et al., 2022]	0.0663	0.0830	0.2857	0.2000	0.0820	0.1007
COT [Lu et al., 2023]	0.0695	0.0808	0.2528	0.1833	0.0858	0.0967
Dis^2 [Rosenfeld and Garg, 2023]						
Using Features	0.2841	0.1886	1.0000	1.0000	0.0000	0.0000
Using Logits	0.1525	0.0881	0.9890	0.7500	0.0171	0.0497
Using Logits w/o δ term	0.0996	0.0937	0.6813	0.2833	0.0779	0.0956
ODD						
Using Features	0.2538	0.1657	0.9890	0.9333	0.0314	0.0318
Using Logits	0.1190	0.0739	0.9341	0.7333	0.0290	0.0738
Using Logits w/o δ term	0.0943	0.1005	0.4945	0.2000	0.0803	0.1079

Table 1: Comparing the ODD bound with Dis^2 bound and other prior methods for predicting accuracy. DA denotes if the representations were learned via a domain-adversarial algorithm (DANN, CDANN). MAE: mean absolute error, Coverage: fraction of predictions correctly bounding the true error, Overest.: MAE among shifts whose accuracy is overestimated. ODD improves on MAE in comparison to Dis^2 with some loss in coverage, but maintains a much higher lead in coverage over prior methods.

follow their setup to calculate our bounds using both: the deep feature space, and the logit space. Our domain classifier (used for measuring domain overlap) is also trained on the same space (details in Appendix B). Lastly, the δ term in Theorem 3.7 may make the bound too conservative in practice, so we also calculate the bound without it for a complete comparison with Dis^2.

Evaluation metrics: We evaluate all methods on mean absolute error (MAE), which measures how close the predicted performance is to the actual performance. In addition, we report coverage, which is the fraction of accuracy predictions that are valid ($\leq$ true accuracies). This measures the reliability of the method. We also report the MAE among predictions which are invalid, to get a sense of how bad the over-estimates are, when they do happen.

Domain-adversarially learnt representations: Domain adversarial representation learning methods like DANN and CDANN, regularize deep representations to be indifferent between source and target domains, to make classifiers robust to distribution shifts. This regularization may cause representations to lose domain specific features resulting in a higher degree of overlap between source and target domain features. A higher degree of overlap typically means a more lenient bound using our ODD method, as it suggests that the two domains are similar (see Figure 4). Therefore, when the method produces a classifier which actually has low target error, the ODD-based bound will be tighter. However, when the actual target error is high, it means that the representations have lost important information from the target domain and the high overlap hurts, causing the ODD-based bound to overestimate accuracy. Therefore, we present our findings in two categories: **DA? ✓**: representing the case when the representations were learnt using a domain adversarial algorithm, and **DA? ✗** otherwise.

Results: We report our comprehensive evaluation metrics in Table 1. We find that *ODD-based bounds are more accurate than Dis^2-based bounds* in general; and significantly improve the estimate in many cases (especially with non-DA methods). Moreover, this improved MAE does not come at a huge cost of coverage, where the ODD-based bound maintains a much higher coverage compared to other baselines. We discuss non-DA results in Appendix B.

Improvements in Valid vs Invalid predictions: In Table 2, we segregate the predictions on the basis of validity to investigate failure cases. In most cases, both Dis^2 and ODD based bounds are valid and, the ODD-based bound is significantly more tight compared to Dis^2. In a small number of cases, ODD overshoots the target accuracy while Dis^2 remains valid. Even in these cases, the MAE of ODD is better than Dis^2, meaning it only slightly overestimates the performance of $\hat{h}$. And in a few cases, ODD achieves a valid bound when Dis^2 does not. We conducted a paired Student's t-test on the distributions of (target accuracy − predicted lower bound) for ODD and Dis^2 (for non-DA algorithms). Out of 94 total predictions, 89 were valid (predicted lower bound was actually lower than target accuracy) for both ODD and Dis^2. To remove any influence of invalid predictions, we show the t-statistic for both valid and all predictions cases:

MAE (↓)	DA: ✗		DA: ✓	
	DIS^2	ODD	DIS^2	ODD
Prediction set				
DIS^2 invalid, ODD valid	0.0171	**0.0029**	0.0489	**0.0032**
DIS^2 valid, ODD invalid	0.0540	**0.0248**	0.0325	**0.0215**
Both valid	0.1611	**0.1234**	0.1058	**0.0772**
Both invalid	0.0000	0.0000	**0.0498**	0.0822

Table 2: Comparing MAE of DIS^2 and ODD conditioned on coverage. When ODD-based bound is invalid, it typically overestimates by a small amount. In comparison, DIS^2 typically has a larger MAE when it is invalid. In the most popular case, when both are valid, ODD consistently outperforms DIS^2.

Valid predictions: t-statistic: **-5.47**, p-value: 4.14e-07
All predictions: t-statistic: **-5.62**, p-value: 2.02e-07

As evident, we achieve a high t-statistic with a very low p-value, giving strong evidence against the null-hypothesis (i.e., no significant difference between the two methods).

Variation due to degree of overlap: We plot point-wise bound estimates (using logits) for all non-DA trained $\hat{h}$ (Figure 5) for more analysis. When actual target accuracy is low (left part of the plot), the overlap between source and target domains is likely low. Hence, we see little improvement over DIS^2 in this region. When actual target accuracy is high (right part of the plot), domain overlap is likely also high. Therefore, we see a lot more improvement in this region. Around the middle is when overlap estimate is likely the hardest. This is where ODD sometimes overestimates the overlap and the target accuracy (only slightly) as a result.

6 DISCUSSION

While ODD shows consistent improvement in prediction accuracies, we discuss some potential limitations and improvements to our method.

Estimating overlap: We find using a small domain-classifier's softmax probabilities as weights to discount the DIS^2 contribution of overlapping target samples effective in improving the bound estimation accuracy. However, this approach does not guarantee a good quantification of overlap and is affected by factors like the training hyper-parameters of the domain-classifier (underfitting/overfitting) and the representations on which the classifier is being trained (like DA vs non-DA). A true overlap-aware bound would require accurate density estimation of the source and target densities, which is challenging for high-dimensional data. Nonetheless, our positive results with the CivilComments dataset using BERT embeddings paint a promising picture for the future with increasingly better representation learning.

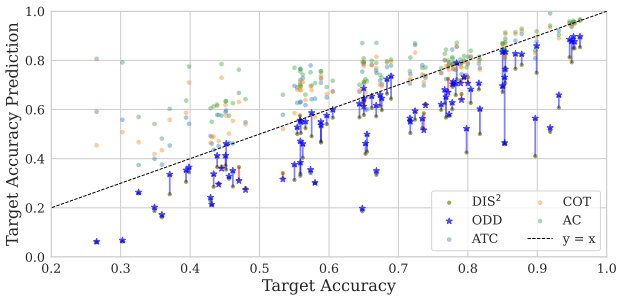

Figure 5: Our ODD-based bound (using logits, **DA? ✗**) improves the DIS^2 bound on multiple dataset-method combinations, achieving an improved average MAE. Improvement is shown with ↑ (deterioration with ↓). Notice that our bound, like DIS^2, still almost always maintains its reliability, unlike other baseline methods.

When does ODD fail? In Figure 6, we compare the estimated bound and the true accuracy on a variety of non-DA representations in both the feature and logit space. We find that both ODD and DIS^2 always remain valid in the case of feature space, and almost always valid in the case of logit space. In both cases, ODD improves on DIS^2 with a more accurate prediction of the bound. However, without the correction δ term in Theorem 3.7, both methods overestimate the accuracy many times. We observe that for a small range of δ values, ODD marginally overshoots the allowed probability of violations. We suspect that this could be caused by an over-estimation of the overlap between the two domains in some cases, making the critic agree more with $\hat{h}$ than it should have. Better overlap estimation should be able to mitigate this issue, which we defer to future work.

Sample Complexity: Our theory does not include sample complexity analysis of training the evaluated classifier. We assume that the classifier is given to us and we focus on estimating the target performance with finite samples. We restricted our search for the critic in the linear hypothesis class which has a uniform sample complexity, but sample complexity for linear classifiers can still be measured and analyzed in a non-uniform way. In our analysis, following Rosenfeld and Garg [2023], we repeated our experiments 30 times to dilute variance (for each dataset, the critic used in bound calculation was chosen from a set of 30 critics learnt on the same data, based on max discrepancy). Regarding analyzing the variance and how it can in turn affect our discrepancy measure and the final bound with a non-uniform sample complexity, we refer to future work.

7 RELATED WORK

Generalization bounds under distribution shift: Finding generalization bounds under distribution shift is a classic machine learning problem with a long history of de-

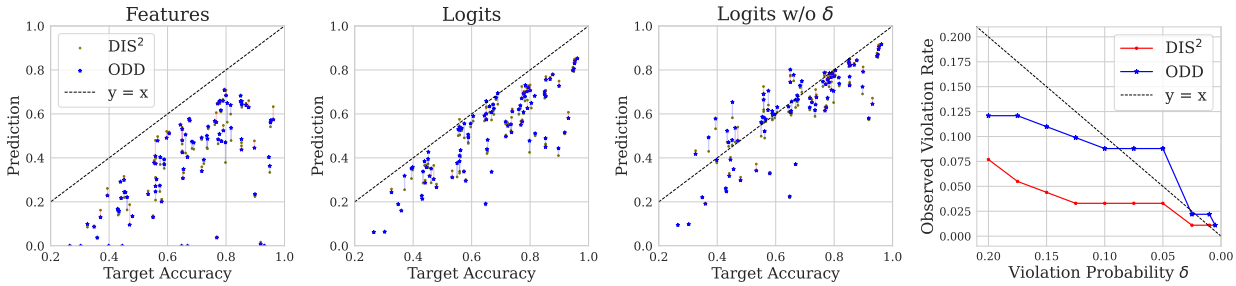

Figure 6: **(First 3)**: Comparing Dis^2 and ODD on estimated bound based on features and logits vs true accuracy. "w/o δ" drops the δ term of Theorem 3.7. **(4th)** Observed bound violation rate vs. desired probability δ from Theorem 3.7 based on logits. The observed violation rate for both Dis^2 and ODD remains under the allowed violation probability for most δ's, but ODD overshoots it for a small set.

velopment [Redko et al., 2019]. Early works using $\mathcal{H}$ and $\mathcal{H}\Delta\mathcal{H}$ divergence [Ben-David et al., 2006, Mansour et al., 2009, Ben-David et al., 2010a] established domain adaptation bounds through uniform convergence. These works inspired many efforts in making classifiers robust to distribution shifts [Zhang, 2019, Ganin et al., 2016, Long et al., 2018, Rahimian and Mehrotra, 2019, Sagawa et al., 2020, Arjovsky et al., 2019]. In particular, a major line of work leverages representation invariance across domains for domain adaptation and generalization [Johansson et al., 2019, Zhao et al., 2019] Theoretically, the $\mathcal{H}\Delta\mathcal{H}$ divergence motivates further improvement on the definition of divergence to measure the difference between domains, with respect to a particular function class [Zhang et al., 2019]. Application to various types of distribution shift [Awasthi et al., 2023] and new methods for the computation of the discrepancy [Kuroki et al., 2019] also followed. In addition, PAC-Bayesian analysis has also been applied to this problem [Germain et al., 2013, 2016] which was recently extended to provide non-uniform sample complexity analysis [Sicilia et al., 2022]. However, due to the general difficulty of estimating the divergence, it is usually hard to directly estimate the target performance following theoretical bounds. [Rosenfeld and Garg, 2023] developed Dis^2 as a theoretically sound but practical method to bound the target risk of a given source classifier, which is the setting of focus in this work.

Another line of work aims to bound the generalization risk of models (espeically neural networks) by measuring and using the complexity of these models [Bartlett et al., 2017, Dziugaite and Roy, 2017, Zhou et al., 2019]. By evaluating the complexity, these works estimate the expressive capacity of the models providing insights into their worst-case behavior, but often lead to loose bounds.

Predicting error in a target domain: Other works focus on predicting the error of a trained classifier/network on unlabeled samples from a target domain. They either provide instance-level estimates using ensembles or data augmentations [Chen et al., 2021, Deng and Zheng, 2021] or an overall domain-level error using empirical observations [Baek et al., 2022], domain-invariant representations [Chuang et al., 2020], average thresholded confidence [Garg et al., 2022] or difference of confidences [Guillory et al., 2021]. Some of these methods require labeled target samples for calibration, which makes their applicability limited. In our work, we mainly focus our effort on improving Dis^2 [Rosenfeld and Garg, 2023] as it presents a strong practical method to give reliable performance bounds estimates. Reliability is a major concern especially when the actual target performance is low. With this work, we push to improve the prediction accuracy of Dis^2 without compromising reliability.

8 CONCLUSION

In this work, we identified a potential for improvement in recent work which uses disagreement discrepancy (Dis^2) to calculate practical bounds on the performance of a source trained classifier in unseen target domains. Dis^2 suffers from a competition between source and target samples for agreement and disagreement respectively in the overlapping region, which may results in a suboptimal critic due to unstable optimization. We mitigate this issue by using overlap-aware disagreement discrepancy (ODD). By restricting the disagreement in the non-overlapping target domains, we eliminate the instability in the overlapping region, and derive a new bound on the basis of ODD. We train a domain classifier to discriminate between source and target samples. We use the softmax probabilities of this domain classifier to act as weights to discount the disagreement of target samples in the overlapping region. This makes the critic agree with source samples in the overlapping region and improves performance bounds as a result, without much compromise in the reliability of the bound.

In the future, we hope to further improve the overlap estimation step. Moreover, in the era of Large Language Models (LLMs), where the input/output space is highly complex, we aim to extend our method to develop generic definitions of disagreement and overlap beyond classification settings; and provide performance estimation for LLMs.

Acknowledgements

We thank the reviewers for their insightful feedback. We also benefit from the highly reproducible work by Rosenfeld and Garg [2023] in our experiments. Both authors are partially supported by a seed grant from JHU Institute of Assured Autonomy (IAA). AL is also partially supported by an Amazon Research Award.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization. In *International Conference on Machine Learning*, pages 244–253. PMLR, 2018.

Pranjal Awasthi, Corinna Cortes, and Christopher Mohri. Theory and Algorithm for Batch Distribution Drift Problems. In *International Conference on Artificial Intelligence and Statistics*, pages 9826–9851. PMLR, 2023.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the Performance of Neural Networks under Distribution Shift. In *Advances in Neural Information Processing Systems*, 2022.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010a.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility Theorems for Domain Adaptation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010b.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting Errors and Estimating Accuracy on Unlabeled Data with Self-training Ensembles. *Advances in Neural Information Processing Systems*, 34: 14980–14992, 2021.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating Generalization under Distribution Shifts via Domain-Invariant Representations. *International Conference on Machine Learning*, 2020.

Weijian Deng and Liang Zheng. Are Labels Always Necessary for Classifier Accuracy Evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15069–15078, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Gintare Karolina Dziugaite and Daniel M Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *Uncertainty in Artificial Intelligence*, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.*, 17 (1):2096–2030, jan 2016. ISSN 1532-4435.

Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging Unlabeled Data to Predict Out-of-Distribution Performance. In *International Conference on Learning Representations*, 2022.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers. In *International Conference on Machine Learning*, pages 738–746. PMLR, 2013.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A New PAC-Bayesian Perspective on Domain Adaptation. In *International Conference on Machine Learning*, pages 859–868. PMLR, 2016.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1144, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*, 2017.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The Low-Rank Simplicity Bias in Deep Networks. *Transactions of Machine Learning Research*, 2024.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and Invertibility in Domain-Invariant Representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. *International Conference on Learning Representations*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. Unsupervised Domain Adaptation Based on Source-Guided Discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4122–4129, 2019.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting Batch Normalization For Practical Domain Adaptation. *International Conference on Learning Representations*, 2017.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional Adversarial Domain Adaptation. *Advances in Neural Information Processing Systems*, 31, 2018.

Yuzhe Lu, Zhenlin Wang, Runtian Zhai, Soheil Kolouri, Joseph Campbell, and Katia Sycara. Predicting Out-of-Distribution Error with Confidence Optimal Transport. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain Adaptation: Learning Bounds and Algorithms. *Citeseer*, 2009.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. VisDA: The Visual Domain Adaptation Challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2Real: A New Benchmark forSynthetic-to-Real Visual Domain Adaptation. *CoRR*, abs/1806.09755, 2018.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment Matching for Multi-Source Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Hamed Rahimian and Sanjay Mehrotra. Distributionally Robust Optimization: A Review. *Open Journal of Mathematical Optimization*, 2019.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younes Bennani. *Advances in Domain Adaptation Theory*. Elsevier, 2019.

Elan Rosenfeld and Saurabh Garg. (Almost) Provable Error Bounds Under Distribution Shift via Disagreement Discrepancy. *Advances in Neural Information Processing Systems*, 36:28761–28784, 2023.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-Adjusted Regression or: ERM May Already Learn Features Sufficient for Out-of-Distribution Generalization. *arXiv preprint arXiv:2202.06856*, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. *International Conference on Learning Representations*, 2020.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for Subpopulation Shift. *International Conference on Learning Representations*, 2021.

Anthony Sicilia, Katherine Atwell, Malihe Alikhani, and Seong Jae Hwang. PAC-Bayesian Domain Adaptation Bounds for Multiclass Learners. In *Uncertainty in Artificial Intelligence*, pages 1824–1834. PMLR, 2022.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems*, 33, 2020.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep Hashing Network for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

Richard Zhang. Making Convolutional Networks Shift-Invariant Again. In *International Conference on Machine Learning*, 2019.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*. PMLR, 2019.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On Learning Invariant Representations for Domain Adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-Vacuous Generalization Bounds at the ImageNet Scale: A PAC-Bayesian Compression Approach. *International Conference on Learning Representations*, 2019.

ODD: Overlap-aware Estimation of Model Performance under Distribution Shift (Supplementary Material)

Aayush Mishra[1] **Anqi Liu**[1]

[1]Department of Computer Science,
Johns Hopkins University, Baltimore, Maryland, USA
{amishr24, aliu.cs}@jhu.edu

A PROOFS

Lemma A.1. *For any classifier h,*
$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \Delta(h, y^*) + \lambda$.

Proof.

$$\begin{aligned}
\epsilon_{\mathcal{T}}(h) &\leq \epsilon_{\mathcal{T}}(h, y^*) + \epsilon_{\mathcal{T}}(y^*) \quad \text{[Triangle Inequality]} \\
&= \epsilon_{\mathcal{S}}(h, y^*) + (\epsilon_{\mathcal{T}}(h, y^*) - \epsilon_{\mathcal{S}}(h, y^*)) + \epsilon_{\mathcal{T}}(y^*) \\
&\leq \epsilon_{\mathcal{S}}(h) + \Delta(h, y^*) + (\epsilon_{\mathcal{T}}(y^*) + \epsilon_{\mathcal{S}}(y^*)) \\
&= \epsilon_{\mathcal{S}}(h) + \Delta(h, y^*) + \lambda
\end{aligned}$$

□

Theorem A.2 (Dis^2 Bound)**.** *Under Assumption 2.4, with probability* $\geq 1 - \delta$,

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\hat{\mathcal{S}}}(\hat{h}) + \hat{\Delta}(\hat{h}, \bar{h}) + \sqrt{\frac{(n_S + 4n_T)\log\frac{1}{\delta}}{2n_S n_T}} + \lambda.$$

Proof. We follow Rosenfeld and Garg [2023] with our notion of *adaptability* through the ideal joint hypothesis y^*. From Lemma A.1 and Assumption 2.4, we have, $\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \Delta(h, \bar{h}) + \lambda$.

To upper bound the first two terms using empirical estimates, we define the following random variables:

$$r_{\mathcal{S},i} = \begin{cases} 0, & \bar{h}(x_i) = y_i^{\mathcal{S}}, \\ \frac{1}{n_S}, & \bar{h}(x_i) = \hat{h}(x_i) \neq y_i^{\mathcal{S}}, \\ \frac{-1}{n_S}, & \bar{h}(x_i) \neq \hat{h}(x_i) = y_i^{\mathcal{S}}, \end{cases} \qquad r_{\mathcal{T},i} = \frac{\mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}}{n_T}$$

Here, $y_i^{\mathcal{S}}$ is the source ground truth label for the ith sample (according to $y^*_{\mathcal{S}}$). Consider the following sums:

$$\sum_{\mathcal{S}} r_{\mathcal{S},i} = \frac{1}{n_{\mathcal{S}}} \sum_{\mathcal{S}} [\mathbf{1}\{\hat{h}(x_i) \neq y_i^{\mathcal{S}}\} - \mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}] = \epsilon_{\hat{\mathcal{S}}}(\hat{h}, y^*_{\mathcal{S}}) - \epsilon_{\hat{\mathcal{S}}}(\hat{h}, \bar{h}),$$

$$\sum_{\mathcal{T}} r_{\mathcal{T},i} = \frac{1}{n_{\mathcal{T}}} \sum_{\mathcal{T}} [\mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}] = \epsilon_{\hat{\mathcal{T}}}(\hat{h}, \bar{h}).$$

Their sum $\epsilon_{\hat{\mathcal{S}}}(\hat{h}, y^*_{\mathcal{S}}) - \epsilon_{\hat{\mathcal{S}}}(\hat{h}, \bar{h}) + \epsilon_{\hat{\mathcal{T}}}(\hat{h}, \bar{h}) = \epsilon_{\hat{\mathcal{S}}}(\hat{h}) + \hat{\Delta}(\hat{h}, \bar{h})$ gives us the empirical estimate of the first two terms. The sum of their corresponding population terms:

$$\mathbb{E}_{\mathcal{S}}\left[\sum_{\mathcal{S}} r_{\mathcal{S},i}\right] = \epsilon_{\mathcal{S}}(\hat{h}) - \epsilon_{\mathcal{S}}(\hat{h}, \bar{h}),$$
$$\mathbb{E}_{\mathcal{T}}\left[\sum_{\mathcal{T}} r_{\mathcal{T},i}\right] = \epsilon_{\mathcal{T}}(\hat{h}, \bar{h}),$$

and λ gives us the bound: $\epsilon_{\mathcal{T}}(h) = \epsilon_{\mathcal{S}}(\hat{h}) + \Delta(\hat{h}, \bar{h}) + \lambda$. Applying Hoeffding's inequality: the probability that the expectation exceeds their sum by t is no more than $\exp\left(-\frac{2t^2}{n_S\left(\frac{2}{n_S}\right)^2 + n_T\left(\frac{1}{n_T}\right)^2}\right)$ and solving for t completes the proof. □

Note that the λ term is incalculable in our setting (without access to target labels), so this bound only provides a qualitative relationship to the target risk. In practice, if λ is estimated through held out target samples, the adjustment due to finite sampling (through Hoeffding's inequality) will also change accordingly.

Theorem A.3. *Under Assumption 2.4, with probability* $\geq 1 - \delta$,

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\hat{\mathcal{S}}}(\hat{h}) + \hat{\Delta}(\hat{h}, \bar{h}, \alpha) + \underline{\hat{\Delta}}(h, \bar{h}, \alpha) + \sqrt{\frac{(n_S + 4n_T)\log\frac{1}{\delta}}{2n_S n_T}} + \lambda.$$

Proof. As $\Delta(h, h') = \Delta(h, h', \alpha) + \underline{\Delta}(h, h', \alpha)$ for any h, h' and α, the random variables defined in the proof for Theorem A.2 can be split in two mutually exclusive and exhaustive sets (overlapping and non-overlapping regions). The rest of the proof follows similarly. □

Theorem A.4 (ODD Bound)**.** *Under Assumption 3.6 and Assumption 3.5, with probability* $\geq 1 - \delta$,

$$\epsilon_{\mathcal{T}}(\hat{h}) \leq \epsilon_{\hat{\mathcal{S}}}(\hat{h}) + \hat{\Delta}(\hat{h}, \bar{h}, \alpha) + \sqrt{\frac{(n_S + 4n_T)\log\frac{1}{\delta}}{2n_S n_T}}$$

Proof. With the common y^*, λ goes to zero. Hence, Lemma A.1 implies: $\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \Delta(h, \bar{h})$.

We now define the following random variables:

$$r_{\mathcal{S}\setminus\mathcal{D}_\alpha,i} = \begin{cases} 0, & \bar{h}(x_i) = y^*_i, \\ \frac{1}{n_S}, & \bar{h}(x_i) = \hat{h}(x_i) \neq y^*_i, \\ \frac{-1}{n_S}, & \bar{h}(x_i) \neq \hat{h}(x_i) = y^*_i, \end{cases} \qquad r^{\mathcal{S}}_{\mathcal{D}_\alpha,i} = \begin{cases} 0, & \bar{h}(x_i) = y^*_i, \\ \frac{1}{n_S}, & \bar{h}(x_i) = \hat{h}(x_i) \neq y^*_i, \\ \frac{-1}{n_S}, & \bar{h}(x_i) \neq \hat{h}(x_i) = y^*_i, \end{cases}$$

$$r_{\mathcal{T}\setminus\mathcal{D}_\alpha,i} = \frac{\mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}}{n_T} \qquad r^{\mathcal{T}}_{\mathcal{D}_\alpha,i} = \frac{\mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}}{n_T}$$

Now we have the following empirical sums:

$$\sum_{\mathcal{S}\setminus\mathcal{D}_\alpha} r_{\mathcal{S}\setminus\mathcal{D}_\alpha,i} = \frac{1}{n_{\mathcal{S}}}\sum_{\mathcal{S}\setminus\mathcal{D}_\alpha}[\mathbf{1}\{\hat{h}(x_i) \neq y^*_i\} - \mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}] = \epsilon^{\mathcal{S}}_{\widehat{\mathcal{S}\setminus\mathcal{D}_\alpha}}(\hat{h}, y^*) - \epsilon^{\mathcal{S}}_{\widehat{\mathcal{S}\setminus\mathcal{D}_\alpha}}(\hat{h}, \bar{h}),$$
$$\sum_{\mathcal{T}\setminus\mathcal{D}_\alpha} r_{\mathcal{T}\setminus\mathcal{D}_\alpha,i} = \frac{1}{n_{\mathcal{T}}}\sum_{\mathcal{T}\setminus\mathcal{D}_\alpha}[\mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}] = \epsilon^{\mathcal{T}}_{\widehat{\mathcal{T}\setminus\mathcal{D}_\alpha}}(\hat{h}, \bar{h}),$$
$$\sum_{\mathcal{D}_\alpha} r^{\mathcal{S}}_{\mathcal{D}_\alpha,i} = \frac{1}{n_{\mathcal{S}}}\sum_{\mathcal{D}_\alpha}[\mathbf{1}\{\hat{h}(x_i) \neq y_i\} - \mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}] = \epsilon^{\mathcal{S}}_{\hat{\mathcal{D}}_\alpha}(\hat{h}, y^*) - \epsilon^{\mathcal{S}}_{\hat{\mathcal{D}}_\alpha}(\hat{h}, \bar{h}),$$
$$\sum_{\mathcal{D}_\alpha} r^{\mathcal{T}}_{\mathcal{D}_\alpha,i} = \frac{1}{n_{\mathcal{T}}}\sum_{\mathcal{D}_\alpha}[\mathbf{1}\{\hat{h}(x_i) \neq \bar{h}(x_i)\}] = \epsilon^{\mathcal{T}}_{\hat{\mathcal{D}}_\alpha}(\hat{h}, \bar{h}),$$

and their corresponding population terms:

$$\mathbb{E}_{\mathcal{S}}\left[\sum_{\mathcal{S}\setminus\mathcal{D}_\alpha} r_{\mathcal{S}\setminus\mathcal{D}_\alpha,i}\right] = \epsilon^{\mathcal{S}}_{\mathcal{S}\setminus\mathcal{D}_\alpha}(\hat{h}, y^*) - \epsilon^{\mathcal{S}}_{\mathcal{S}\setminus\mathcal{D}_\alpha}(\hat{h}, \bar{h}),$$

$$\mathbb{E}_{\mathcal{T}}\left[\sum_{\mathcal{T}\setminus\mathcal{D}_\alpha} r_{\mathcal{T}\setminus\mathcal{D}_\alpha,i}\right] = \epsilon^{\mathcal{T}}_{\mathcal{T}\setminus\mathcal{D}_\alpha}(\hat{h}, \bar{h}),$$

$$\mathbb{E}_{\mathcal{S}}\left[\sum_{\mathcal{D}_\alpha} r^{\mathcal{S}}_{\mathcal{D}_\alpha,i}\right] = \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, y^*) - \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h}),$$

$$\mathbb{E}_{\mathcal{T}}\left[\sum_{\mathcal{D}_\alpha} r^{\mathcal{T}}_{\mathcal{D}_\alpha,i}\right] = \epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h}).$$

The sum of these terms

$$\begin{aligned}
&= \epsilon^{\mathcal{S}}_{\mathcal{S}\setminus\mathcal{D}_\alpha}(\hat{h}, y^*) - \epsilon^{\mathcal{S}}_{\mathcal{S}\setminus\mathcal{D}_\alpha}(\hat{h}, \bar{h}) + \epsilon^{\mathcal{T}}_{\mathcal{T}\setminus\mathcal{D}_\alpha}(\hat{h}, \bar{h}) + \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, y^*) - \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h}) + \epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h}) \\
&= (\epsilon^{\mathcal{S}}_{\mathcal{S}\setminus\mathcal{D}_\alpha}(\hat{h}, y^*) + \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, y^*)) + (\epsilon^{\mathcal{T}}_{\mathcal{T}\setminus\mathcal{D}_\alpha}(\hat{h}, \bar{h}) - \epsilon^{\mathcal{S}}_{\mathcal{S}\setminus\mathcal{D}_\alpha}(\hat{h}, \bar{h})) + (\epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h}) - \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h})) \\
&= \epsilon_{\mathcal{S}}(\hat{h}, y^*) + \Delta(\hat{h}, \bar{h}, \alpha) + (\epsilon^{\mathcal{T}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h}) - \epsilon^{\mathcal{S}}_{\mathcal{D}_\alpha}(\hat{h}, \bar{h})) \\
&\leq \epsilon_{\mathcal{S}}(\hat{h}, y^*_{\mathcal{S}}) + \Delta(\hat{h}, \bar{h}, \alpha) + \underline{\Delta}(\hat{h}, \bar{h}, \alpha) \\
&\leq \epsilon_{\mathcal{S}}(\hat{h}, y^*_{\mathcal{S}}) + \Delta(\hat{h}, \bar{h}, \alpha) \qquad \left(\underline{\Delta}(\hat{h}, \bar{h}, \alpha) \leq 0 \text{ from Assumption 3.5.}\right)
\end{aligned}$$

Now, applying Hoeffding's inequality completes the proof.

□

Remark: Assumption 3.5 bounds the target risk only in terms of non-overlapping disagreement discrepancy. It may be counter-intuitive to think that target disagreement is bounded by source disagreement. However, we have shown that $\Delta(\hat{h}, \bar{h}, \alpha)$ accounts for nearly all of $\Delta(\hat{h}, \bar{h})$ (Figure 4) for $\bar{h}$ found through either DIS^2 or ODD. This implies that even if the target disagreement exceeds source disagreement, it doesn't do it by a lot. Therefore, even without Assumption 3.5, the $\underline{\Delta}$ term from Theorem 3.4 should be expected to be negligible and inconsequential in most practical cases.

B EXPERIMENT DETAILS

B.1 DATASET DETAILS

Synthetic Datasets: The Gaussians are randomly initialized with varying means and covariance matrices. The target gaussian is brought close to the source gaussian by translating its mean using $\mu_{\mathcal{T}} \leftarrow \mu_{\mathcal{T}} + (\mu_{\mathcal{S}} - \mu_{\mathcal{T}}) *$ overlap factor. For a point $X = (x_1, x_2)$, its class label (y^*) is decided using a complex function:

$$y^*(X) = \begin{cases} 0, & \text{if } x_1 \leq \cos(a\sin(bx_2) + ce^{dx_2} + \frac{x_2^2+2x_2-5}{2}) + \epsilon \\ 1, & \text{otherwise} \end{cases}$$

where, a, b, c, d are chosen randomly from a small range and ϵ is small random noise to make the decision boundary noisy. This allowed for simple as well as extremely complex decision surfaces, making for a suitable test bed for our method and its comparison with DIS^2. We show 5 more samples of these datasets in Figure 7.

Real Datasets: We use the datasets provided by Rosenfeld and Garg [2023] on their github repository. This includes CIFAR10 [4 shift variations], CIFAR100 [4 shift variations], DomainNet [3 shift variations], Entity13 [3 shift variations], Entity30 [3 shift variations], FMoW [2 shift variations], Living17 [3 shift variations], NonLiving26 [3 shift variations], OfficeHome [3 shift variations] and Visda [2 shift variations]. All datasets were trained with different deep neural network backbones as described in the DIS^2 paper (Appendix A) with ERM, DANN, CDANN, BN-adapt and FixMatch methods. The deep representations of the trained networks are provided directly for download making replication seamless.

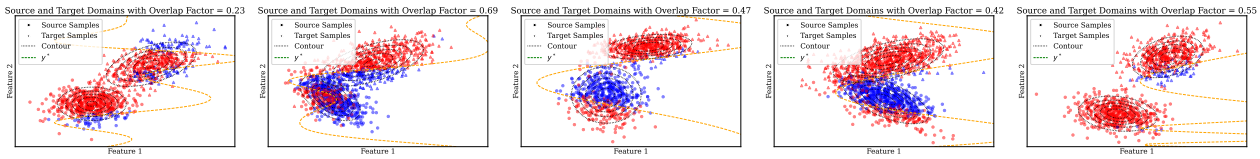

Figure 7: Variations in synthetic datasets.

CivilComments: We used CivilComments Borkan et al. [2019] to test if our method improves over DIS^2 in natural language domain, which it had not been tested on before. This dataset contains sentences labeled *toxic* or not. We created source and target domains using subpopulation shift in this dataset. We filtered all rows with the *black* label True into one set (target) and the rest in another (source), and evenly balanced the dataset to avoid class imbalance issues in evaluation. Train Set: $\sim$ 27k samples in each class, Validation Set: $\sim$ 4.5k samples in each class. We collected the BERT [Devlin et al., 2019] embeddings of all sentences in the source and target domain and trained a linear model on the source domain, which achieved a source validation accuracy of $\sim$ 75%. The source model achieved a target validation accuracy of $\sim$ 65%, which would be the target to predict by our bound. We found that the DIS^2-based bound predicted an accuracy of $\sim$ 58%, while ODD improved it to $\sim$ 63%.

B.2 MORE EXPERIMENTAL DETAILS

Training domain classifiers: For each experiment: (dataset, shift, training method) combination, we train a small 3-layer MLP, with number of neurons = the dimensionality of the representation. We train the network on a balanced training set consisting of equal number of samples from both source and target domains. We randomly subsample the majority domain to make both classes balanced to not skew the classifier towards one of the classes. We train the classifier with Adam Optimizer, with learning rate = 10^{-4}, until convergence (difference in loss $< 10^{-4}$ for 10 epochs).

B.3 MORE EXPERIMENTAL RESULTS

Results on non-DA algorithms: It is noticeable that the MAE for domain-adversarial algorithm based representations is much better than the other. We plot the point-wise bound estimates (using logits) for all DA-method predictions in the appendix Figure 8 and find that like DIS^2, ODD overestimates accuracy in many cases. This is expected as explained in subsection 5.2, but our method still maintains a significantly higher coverage compared to baselines.

Discounting loss of overlapping source samples: We ran an additional experiment on the real datasets where we weighted the source samples in addition to the target samples while optimizing for the critic. We found no significant difference in MAE under this setting (MAE without source weighting: 0.1224, MAE with source weighting: 0.1244). This is probably

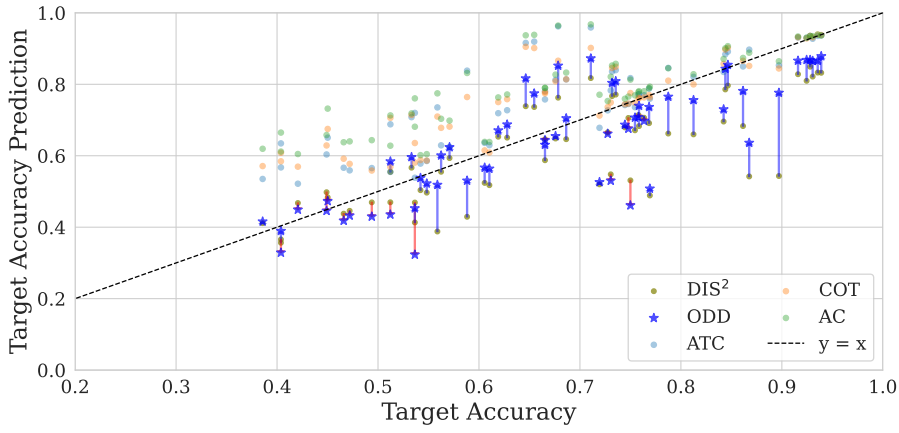

Figure 8: Like DIS^2, **DA?✗** based predictions overestimate accuracy in many cases but still by a smaller margin compared to other baselines.

because we are searching in the simple space of linear models, and ignoring a few samples does not impact the selection of the critic much on average.

Measuring quality of overlap estimation: To measure the quality of overlap estimation, we split the source and target data (of a few studied datasets) into train and dev sets, and early stopped the domain classifier training if the dev set loss did not fall for 10 iterations in a row. As we do not have access to true densities, the dev set performance acts as a proxy for the quality of overlap estimation. We tried multiple sized MLPs and found that the mean dev set accuracy across datasets was almost the same (apart from the linear classifier, which probably underfits in some cases). For details, see Table 3 where we also measure the expected calibration error (ECE) [Guo et al., 2017] of the model. Although we could improve ODD-based bounds by tuning the classifier's hyperparameters for each individual dataset, we used a single setting (3 layer deep, num_features wide network) for all experiments to reflect the robustness of our approach.

Table 3: We find that training domain classifiers on representations is fairly robust to the model architecture.

num_layers	width	Accuracy (mean)	Accuracy (std)	ECE (mean)	ECE (std)	Final MAE
1	-	0.66	0.15	0.09	0.08	0.1234
2	num_features	0.71	0.16	0.11	0.09	0.1269
2	num_features // 2	0.69	0.17	0.10	0.09	0.1236
3	num_features	0.71	0.15	0.15	0.10	0.1223
3	num_features // 2	0.70	0.16	0.14	0.09	0.1282
4	num_features	0.70	0.16	0.19	0.12	0.1229
4	num_features // 2	0.69	0.16	0.18	0.11	0.1255