Diagnosing Model Performance Under Distribution Shift

Tiffany (Tianhui) Cai[1] Hongseok Namkoong[2] Steve Yadlowsky[3]

Columbia University[1,2] Google Research[3]

Department of Statistics[1], Decision, Risk, and Operations Division[2], Brain Team[3]

tiffany.cai@columbia.edu, namkoong@gsb.columbia.edu, yadlowsky@google.com

Abstract

Prediction models can perform poorly when deployed to target distributions different from the training distribution. To understand these operational failure modes, we develop a method, called DIstribution Shift DEcomposition (DISDE), to attribute a drop in performance to different types of distribution shifts. Our approach decomposes the performance drop into terms for 1) an increase in harder but frequently seen examples from training, 2) changes in the relationship between features and outcomes, and 3) poor performance on examples infrequent or unseen during training. Empirically, we demonstrate how our method can 1) inform potential modeling improvements across distribution shifts for employment prediction on tabular census data, 2) measure the degree to which an image classification model is invariant across a distribution shift, and 3) help to explain why certain domain adaptation methods fail to improve model performance for satellite image classification.

1 Introduction

Prediction models operate on data distributions different from those seen during training, but often perform worse on these *target* distributions than on the original *training* distribution. For example, Wong et al. [105] observed EPIC's sepsis risk assessment model, which is deployed across hundreds of hospitals in the US, performs "substantially worse" in the wild compared to vendor claims. The lack of reliability in predictive performance across distributions has been documented in many domains including healthcare [57, 4, 111, 16, 105], loan approval [40], wildlife conservation [6], and education [1]. Similar performance degradation has been widely observed in computer vision and natural language processing (NLP) settings [78, 62, 97, 89, 63, 56]. As decisions are increasingly made based on model predictions, we need rigorous and scalable tools for diagnosing model failures in order to guide resources toward effective model improvements.

Different distribution shifts require different solutions. Understanding *why* model performance worsened is a fundamental step for informing subsequent methodological and operational interventions. While there are multiple taxonomies for discussing distribution shifts [60, 87, 101], we focus on a popular one: we categorize distribution shifts as either a change in the marginal distribution of the covariates (X), or a change in the conditional relationship between the label and covariate ($Y \mid X$). Since real distribution shifts occur as a combination of both types, we develop a framework for understanding how much performance degradation is assigned to each type in order to best improve model performance.

There are many reasons why the marginal distribution of the covariates X can change. For example, training and target data may be collected from different points in time and space, marginalized groups may be underrepresented in the training data, or population demographics may change over time [91, 7, 17, 14]. If the shift in the distribution of covariates X is small, then domain adaptation [91], warm-starting/fine-tuning [21, 106], or importance sampling-based

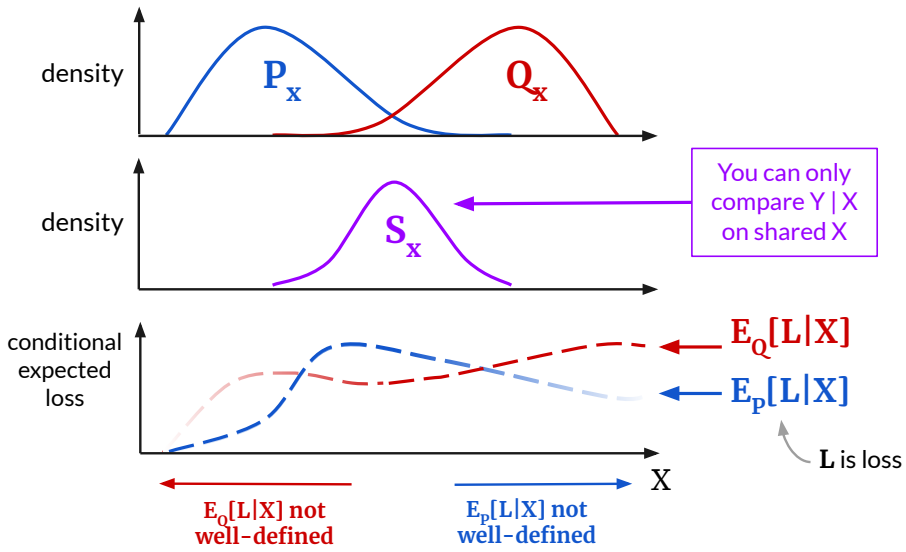


Figure 1. Diagnosing distribution shifts: Illustration of a key idea on a one-dimensional covariate X: we can only compare $Y \mid X$ for training vs test, and predictive performances thereof, on X's that are common between both distributions. We use L to denote the loss.

reweighting [3, 70, 28] approaches can be effective. If there is a more extreme form of covariate shift, such as the presence of target covariate values X that were unseen during training, then it may be necessary to collect data and labels over this new group.

On the other hand, shifts in the relationship between the outcomes and covariates $Y \mid X$ can be caused by changes in measurement errors, changes in user behavior, and unrecorded confounding factors whose distributions change [40]. In these cases, recent works on explicit causal or worst-case modeling may be helpful [107, 37, 12, 82, 9, 53, 104, 29]. When methods like these cannot address the performance drop, existing data from the training distribution may not be useful for the target distribution, so that further data collection and labeling may be necessary. Alternatively, the covariates X may not be predictive of the outcome between both distributions, requiring changes in feature engineering.

Towards diagnosing model performance degradations, we propose a new framework which we call DIstribution Shift DEcomposition (DISDE). At a high level, performance degradation attributed to $Y \mid X$ shift is obtained by varying $Y \mid X$ between training and test while keeping the distribution of X fixed, and performance degradation attributed to X shift is obtained by varying the distribution of X while keeping $Y \mid X$ fixed. In order to make these comparisons, observe that the training and target distributions of $Y \mid X$—and predictive performances thereof—can only be compared over values of X that are sufficiently common between training and target (see Figure 1). Thus, one core aspect of our framework is the notion of a *shared distribution* between X's of the training and target distributions over which it is easy to estimate model performace (Section 3). Using this shared distribution, we decompose the performance degradation into three terms: one for changes in the relationship between X and Y on this shared distribution, and two for changes in the distribution of X from both the training and target distributions to the shared distribution.

By attributing loss to different distribution shifts, our approach goes beyond detecting changes in the distribution, as DISDE quantifies how much each type of shift affects performance. The first term in our decomposition (Equation (2.1) to come) measures performance degradation due to

an increase in the frequency of harder values of X that are common between training and target distributions (an X shift). The second term measures performance degradation due to changes in the relationship between X and Y (a $Y \mid X$ shift), over values of X common to both distributions. The third measures performance degradation due to poor performance on values of X that are infrequent or unseen during training (an X shift).

One important aspect of DISDE is precisely defining and estimating the shared distribution. As we discuss in Section 2.2, we define this shared distribution so that it has the largest mass on values of X that are most likely in both training and target. Our approach is motivated by weighting schemes in causal inference that focus on a subset of the population where treatment and control units are similar enough to compare [23, 59]. To estimate this shared distribution, we train a binary classifier to predict between the training and target distributions using features X, which we refer to as the *auxiliary domain classifier*. Using this classifier, we can reweight samples from training and test into the shared distribution; we operationalize "common" examples across train and target as those that cannot be confidently and correctly classified as being from either training or target.

Our approach is general, and DISDE can be extended to attribute performance degradation to use Z in place of X for *any* flexible feature vector Z (Section 2.3). For example, Z can be a subset of covariates (Section 4.1.1, a (learned) feature mapping (Section 4.3), some other piece of metadata, or even the label Y itself, turning DISDE into a method for diagnosing label shift.

We apply our method to three real distribution shifts in Section 4. In the first example (Section 4.1), we predict employment from tabular census data [27] and study how predictive performance deteriorates over a variety of natural and semi-synthetic distribution shifts. We illustrate how DISDE gives correct performance attributions, and how it can inform different ways of improving performance on the target distribution. In the second example (Section 4.2), we use DISDE to measure the degree to which image classification models are invariant across a distribution shift, as invariance is often thought of as a desirable property [73, 83, 2, 81]. We focus on zero-shot CLIP models [77], which are interesting because of their exceptional performance across diverse datasets. In the last example (Section 4.3), we study satellite image classification for land or building use (FMoW-WILDS [52]). We use DISDE to diagnose why a popular domain adaptation method, DANN [36], does not do better than standard empirical risk minimization over spatiotemporal shifts, even though DANN is designed to perform well over distribution shifts. This example also demonstrates how DISDE can be scaled to high-dimensional, complex data such as images. Overall, our experiments show how DISDE can be used to guide modeling improvements and to provide insight where specific interventions fail, and also to understand learned feature representations.

Finally, we analyze the large-sample statistical properties of the proposed estimation method in Section 5. By leveraging results from the semiparametric statistics literature [68], we give a central limit theorem for our estimator even when the auxiliary domain classifier is estimated nonparametrically. This provides a principled approach for generating confidence intervals in DISDE. We compare the asymptotic variance of our estimator to lower bounds under a well-developed calculus for statistical functionals [10, 66, 50] to confirm that our proposed approach is statistically efficient, under suitable regularity assumptions on the auxiliary classifier.

To summarize, our main contributions are as follows:

- We introduce a simple method, DIstribution Shift DEcomposition (DISDE), to attribute a change in model performance across a distribution shift to X and $Y \mid X$ shifts.
- We empirically demonstrate how DISDE can inform model improvements across distribution shifts on both tabular and image data.

- We analyze the asymptotic properties of DISDE, provide a principled way to estimate confidence intervals, and show efficiency of our estimator under certain conditions.

As there is a large body of work on distribution shifts across multiple fields, we defer a discussion of related literature to Section 6 where we situate the current work in a broader context.

2 Decomposing the performance degradation

Consider a model f that predicts outcome Y from covariates X. Let f be trained on data (X, Y) drawn from the *training* distribution P. Let $\ell(f(x), y)$ be a loss function denoting a notion of predictive error (e.g. 0-1 error, cross-entropy loss). We consider a situation in which we observe a performance degradation from P to a *target* distribution Q, i.e. $\mathbb{E}_Q[\ell(f(X), Y)] > \mathbb{E}_P[\ell(f(X), Y)]$. As a motivating example, consider a model trained to predict employment status based on demographic information, which could be important for economic policy-making and marketing. Due to operational constraints, the model may be trained using data collected from one state, such as West Virginia (training distribution), but used across a range of different states, including Maryland (target distribution). In such cases, we may only have enough labeled data from Maryland to evaluate a model, but not enough to train a new model. As we show in Table 1 and further investigate in Section 4.1, the model trained on West Virginia performs poorly on Maryland. As collecting more training data from Maryland is costly, we seek to understand the performance drop to determine how best to improve our models for use on Maryland.

Training accuracy	70%
Target accuracy	61%

Table 1. Performance degradation in the employment prediction example

To understand how the distribution shift led to such a performance degradation, we attribute the performance degradation to shifts in the marginal distribution of X and to the conditional distribution of $Y \mid X$. The natural objects to quantify X and $Y \mid X$ shifts are, respectively, the marginal distributions of X, $P_X(\cdot)$ and $Q_X(\cdot)$, and the conditional risks on P and Q,

$$R_\mu(x) := \mathbb{E}_\mu[\ell(f(X), Y) \mid X = x] \text{ for } \mu = P, Q.$$

However, there are two practical issues with studying these quantities directly. First, estimation of these functions is a challenging density estimation or nonparametric regression problem. As in our motivating example above, one of our primary use cases is when we only have limited labeled target data from Q. Second, even if we could derive reasonable estimates of these functions, it would be difficult to display them in a way that would be useful for a human to understand, especially for X of larger dimension. Therefore, we need to summarize the infinite dimensional quantities $\{R_P(\cdot), R_Q(\cdot), P_X, Q_X\}$ in a practical and statistically efficient way.

Our approach is the following: to attribute loss to X shift, compare the conditional risk $R_P(x)$ (resp. $R_Q(x)$) under different marginal distributions of X. Then, to attribute loss to $Y \mid X$ shift, compare $R_P(x)$ and $R_Q(x)$ under the same marginal distribution of X. Recalling Figure 1, the main challenge in doing this is that $R_P(x)$ (resp. $R_Q(x)$) is only well-defined for values of x that are in the support of P_X (resp. Q_X). Therefore, when comparing performance $R_Q(x) - R_P(x)$ across different $Y \mid X$ distributions, we need to define a *shared distribution* S_X with support contained in both P_X and Q_X. We temporarily defer a discussion on different choices of formulations for shared distributions to Section 2.2 and first outline our approach for any fixed definition of the shared distribution S_X.

We thus attribute performance degradation to $Y \mid X$ by comparing the averages of the conditional risks $R_P(X)$ and $R_Q(X)$ on the marginal distribution S_X as

$$\mathbb{E}_{S_X}[R_{\boldsymbol{P}}(X)] \text{ versus } \mathbb{E}_{S_X}[R_{\boldsymbol{Q}}(X)].$$

Then, for X shift, we can construct a well-defined comparison of the average of $R_P(X)$ under the two marginal distributions P_X and S_X,

$$\mathbb{E}_{\boldsymbol{P_X}}[R_P(X)] \text{ versus } \mathbb{E}_{\boldsymbol{S_X}}[R_P(X)],$$

and similarly for the average of $R_Q(X)$ under the marginal distributions Q_X and S_X. Altogether, we can decompose the change in expected loss from P to Q as the telescoping sum

$$\begin{aligned}
\mathbb{E}_Q[\ell(f(X),Y)] - \mathbb{E}_P[\ell(f(X),Y)] &= \mathbb{E}_{S_X}[R_P(X)] - \mathbb{E}_P[R_P(X)] && X \text{ shift } (P \to S) && (2.1a)\\
&\quad + \mathbb{E}_{S_X}[R_Q(X) - R_P(X)] && Y \mid X \text{ shift} && (2.1b)\\
&\quad + \mathbb{E}_Q[R_Q(X)] - \mathbb{E}_{S_X}[R_Q(X)]. && X \text{ shift } (S \to Q) && (2.1c)
\end{aligned}$$

The telescoping sum is illustrated in Figure 2, where the bolded arrows represent the terms in Equation (2.1).

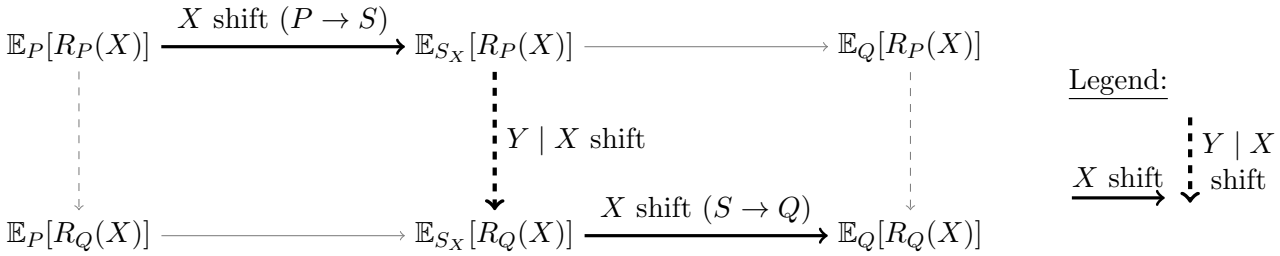


Figure 2. Diagram of decomposition: left to right represents changing X distribution from P_X to S_X to Q_X. Up to down represents changing $Y \mid X$ from $P_{Y|X}$ to $Q_{Y|X}$, by replacing $R_P(X)$ with $R_Q(X)$. The bolded arrows correspond to the decomposition terms we propose, where each arrow $A \to B$ represents the difference $B - A$.

2.1 Interpreting the decomposition (2.1)

We provide intuition for the terms in Equation (2.1) and discuss ways to improve test performance for when each term is large.

2.1.1 X shift $(P \to S)$ term

$\mathbb{E}_{S_X}[R_P(X)] - \mathbb{E}_{P_X}[R_P(X)]$ is the expected difference in loss under $P_{Y|X}$ between examples with $X \sim P_X$ and $X \sim S_X$. This term is large if the model performs worse on covariates X common in both P_X and Q_X under $P_{Y|X}$, compared to covariates from P_X. This can happen if e.g. Q_X has more of the hard examples from P_X. If $P_{Y|X} = Q_{Y|X}$, then we expect the shared data $X \sim S_X$ to be able to inform good models under Q_X, e.g. by using domain adaptation methods.

In the context of the employment prediction problem described at the beginning of the section, a large X shift $(P \to S)$ term could indicate that the change in loss between West Virginia and Maryland is largely due to Maryland having more people than West Virginia of demographics that

were difficult for the model to predict. In this case, we may refit the model by reweighting training observations from West Virginia to resemble the population in Maryland.

2.1.2 $Y \mid X$ shift term

$\mathbb{E}_{S_X}[R_Q(X) - R_P(X)]$ is the expected difference in loss between $Q_{Y|X}$ and $P_{Y|X}$ over the shared covariates $X \sim S_X$ common across P_X and Q_X. We expect this term to be large if the loss is larger under $Q_{Y|X}$ than under $P_{Y|X}$ over shared covariates $X \sim S_X$. Loosely speaking, this can happen if the relationship $Y \mid X$ changes unfavorably from P to Q for the prediction model f. In cases with a large $Y \mid X$ shift term, it may be necessary to collect and label new data over $Q_{Y|X}$, or to modify the set of covariates.

In the context of employment prediction, a large X shift ($S \to Q$) term may indicate that the change in model loss is largely due to a difference in the likelihood of whether a person *of the same demographics* is employed, in West Virginia vs Maryland, for people whose demographics are common in both states. Solving this may require investing in a new data collection campaign in Maryland large enough to train a new model. Alternatively, we may search for new covariates X' such that $Y \mid (X, X')$ remains similar across West Virginia (training) and Maryland (target) and train a new model that also uses the additional features X'.

2.1.3 X shift ($S \to Q$) term

$\mathbb{E}_{Q_X}[R_Q(X)] - \mathbb{E}_{S_X}[R_Q(X)]$ is the expected difference in loss under $Q_{Y|X}$ between examples $X \sim S_X$ and $X \sim Q_X$. Consider values of X that are frequent in the target distribution Q_X but infrequent or unseen in the training distribution P_X. We colloquially refer to these as "newer examples", which include unseen examples as well as values of X that are relatively more common in Q_X compared to S_X. The X shift ($S \to Q$) term is large if the model performs poorly incurs higher loss on such "newer examples" under $Q_{Y|X}$. This type of performance degradation can be addressed by collecting "newer examples" and retraining or fine-tuning a model on these examples.

In our employment prediction example, a large X shift ($S \to Q$) term could indiciate that the change in model loss is largely due to Maryland having people of demographics that are unseen or uncommon West Virginia, that are also difficult for the model to predict. A potential solution involves conducting a survey of employment status in Maryland targeting demographic groups that were rarely observed in West Virginia.

2.1.4 Comparisons

The two X shift terms both attribute performance degradation to X shift, but differ in the following way: the X shift ($P \to S$) term contains performance degradation from P_X having support on examples with low loss where Q_X does not, and the X shift ($S \to Q$) term contains performance degradation from Q_X having support on examples with high loss where P_X does not. However, beyond this distinction, how much a value of X is "allocated" to either of these terms may not necessarily be meaningful. Another difference between the two X shift terms is that the X shift ($P \to S$) term uses $P_{Y|X}$ while the X shift ($S \to Q$) term uses $Q_{Y|X}$. Because of this, one part of the X shift (from P_X to S_X) is measured under $P_{Y|X}$, while the rest (fom S_X to Q_X) is measured under $Q_{Y|X}$. Again, this distinction can be somewhat arbitrary. Despite these subtleties, in practice, it may be helpful to think of these two terms as qualitatively similar when diagnosing a model's failures in practice.

For concrete examples of decompositions for different $Y \mid X$ vs X shifts and the corresponding modeling interventions, please refer to the experiments in Section 4.1.

2.2 Defining the shared distribution S_X

To operationalize our approach, we choose a specific shared distribution S_X over X whose support is contained in that of P_X and Q_X. An ideal choice of shared distribution is one that has higher density when the P_X and Q_X both have higher density, and lower density when either of P_X or Q_X has a low density. This sharpens the above requirement of shared support to a stronger notion of sharing regions of high density. Letting p_X, q_X, and s_X be the densities of X under P, Q, and S under a suitable base measure, consider choosing

$$s_X(x) \propto \frac{p_X(x) q_X(x)}{p_X(x) + q_X(x)}. \tag{2.2}$$

However, there are many shared distributions that satisfy our criteria, and the above choice is not unique. For example, other choices include

$$s_X(x) \propto \min\{p_X(x), q_X(x)\} \quad \text{or} \tag{2.3a}$$

$$s_X(x) \propto \begin{cases} p_X(x) + q_X(x) & \text{if } \min\left\{\frac{p_X(x)}{q_X(x)}, \frac{q_X(x)}{p_X(x)}\right\} \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{2.3b}$$

for some $\epsilon > 0$. The shared distribution (2.3b) is particularly intuitive as it defines "shared" examples as those that have sufficiently high likelihood ratios. However, it is difficult to operationalize as it depends on an arbitrary choice of $\epsilon > 0$. Observe that for all of the above definitions of S_X, if $P_X = Q_X$, then $S_X = P_X = Q_X$. Additionally observe that Equations (2.2) and (2.3a) become similar (up to a scaling factor) for regions of x where $p_X(x) \gg q_X(x)$ or $p_X(x) \ll q_X(x)$.

For the shared distributions listed above, our decomposition can be estimated by reweighting data from P and Q to approximate samples from the shared distribution, as described in more detail in Section 3. We use an auxiliary binary classifier between P and Q to learn these importance weights, which are bounded under weak assumptions and provide low (asymptotic) variance estimates for $\mathbb{E}_{S_X}[R_P(X)]$ and $\mathbb{E}_{S_X}[R_Q(X)]$. Most of our method extends to the alternative shared distributions (2.3) in place of the one in Equation (2.2), and in practice we find that the qualitative conclusions obtained from our approach are not very sensitive to the specific choice of shared distribution (Appendix F). For ease of exposition, we focus on the shared distribution defined in Equation (2.2).

2.3 Generalization to other decompositions

Our approach generalizes to any choice of mapping $(X, Y) \mapsto Z$ in place of X. This generalization attributes performance degradation from P to Q to changes in the marginal distributions of Z, and to changes in the conditional distribution of the data (X, Y) given Z. Specifically, we can generalize $R_\mu(z)$ to be

$$R_\mu(z) := \mathbb{E}_\mu[\ell(f(X), Y) | Z = z] \text{ for } \mu = P, Q,$$

and re-write the decomposition as

$$\mathbb{E}_Q[\ell(f(X), Y)] - \mathbb{E}_P[\ell(f(X), Y)] = \mathbb{E}_{S_Z}[R_P(Z)] - \mathbb{E}_P[R_P(Z)] \qquad Z \text{ shift } (P \to S) \tag{2.4a}$$

$$+ \mathbb{E}_{S_Z}[R_Q(Z) - R_P(Z)] \qquad (X, Y) \mid Z \text{ shift} \qquad (2.4b)$$
$$+ \mathbb{E}_Q[R_Q(Z)] - \mathbb{E}_{S_Z}[R_Q(Z)]. \qquad Z \text{ shift } (S \to Q) \qquad (2.4c)$$

This generalization allows us to provide a richer diagnostic for machine learning applications. For example, when $Z = \phi(X)$ where $\phi(\cdot)$ is a feature mapping (e.g. from the prediction model f), large performance degradation attributed to $Y \mid \phi(X)$ shift may imply the relationship between Y and $\phi(X)$ is not stable between training and target. This could suggest that those features are not appropriate for prediction across both distributions. This is demonstrated in experiments in Section 4.1.1, where Z is a subset of X, and Section 4.3, where Z is a learned feature representation.

We can also use metadata as Z. In this case, we can think of X as containing both metadata and "regular" data, and $f(\cdot)$ to only use regular data. As an additional example, we can use Y as Z, in which case the decomposition measures label shift [60].

The methodological (Section 3) and theoretical (Section 5) results in the rest of the paper extend immediately to using Z instead of X; we would replace $\pi(x), R_P(x), R_Q(x)$ with $\pi(z), R_P(z), R_Q(z)$ and similarly for their estimates.

3 Estimation

In our decomposition (2.1), each term corresponds to the difference of consecutive pairs of $\mathbb{E}_P[R_P(X)]$, $\mathbb{E}_{S_X}[R_P(X)]$, $\mathbb{E}_{S_X}[R_Q(X)]$, $\mathbb{E}_Q[R_Q(X)]$, as illustrated in Figure 2. The first and last terms are easily estimated by the empirical average of losses over $(X_i, Y_i) \overset{\text{iid}}{\sim} P$ and $(X_j, Y_j) \overset{\text{iid}}{\sim} Q$, respectively. Thus, we focus on the estimation of the middle two terms

$$\mathbb{E}_{S_X}[R_P(X)] \quad \text{and} \quad \mathbb{E}_{S_X}[R_Q(X)]. \qquad (3.1)$$

A key statistical challenge is that we do not have access to samples from the shared distribution S_X since it is a fictitious quantity defined for interpretating the performance degradation. As we only have samples $X_i \overset{\text{iid}}{\sim} P$ and $X_j \overset{\text{iid}}{\sim} Q$, our strategy is to reweight them to approximate samples from S_X, which in turn allows us to estimate the two terms (3.1). To see this, rewrite our estimands as

$$\theta_P := \mathbb{E}_{S_X}[R_P(X)] = \mathbb{E}_P\left[R_P(X)\frac{dS_X}{dP_X}(X)\right] = \mathbb{E}_P\left[\ell(f(X), Y)\frac{dS_X}{dP_X}(X)\right] \qquad (3.2a)$$
$$\theta_Q := \mathbb{E}_{S_X}[R_Q(X)] = \mathbb{E}_Q\left[R_Q(X)\frac{dS_X}{dQ_X}(X)\right] = \mathbb{E}_Q\left[\ell(f(X), Y)\frac{dS_X}{dQ_X}(X)\right]. \qquad (3.2b)$$

Once we obtain estimates for the importance weights $\frac{dS_X}{dP_X}$ and $\frac{dS_X}{dQ_X}$, we estimate θ_P and θ_Q by using an empirical plug-in of Equation (3.2) above, and as described in Algorithm 1.

To calculate the importance weights $\frac{dS_X}{dP_X}$, $\frac{dS_X}{dQ_X}$, we can use an auxiliary domain classifier. Recalling the definition of the shared distribution S_X (2.2), we have

$$\frac{dS_X}{dP_X}(x) \propto \frac{q(x)}{p(x) + q(x)} \quad \text{and} \quad \frac{dS_X}{dQ_X}(x) \propto \frac{p(x)}{p(x) + q(x)}. \qquad (3.3)$$

Now observe that $\frac{dS_X}{dP_X} \propto \frac{q(x)}{p(x)+q(x)}$ corresponds to the probability that a data point with value x drawn from an even mixture distribution of P_X and Q_X actually came from Q_X, which can be estimated using any machine learning classifier that outputs class probabilities (e.g. neural networks, random forests, logistic regression), as we can consider coming from Q_X vs P_X to be the classes for

classification. We can thus leverage a broad set of scalable and effective machine learning methods to estimate these probabilities.

Observe also that weighting samples from P_X by the classifier probability that it came from Q_X is consistent with the intuition that data points likely to be from the shared distribution S_X are those that are difficult to classify correctly as being from P_X or Q_X.

In practice, we may not have an even mixture of samples from P_X and Q_X on which to learn the auxiliary domain classifier, e.g. if we have fewer samples from target than training. In such cases, we can use standard techniques from the label-shift literature to adjust the base rate [33]. To help formalize this, define the following:

$$\begin{aligned} &\alpha^* \text{ is the proportion of the pooled data that comes from } Q_X \\ &T = \begin{cases} 0 & \text{if } \tilde{X} \text{ is from } P_X \\ 1 & \text{if } \tilde{X} \text{ is from } Q_X \end{cases} \\ &\pi^*(x) := \mathbb{P}(T = 1 \mid \tilde{X} = x) = \frac{\alpha^* q(x)}{\alpha^* q(x) + (1-\alpha^*)p(x)} \end{aligned} \tag{3.4}$$

where the last equality is by Bayes' rule, and $\mathbb{P}(\cdot)$ is probability under the distribution of the observed data. As before, $\pi^*(x)$ can also be interpreted as the probability output from the true domain classifier. Then we can express the importance ratios as follows:[1]

$$\frac{dS_X}{dP_X}(x) \propto \frac{\pi(x)}{(1-\alpha)\pi(x) + \alpha(1-\pi(x))} =: w_P(\pi(x), \alpha) \tag{3.5a}$$

$$\frac{dS_X}{dQ_X}(x) \propto \frac{1-\pi(x)}{(1-\alpha)\pi(x) + \alpha(1-\pi(x))} =: w_Q(\pi(x), \alpha). \tag{3.5b}$$

Since $w_Q(\pi^*(x), \alpha^*) \propto \frac{dS_X}{dQ_X}$ and $w_P(\pi^*(x), \alpha^*) \propto \frac{dS_X}{dP_X}$, we rewrite the estimands θ_P and θ_Q from (3.2) as functionals of the observed data distributions P and Q, the conditional probability $\pi^*(x)$, and the mixture proportion α^*:

Proposition 1. *With w_P, w_Q as defined in Equation (3.5),*

$$\theta_P = \frac{\mathbb{E}_P\left[\ell(f(X), Y) w_P(\pi^*(X), \alpha^*)\right]}{\mathbb{E}_P\left[w_P(\pi^*(X), \alpha^*)\right]} \quad \textit{and} \quad \theta_Q = \frac{\mathbb{E}_Q\left[\ell(f(X), Y) w_Q(\pi^*(X), \alpha^*)\right]}{\mathbb{E}_Q\left[w_Q(\pi^*(X), \alpha^*)\right]}. \tag{3.8}$$

The reformulation (3.8) lends itself well to a two-stage semiparametric estimation technique. In the first stage, we can use various ML methods or nonparametric statistical estimators to estimate $\pi^*(x)$ as $\widehat{\pi}(x)$, and an empirical mean to estimate α^* as $\widehat{\alpha}$. Then, we plug the estimated quantities into the *empirical* expectations

$$\widehat{\theta}_P = \frac{\frac{1}{n_P}\sum_{i=1}^{n_P} \ell(f(X_i), Y_i) w_P(\widehat{\pi}(X_i), \widehat{\alpha})}{\frac{1}{n_P}\sum_{i=1}^{n_Q} w_P(\widehat{\pi}(X_i), \widehat{\alpha})} \quad \text{and} \quad \widehat{\theta}_Q = \frac{\frac{1}{n_Q}\sum_{j=1}^{n_Q} \ell(f(X_j), Y_j) w_Q(\widehat{\pi}(X_j), \widehat{\alpha})}{\frac{1}{n_Q}\sum_{i=1}^{n_Q} w_Q(\widehat{\pi}(X_j), \widehat{\alpha})}. \tag{3.9}$$

See Algorithm 1 for a summary of the procedure, Appendix D for discussion of data splits and additional implementation details, and Section 5 for statistical properties of the estimators.

Note that the domain classifier probability $\pi^*(x)$ is analogous to the propensity score in causal inference. Our method is motivated by weighting schemes in causal inference that focus on a subset of the population where treatment and control units are similar enough to compare [23, 59].

[1] An easy way to see this is that $\frac{dS_X}{dQ_X}(x), \frac{dS_X}{dP_X}(x), \pi^*(x)$ can all be written in terms of $q(x)/p(x)$.

Algorithm 1: DECOMPOSE CHANGE IN LOSS UNDER SHIFT FROM P TO Q

1 Estimate $\mathbb{E}_P[R_P(X)]$ and $\mathbb{E}_Q[R_Q(X)]$ using data $(X_i, Y_i) \sim P$, with $i = 1, \ldots, n_P$, and $(X_j, Y_j) \sim Q$, with $j = 1, \ldots, n_Q$:

$$\mathbb{E}_P[R_P(X)] \approx \frac{1}{n_P} \sum_{i=1}^{n_P} \ell(f(X_i), Y_i) \quad \text{and} \quad \mathbb{E}_Q[R_Q(X)] \approx \frac{1}{n_Q} \sum_{j=1}^{n_Q} \ell(f(X_j), Y_j).$$

2 Estimate $\widehat{\alpha} = n_Q/(n_P + n_Q)$.

3 Estimate $\widehat{\pi}(x) \approx \mathbb{P}(T = 1|X = x)$ by training a classifier on samples from P_X and Q_X.

4 Calculate importance weights proportional to $\frac{dS_X}{dP_X}$ and $\frac{dS_X}{dQ_X}$:

$$w_P(\widehat{\pi}(x), \widehat{\alpha}) = \frac{\widehat{\pi}(x)}{(1-\widehat{\alpha})\widehat{\pi}(x) + \widehat{\alpha}(1-\widehat{\pi}(x))} \quad \text{and} \quad w_Q(\widehat{\pi}(x), \widehat{\alpha}) = \frac{1-\widehat{\pi}(x)}{(1-\widehat{\alpha})\widehat{\pi}(x) + \widehat{\alpha}(1-\widehat{\pi}(x))}.$$

5 Estimate $\mathbb{E}_{S_X}[R_P(X)]$ and $\mathbb{E}_{S_X}[R_Q(X)]$ using these importance weights:

$$\mathbb{E}_{S_X}[R_P(X)] \approx \frac{\sum_{i=1}^{n_P} \ell(f(X_i), Y_i) w_P(\widehat{\pi}(X_i), \widehat{\alpha})}{\sum_{i=1}^{n_P} w_P(\widehat{\pi}(X_i), \widehat{\alpha})}$$
$$\mathbb{E}_{S_X}[R_Q(X)] \approx \frac{\sum_{j=1}^{n_Q} \ell(f(X_j), Y_j) w_Q(\widehat{\pi}(X_j), \widehat{\alpha})}{\sum_{j=1}^{n_Q} w_Q(\widehat{\pi}(X_j), \widehat{\alpha})}.$$

6 Estimate the terms in Equation (2.1) by taking differences between consecutive pairs of estimates for

$$\mathbb{E}_P[R_P(X)],\ \mathbb{E}_{S_X}[R_P(X)],\ \mathbb{E}_{S_X}[R_Q(X)],\ \mathbb{E}_Q[R_Q(X)]. \tag{3.7}$$

4 Experiments

We demonstrate the versatility of our approach with three sets of experiments. In the first experiment setting (Section 4.1), we predict employment status using tabular census data, as in our motivating example. We consider various known distribution shifts and verify that our decomposition attributes the observed performance degradation to the appropriate types of shift. We also illustrate how DISDE can help guide the allocation of resources toward more effective modeling interventions.

In the second experiment setting (Section 4.2), we study a pair of image classification benchmark datasets where the labeling method is the same between both, so that there is no $Y \mid X$ shift. We investigate the quality of learned feature representations $\phi(X)$ to see if the feature representations also capture a lack of $Y \mid \phi(X)$ shift, and we focus on zero-shot CLIP [77] models (and their corresponding image feature representations) which were found to be particularly robust to distribution shift.

In the third experiment setting (Section 4.3), we study a satellite image classification problem where the baseline model performs poorly under spatiotemporal distribution shifts. We use DISDE to understand why a popular domain adaptation method (DANN) meant to improve performance under distribution shift fails to do so. Our experiments thus demonstrate how DISDE can be applied in various stages of the model development process and also help us understand learned feature

representations.

4.1 Case studies: predicting employment from census data

In the following case studies, we consider natural and semi-synthetic $Y \mid X$ and X shifts on tabular datasets. When some interventions are more expensive than others (e.g., collecting and labeling large quantities of new data vs model changes), it is useful to understand which modeling interventions are most likely to help. By studying specific types of distribution shifts, we illustrate how our diagnostic can guide modeling interventions such as reweighting existing data, using domain knowledge to identify and sample a missing covariate, and collecting new data samples from the target distribution.

Our task is to predict whether an individual is employed (Y) based on their (tabular) census data (X). We use the Adult census dataset [27] to train an employment classifier on one set of data (training), and apply the classifier to a different set (target). We use the decomposition (2.1) and focus on data from 2018. Both the employment model $f(x)$ and domain classifier $\hat{\pi}(x)$ are implemented as random forest classifiers from the `sklearn` Python package [72]. To measure performance degradation due to distribution shift and not to overfitting, all evaluated performances are on validation sets, i.e., data that are separate from the training set but are drawn from the same distribution as the training set. Additional implementation details are in Appendix E.1. Confidence intervals are in Table 2 in Section 5.

4.1.1 $Y \mid X$ shift: missing/unobserved covariates

Due to the cost of collecting and labeling enough data, it may be challenging to train a new model from scratch for each target site. Consider a prediction model trained on data collected from one site (West Virginia), which may be applied to other target sites (Maryland). To construct a specific type of $Y \mid X$ shift between sites, we consider a scenario where the initial feature set in the training data did not include educational attainment data. People in Maryland tend to be more educated than people in West Virginia, and educational attainment affects employment. Consequently, when educational attainment is not included in X, a $Y \mid X$ shift arises from marginalizing out the effect of education over a different distribution in West Virginia versus Maryland.

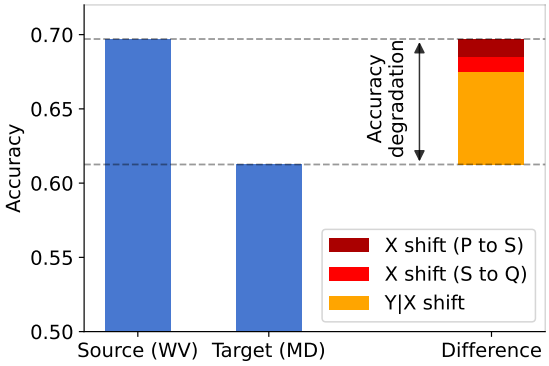
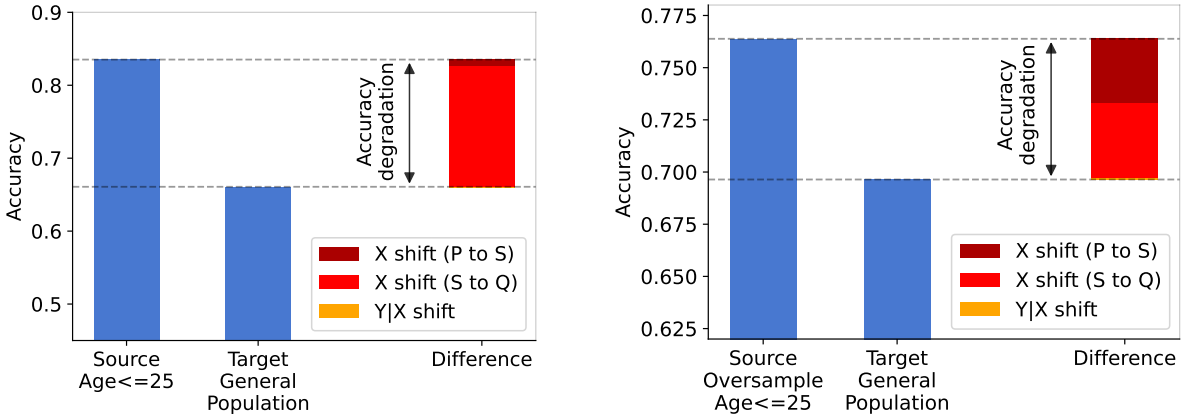


Figure 3. $Y \mid X$ shift: original model trained on West Virginia and evaluated on Maryland

Let us now take the perspective of a modeler who did not know how this distribution shift was constructed. In Figure 3, observe that the prediction model trained in West Virginia performs worse in Maryland than in West Virginia; our goal is to understand this difference in performance. The decomposition depicted in the figure attributes substantial loss to a $Y \mid X$ shift. In response to such a decomposition, one might consider reasons why there is a $Y \mid X$ shift, including the possibility of a shift in the distribution of important unobserved variables. Analysts could use domain knowledge to consider educational attainment as one such unobserved variable, and obtain and append this feature to the original training set. Notably, collecting more features can often be less costly than collecting enough data from Maryland



(a) Model trained on data with age $\leq$25

(b) Model trained on data over-sampling age $\leq$25

Figure 4: Moderate vs. severe X-shift. Both models are evaluated on the general population.

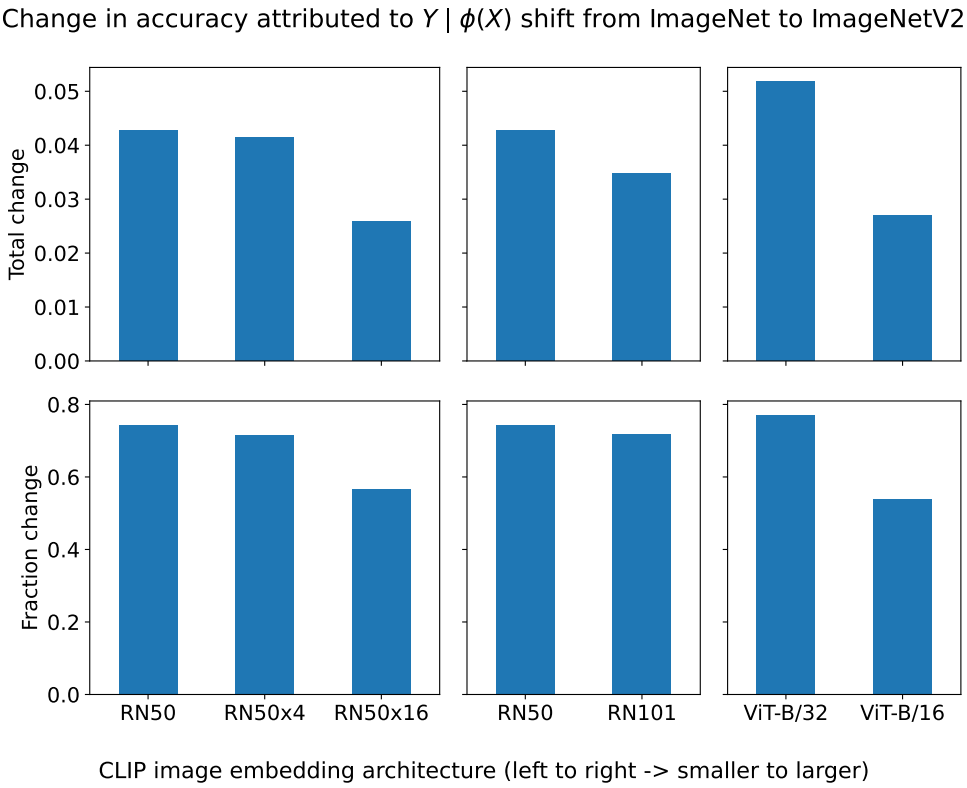
to fit a new model, and has the additional benefit of improving the model accuracy overall. Indeed, when we fit a new model on West Virginia incorporating educational attainment, we find that the new model performs well on both West Virginia and on Maryland, with an accuracy of 70% on Maryland, which is an improvement over the previous 61%.

4.1.2 X shift: selection bias in age

Data can often be collected with unrecognized selection bias. Here, we consider a data collection process that oversamples people up to the age of 25 (e.g. from sampling near a college campus). We expect that it is easier to predict whether people of age ≤ 25 are employed, as most of them are still in school. In contrast, older adults may encounter new situations that affect employment. Let us take the perspective of a modeler who notices that there is a difference between the training data and the general population (target Q) after the model is trained, and who obtains a carefully collected dataset over the general population for evaluation purposes. We demonstrate our approach on two training distributions that differ in their magnitude of selection bias and illustrate how the X shift terms in the decomposition (2.1) can reflect such differences, and consequently inform future modeling interventions.

We first focus on the extreme case where the training data *only* consists of people of up to age 25. In Figure 4a, the trained model performs much worse on the general population, and our decomposition correctly reveals that the decrease in performance is due to a X shift. The performance degradation is primarily attributed to X shift ($S \to Q$), indicating that the model performed poorly on examples in the general population (Q) that were rarely encountered during training. To better understand the shift from P_X to Q_X, we use the feature importance method for random forests [72] on the domain classifier and find that the most important feature, as expected, is age. Our method thus suggests that data collection over an older population may be necessary in order to perform well on the general population. Indeed, a new employment model fitted on more samples from all ages has an improved target accuracy of 73%, compared to the original model's target accuracy of 66%.

Next, we consider a less dramatic shift in the distribution of X, where people up to the age of 25



(a) Total and fraction of change in accuracy attributed to $Y \mid \phi(X)$ for ImageNet vs ImageNetV2 with labels from [90]. Each sub-plot contains a sequence of models of increasing size, from left to right. See Appendix G for full DISDE decompositions.

are merely *overrepresented* in the training data. In Figure 4b, we observe that the model performs worse on the general population as before, and our decomposition again reveals that the decrease in performance is primarily attributed to an X shift. Unlike in the previous example, the training data here includes people of all ages, and we see that the X shift ($P \to S$) term is larger as a result. This suggests that instead of resorting to an expensive data collection process, reweighting the original dataset may be an effective algorithmic solution to adapt to the target. We fit a new model on an appropriately re-weighted version of the original training data and find that the new model achieves an accuracy of 80% on the target, a significant improvement over the previous model's target accuracy of 69%.

4.2 Invariant feature representations in ImageNet vs ImageNetV2

Despite excellent progress on IID datasets, computer vision models perform substantially worse under distribution shift [97]. State-of-the-art supervised models trained on ImageNet [26], a standard image classification benchmark dataset in deep learning, all suffer a universal accuracy drop of 11-14 percentage points on a new dataset, ImageNetV2 [78]. In contrast, human labelers can perform equally well on both datasets [90], suggesting deficiencies in modeling. Similar trends have been widely observed across computer vision and natural language processing (NLP) [62, 97, 89, 63]. In addition, algorithmic robustness interventions do not appear to be a substantial improvement over

supervised models for out-of-domain performance [77]).

In contrast, zero-shot Contrastive Language-Image Pretraining (CLIP [77]) models were recently found to perform much better across a diverse set of datasets for image classification, including ImageNetV2, compared to both supervised models and algorithmic robustness interventions of similar performance on ImageNet. CLIP models use 400 million image-caption pairs to simultaneously learn an image embedding and a language embedding so that the embeddings of an image and its accompanying caption are close, while the embeddings of an image and an unrelated caption are far. These embeddings immediately give rise to a "zero-shot" image classifier: class labels are turned into captions (e.g., "a photo of a Y" where Y is the text label), and the model outputs the class whose caption embedding is closest to that of the image. Using a process of elimination, Fang et al. [34] found that CLIP's exceptional out-of-domain robustness is because of the diversity of the data used in training.

Because of the unprecedented robustness of zero-shot CLIP, we were interested to use our diagnostic to evaluate zero-shot CLIP models and their corresponding representations. We do so by using DISDE on ImageNet and ImageNetV2, with the additional modification that instead of the usual class labels for ImageNet and ImageNetV2, we use the labels constructed by Shankar et al. [90], which ensures that the labeling mechanism $Y \mid X$ is the same across both datasets by construction. Consequently, any dataset shift must be a X shift, and any change in model performance must also be due to X shift. One would hope that better features $\phi(X)$ better capture the relationship among the input X and the label Y across distributions, so that in this setting, decompositions would attribute more performance degradation to $\phi(X)$ shift and less $Y \mid \phi(X)$ shift. In this setting, this property is equivalent to the property of invariance of a predictor

To investigate this, we consider decompositions for several variants of the CLIP model. As CLIP models are flexible and can use various neural network architectures, and as larger architectures have been observed to perform better in various settings, we wonder if larger architectures in zero-shot CLIP might also better capture the relationship between inputs and labels. Indeed, we find that more performance degradation is attributed to $\phi(X)$ shift rather than $Y \mid \phi(X)$ shift for larger architectures, both in absolute and relative terms (Figure 5a). Our decomposition suggests the learned features do a better job of capturing the relationship between input and label as model size increases.

In contrast, this pattern fails to hold for supervised classification models trained on ImageNet. When we apply our decomposition on the last-layer outputs ($\phi(X)$) of supervised ImageNet models, virtually all of the performance degradation is attributed to $Y \mid \phi(X)$ shift, suggesting that features learned from the ImageNet training set alone do not adequately capture the relationship between inputs and labels across distribution shifts. See Appendix G for full DISDE decompositions. See Appendix E.2 for implementation details.

4.3 Diagnosing failures of domain-adaptation methods

We study an image classification problem where the learned feature representation plays a critical role in the distributional robustness of a prediction model. We focus on satellite image classification, which can inform humanitarian and sustainability efforts by tracking deforestation, population density, and other economic metrics [84]. In order to track these metrics effectively, such models must perform reliably over space and time. We consider a spatio-temporal distribution shift for classifying land use from satellite images on a variant of the Functional Map of the World dataset [84], called FMoW-wilds [19]. The training data consists of images before 2013, while the target

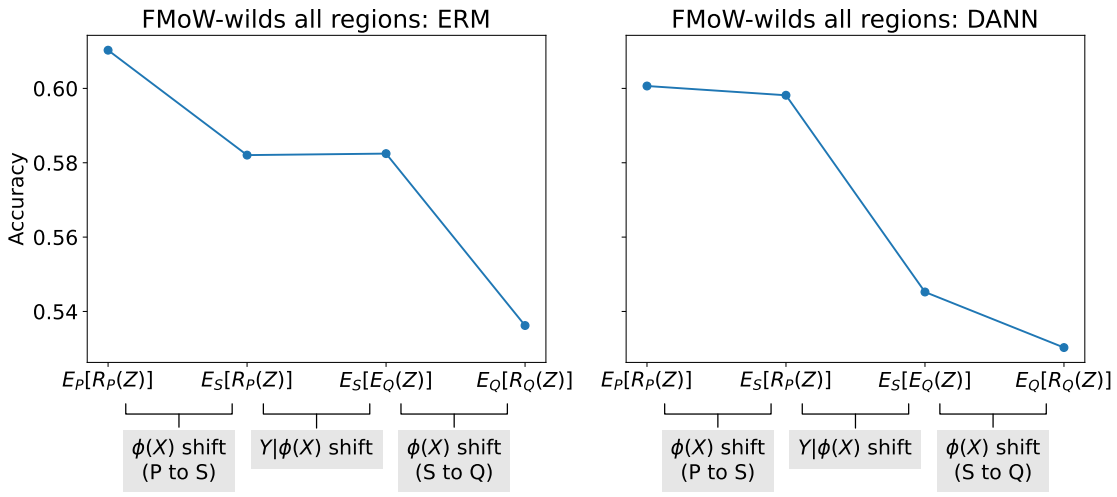


Figure 6. ERM (left) vs DANN (right) in the distribution shift from FMoW-wilds. On the vertical axis is model accuracy; all results hold using accuracy in place of $\ell(\cdot)$. On the horizontal axis are the four values in Equation (2.4). Their consecutive differences are the $Z = \phi(X)$ shift ($P \to S$), $Y \mid \phi(X)$ shift, and $\phi(X)$ shift ($S \to Q$) terms, respectively. Both ERM and DANN perform worse on target than training, and DANN is not better on target data compared to ERM. For ERM, the $\phi(X)$ shift terms are largest; for DANN, the $Y \mid \phi(X)$ terms are largest.

data is from 2016 and after, and the proportion of samples from different geographical regions is different between training and target.

Sagawa et al. [84] recently observed that classifiers trained via the standard empirical risk minimization (ERM) approach (also known as sample average approximation) perform significantly worse on the target distribution compared to training. They also noted that algorithmic interventions do not, in fact, provide meaningful robustness gains over ERM. To illustrate how these methods can fail, we focus on a specific domain adaptation method, called Domain Adversarial Neural Network (DANN) [36]. The main idea of DANN is to learn a feature embedding $\phi(X)$ of X such that the distribution of $\phi(X)$ is difficult to distinguish between training and target. Specifically, DANN uses labeled (X, Y) data from training and unlabeled X data from target to learn neural network feature representations $\phi(X)$ that are not only useful for predicting Y on the training data, but are also difficult to classify as being from training vs target. Intuitively, requiring $\phi(X)$ to be difficult to classify as being from either distribution can reduce $\phi(X)$ shift, and thus also loss attributed to $\phi(X)$ shift.

We investigate why models trained using DANN perform poorly on the target, even though DANN was designed to perform well across a distribution shift. Using the output of the penultimate layer of the neural network as features so that $Z_{\text{ERM}} = \phi_{\text{ERM}}(X)$ and $Z_{\text{DANN}} = \phi_{\text{DANN}}(X)$ for the respective models, we fit $\hat{\pi}$ using a logistic regression and present our decomposition (2.4) in Figure 6. Compared to the ERM model, the DANN counterpart has less performance degradation attributed to $\phi(X)$ shift. However, the performance degradation attributed to $Y \mid \phi(X)$ is much larger for the DANN model, so that DANN performs poorly overall, even though it successfully has less performance degradation attributed to $\phi(X)$ shift. Such a performance degradation can occur if the assumption that the distributions of $\phi(X)$ should be indistinguishable between train and target is inappropriate. This trade-off between different types of distribution shifts highlights

the pitfalls of solely focusing on protecting against $\phi(X)$-shifts, as is common practice in domain adaptation. See Section 5 for confidence intervals.

5 Statistical properties

In this section, we study the statistical properties of the estimators $\hat{\theta}_P$, $\hat{\theta}_Q$ (3.9), and the estimators for the decomposition terms from Equation (2.1) from Algorithm 1. We show asymptotic normality, efficiency, and the validity of bootstrapping methods for estimating confidence intervals. Then, we apply our statistical guarantees to calculate confidence intervals for the experiments in Section 4.

We give conditions under which we know that $\hat{\theta}_P$, $\hat{\theta}_Q$, and the estimators for the decomposition terms from Equation (2.1) from Algorithm 1 are asymptotically normal. Knowing the asymptotic distribution allows us construct confidence intervals for the true parameters θ_P, θ_Q, and the decomposition terms from Equation (2.1). In Appendix C, we use heuristics developed by Kennedy [50] to show that our estimator is asymptotically efficient, suggesting that our estimator cannot be improved upon in terms of asymptotic variance.

Recall our notation from Section 3

$$\alpha^* \text{ is the proportion of the pooled data that comes from } Q_X$$

$$T = \begin{cases} 0 & \text{if } \tilde{X} \text{ is from } P_X \\ 1 & \text{if } \tilde{X} \text{ is from } Q_X \end{cases}$$

$$\pi^*(x) := \mathbb{P}(T = 1 \mid \tilde{X} = x) = \frac{\alpha^* q(x)}{\alpha^* q(x) + (1 - \alpha^*) p(x)}$$

The statistical behavior of $\hat{\theta}_P$, $\hat{\theta}_Q$, and the estimators for the decomposition terms from Equation (2.1) depend on the estimator $\hat{\pi}(x)$ used for the nuisance parameter $\pi^*(x)$. However, as is the case for many other semiparametric estimators, when the nuisance parameter estimates converge sufficiently quickly and smoothly, the asymptotic distributions of $\hat{\theta}_P$ and $\hat{\theta}_Q$ are well-approximated in a way that does not depend on the form of the nuisance parameter estimator [66]. In particular, the nuisance parameters need not be estimated at the $\sqrt{n}$-rate for $\hat{\theta}_P$, $\hat{\theta}_Q$ to be asymptotically normal at the $\sqrt{n}$-rate. We show asymptotic normality of the main estimators when the nuisance parameter $\pi^*(\cdot)$ can only be estimated at the $n^{1/4}$-rate. Key to this, however, is "undersmoothing" of the nuisance parameters [68], as we describe below.

In this section, we focus on understanding the approximation when the nuisance parameter is estimated using nonparametric kernel smoothing. Kernel smoothing is a well-established nonparametric estimation technique [68], where the prediction for a test point x is a locally-weighted average of the outcomes for observations "near" x in the training data. The weighting is defined by a kernel function $K(\cdot)$ and a bandwidth $\sigma > 0$, and the weighting is done using $K_\sigma(t) := \sigma^{-1} K(t/\sigma)$. Applying this approach to estimate $\pi^*(\cdot)$ gives

$$\hat{\pi}(x) = \frac{\frac{1}{n}\sum_{i=1}^n T_i K_\sigma(x - X_i)}{\frac{1}{n}\sum_{i=1}^n K_\sigma(x - X_i)}.$$

While other estimators will give similar results under different assumptions about the data generating process, focusing on this one allows us to take advantage of existing technical results for kernel smoothing and to focus our attention on the behavior of the semiparametric estimators themselves.

We write the kernel smoothing estimate $\hat{\pi}(\cdot)$ as the ratio of two kernel density estimators: let

$A = [1, T]^\top$ and

$$\hat{\gamma}(x) = \frac{1}{n}\sum_{i=1}^{n} AK_\sigma(x - X_i) \tag{5.1}$$

so that $\hat{\gamma}$ is an estimator for the true γ^* with $\gamma_j^*(x) := m_X(x)\mathbb{E}[A_j \mid X = x]$, with $\gamma_1^*(x) = m_X(x)$ the marginal density of z across the pooled data from P and Q, and $\gamma_2^*(x) = m_X(x)\mathbb{E}[T \mid X = x]$. Then $\hat{\pi}(x) = \hat{\gamma}_2(x)/\hat{\gamma}_1(x)$.

To get good rates of convergence, we need to be careful with choosing the kernel and bandwidth.

Assumption A (Kernel density estimator assumptions)**.** *Let d be the dimension of X. For some choice of constants k and p,*

1. *$K(u)$ is differentiable of order p, the derivatives of order p are bounded, $K(u)$ is zero outside of a bounded set, $\int K(u)du = 1$, there is a positive integer k such that for all $j < k$, $\int K(u)[\bigotimes_{l=1}^{j} u]du = 0$.*
2. *The bandwidth $\sigma = \sigma(n)$ satisfies $n\sigma^{2d+4p}/(\ln n)^2 \to \infty$ and $n\sigma^{2k} \to 0$.*

Remark 1. *If the bandwidth $\sigma = \sigma(n) = n^b$, then for Assumption A to hold, we would need*

$$-\frac{1}{2r+4d} < b < -\frac{1}{2k}.$$

This requires the bandwidth to shrink faster than the rate used to get minimax optimal convergence of the nuisance parameters themselves with kernel smoothing [94, 68]. As a result, the pointwise variance of the nuisance parameter estimates will be higher, and the bias will be lower. This turns out to be critical for getting $\sqrt{n}$-consistency of the semiparametric estimators $\hat{\theta}_Q$ and $\hat{\theta}_P$, and is one widely applicable lesson that applies broadly to most applicable methods for nuisance parameter estimation in our setting. Therefore, we recommend users to choose hyperparameters that lead to mild, yet notiable undersmoothing when fitting nuisance parameters for our method. Unfortunately, reliable methods for automatically choosing hyperparameters to achieve these requirements are currently not well-known in the literature.

To ensure that the estimates from nonparametric kernel smoothing converge sufficiently quickly over the entire domain, we additionally make the following assumptions.

Assumption B.

1. *X has support on a compact set $\mathcal{X}$, on which its density $m_X(x)$ is bounded above and below (away from 0): there are $B_{mL} > 0$ and $B_{mU} < \infty$ such that $B_{mL} < m_X(x) < B_{mU}$ for all x.*
2. *$\pi^*(x) := \mathbb{P}(T = 1 \mid X = x)$ is bounded away from 0 and 1: there is a $\delta_\pi > 0$ such that $0 < \delta_\pi < \pi^*(x) < 1 - \delta_\pi < 1$ for almost every x.*[2]
3. *$\pi^*(x)$ is continuous almost everywhere.*
4. *There is a version of $\gamma^*(x)$ that is continuously differentiable to order p with bounded derivatives on an open set containing $\mathcal{X}$.*

[2] As a consequence, $\alpha := \mathbb{P}(T = 1)$ also satisfies $0 < \delta_\pi < \alpha < 1 - \delta_\pi < 1$.

Assumptions like these are standard for kernel smoothing estimators [68, 103], and are necessary to ensure convergence over the entire domain.

With these assumptions, the nuisance parameters will not typically be estimated at the parametric $\sqrt{n}$ rate. However, when plugged into our semiparametric method, the overall estimators $\hat{\theta}_P$ and $\hat{\theta}_Q$ will converge at a $\sqrt{n}$ rate, and be asymptotically normal, so long as the loss $L = \ell(f(X), Y)$ is sufficiently regular. We describe sufficient regularity conditions for L in the assumption below.

Assumption C.

1. $\mathbb{E}[L^4] < B_{L^4}$ *for some* $B_{L^4} < \infty$.
2. $\mu_Q(x) := \mathbb{E}[L \mid T = 1, X = x]$ *and* $\mu_P(x) := \mathbb{E}[L \mid T = 0, X = x]$ *are continuous almost everywhere.*
3. $\mu_Q(x), \mu_P(x) < B_\mu < \infty$ *for some* B_μ *on* $\mathcal{X}$.

Because our estimators (3.9) depend on the product of functions of $\hat{\pi}(\cdot)$ and the loss L, Assumption C ensures that small estimation errors in $\hat{\pi}(\cdot)$ aren't amplified in pathological ways when multiplied by L. With these three assumptions, our estimators are jointly regular and asymptotically linear. Our main result (Theorem 2 to come) makes precise the asymptotic distribution of the estimator $\hat{\theta}_Q$. A symmetric result holds for $\hat{\theta}_P$, and from these we can deduce the asymptotic distribution of the decomposition terms in Equation (2.1).

To state Theorem 2, we introduce additional notation. Building on Equation (3.4), let $M(\cdot)$ denote the distribution of the observed data. Let $M_X(\cdot)$ be the marginal distribution of X under this distribution, and $m_X(x)$ its density. As before, define T as a binary dummy variable that is 1 when the observation comes from the target distribution Q and 0 when it comes from the training distribution P. Then as in Section 3 we can write

$$\theta_Q = \mathbb{E}_M\left[\frac{\mathrm{d}S_X}{\mathrm{d}M_X}\frac{\mathbf{1}\{T=1\}}{\mathbb{P}_M(T=1 \mid X)}L\right] \quad \text{and} \quad \theta_P = \mathbb{E}_M\left[\frac{\mathrm{d}S_X}{\mathrm{d}M_X}\frac{\mathbf{1}\{T=0\}}{\mathbb{P}_M(T=0 \mid X)}L\right].$$

Recalling that $\pi^*(x) := \mathbb{P}_M(T = 1 \mid X = x)$,[3] and using Equation (2.2) and some calculations,

$$\frac{\mathrm{d}S_X}{\mathrm{d}M_X} \propto \lambda(\pi^*(X), \alpha^*) \quad \text{where} \quad \lambda(\pi, \alpha) = \frac{\pi(1-\pi)}{(1-\alpha)\pi + \alpha(1-\pi)}. \tag{5.2}$$

Then we can write our estimands θ_P, θ_Q as functionals of the data distribution M, the conditional probability $\pi^*(X)$, and the marginal probability $\alpha^* = \mathbb{P}_M(T = 1)$ as

$$\theta_P = \frac{\mathbb{E}_M\left[L\frac{1-T}{1-\pi^*(X)}\lambda(\pi^*(X), \alpha^*)\right]}{\mathbb{E}_M\left[\frac{1-T}{1-\pi^*(X)}\lambda(\pi^*(X), \alpha^*)\right]} \quad \text{and} \quad \theta_Q = \frac{\mathbb{E}_M\left[L\frac{T}{\pi^*(X)}\lambda(\pi^*(X), \alpha^*)\right]}{\mathbb{E}_M\left[\frac{T}{\pi^*(X)}\lambda(\pi^*(X), \alpha^*)\right]}. \tag{5.3}$$

For brevity, let W denote (X, T, L). Then we show the asymptotic linearity of $\hat{\theta}_Q$:

Theorem 2. *Let* $\hat{\gamma}$ *be as in Equation* (5.1). *Then, under Assumptions A, B, and C,*

$$\sqrt{n}(\hat{\theta}_Q - \theta_Q) = \sum_{i=1}^{n} \psi_Q(w_i)/\sqrt{n} + o_P(1),$$

[3] In this section we write $\pi^*, \gamma^*, \alpha^*$ to denote true values of nuisance parameters.

so that

$$\sqrt{n}(\hat{\theta}_Q - \theta_Q) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\psi_Q(W))).$$

where

$$\begin{aligned} D_Q &= \mathbb{E}\left[\frac{T}{\pi^*(X)}\lambda(\pi^*(X), \alpha^*)\right] \\ D_Q\psi_Q(w) &= (l - \theta_Q)\frac{t}{\pi^*(x)}\lambda(\pi^*(x), \alpha^*) \\ &\quad + \mathbb{E}\left[(L - \theta_Q)\frac{T}{\pi^*(X)}\frac{\partial}{\partial\alpha}\lambda(\pi^*(X), \alpha)\Big|_{\alpha=\alpha^*}\right](t - \alpha^*) \\ &\quad + (\mu_Q(x) - \theta_Q)\pi^*(x)(t - \pi^*(x))\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi, \alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}. \end{aligned}$$

See Appendix A for the proof. An analogous result holds for $\hat{\theta}_P$, and the decomposition terms from Equation (2.1) follows immediately as $\hat{\theta}_P$, $\hat{\theta}_Q$ are asymptotically linear. Theorem 2 can be further generalized: Yadlowsky [108] shows that the assumptions needed to guarantee asymptotic linearity with the same influence function extend immediately to cross-fitted analogues of these estimators (see Appendix D.1 for more on cross-fitting). Additionally, the literature has many results showing asymptotic linearity of semiparametric methods using other nuisance parameter estimators under more general conditions than those assumed in the proofs provided here [67, 18]. This is because many of these results have a similar form, with the first order effect of estimating the nuisance parameters is the influence function in Theorem 2, and the remaining error terms that depend on the specific choice of nuisance parameter estimator are lower order. See, for example, Newey [66, 67], Kennedy [50], Yadlowsky [108].

The asymptotic linearity in our result directly ensures that subsampling procedures such as half-sampling [20] are justified for constructing calibrated confidence intervals. The half-sample bootstrap [30, 74] is a simple procedure of this form, which Yadlowsky et al. [109, Lemma 4] showed directly applies here (Corollary 1). The half-sample bootstrap is constructed by repeatedly drawing samples of half the data without replacement, constructing the estimator $\hat{\theta}^*_{P,Q}$ with each sample, and analyzing the distribution of the errors $\hat{\theta}^*_{P,Q} - \hat{\theta}_{P,Q}$. Our next result shows that this provides be a good approximation of the sampling distribution of $\hat{\theta}_{Q,P} - \theta_{P,Q}$. In this corollary (of Yadlowsky et al. [109, Lemma 4]), P^* refers to the observed Monte Carlo distribution of the procedure, and is conditional on the observed data (and therefore, $\hat{\theta}_{P,Q}$).

Corollary 1. *Under the conditions of Theorem 2, conditionally on the observed data, the Monte Carlo distribution of* $\sqrt{n}(\hat{\theta}^*_Q - \hat{\theta}_Q) \overset{P^*}{\to} N(0, \mathrm{Var}(\psi_Q(W)))$.

Confidence intervals in practice Showing that the nonparametric bootstrap [30] gives valid confidence intervals requires additional work. However, given the regularity of our functional, we conjecture that the nonparametric bootstrap also provides calibrated confidence intervals. The advantage of the nonparametric bootstrap is its simplicity and reduced sensitivity to finite sample defects resulting from each subsample being larger in size. In this subsection, we empirically verify that the nonparametric bootstrap provides similar inference to the theoretically-justified half-sample bootstrap, so we recommend this procedure for routine practice.

		Standard error		
Term	Estimate	NP boot	HS boot	IF est
X shift ($P \to S$)	0.78%	0.66%	0.71%	0.96%
$Y \mid X$ shift	6.29%	0.97%	1.17%	1.13%
X shift ($S \to Q$)	1.35%	0.68%	0.75%	0.86%

(a) $Y|X$ shift: original model trained on West Virginia and evaluated on Maryland

		Standard error		
Term	Estimate	NP boot	HS boot	IF est
X shift ($P \to S$)	0.23%	0.38%	0.43%	0.53%
$Y \mid X$ shift	-0.01%	0.43%	0.53%	0.56%
X shift ($S \to Q$)	17.21%	0.33%	0.34%	0.40%

(b) X shift: model trained on only age $\leq$25 and evaluated on general population

		Standard error		
Term	Estimate	NP boot	HS boot	IF est
X shift ($P \to S$)	2.03%	0.79%	0.67%	1.13%
$Y \mid X$ shift	0.27%	0.97%	1.05%	1.26%
X shift ($S \to Q$)	5.15%	0.64%	0.63%	0.89%

(c) X shift: model trained on over-sampling age $\leq$25 and evaluated on general population

Table 2. Point estimate and estimates of standard error for estimators of decomposition terms for Section 4.1. The standard error column names are abbreviated: "NP boot" is short for nonparametric boostrap and "HS boot" is short for half-sample bootstrap, as described in Section 5. Each type of bootstrapping is done with 500 bootstrap re-samples. "IF est" is short for plug-in estimator of influence function: we calculate the standard error using influence functions as in results like Theorem 2 and plug in estimates $\hat{\pi}(x), \hat{\mu}_Q(x), \hat{\alpha}$ in place of $\pi^*(x), \mu_Q(x), \alpha^*$.

In Table 2, we first compare different ways of calculating confidence intervals for Section 4.1 using methods from Section 5. Next, we calculate nonparametric bootstrap confidence intervals for Section 4.3. The confidence intervals in Table 3 show that the conclusions drawn from our previous analysis are not due to happenstance and are statistically sound.

	ERM		DANN	
Term	Estimate	NP boot SE	Estimate	NP Boot SE
$\phi(X)$ shift $(P \to S)$	2.82%	0.15%	0.25%	0.13%
$Y \mid \phi(X)$ shift	-0.04%	0.40%	5.29%	0.40%
$\phi(X)$ shift $(S \to Q)$	4.62%	0.15%	1.49%	0.13%

Table 3. Estimates for ERM vs DANN, corresponding to Figure 6. The standard errors are calculated using non-parametric bootstrap with 100 bootstrap samples ("NP boot SE")

6 Discussion

Distribution shift is a fundamental problem in data-driven decision-making. Still, there are limited tools for understanding changes in model performance with respect to different types of distribution shifts. We introduce a simple method, DIstribution Shift DEcomposition (DISDE) that can diagnose out-of-distribution performance by attributing the change in loss to three terms, corresponding to 1) an increase in harder but frequently seen examples from training, 2) changes in the relationship between features and outcomes, and 3) poor performance on examples infrequent or unseen during training (Sections 2, 3). We empirically demonstrate how DISDE can inform modeling improvements on tabular and image data, and also provide insight on the invariance properties of newer models (Section 4), and we analyze its asymptotic properties (Section 5). In this final section, we situate the current work in the large body of literature on distribution shifts (Section 6.1) and discuss the limitations of our approach alongside future research directions (Section 6.2).

6.1 Related work

Distribution shift is an important topic across multiple research communities. Much of the work is on training better models through methodological improvements. Like ours, some works focus on taxonomies for discussing distribution shifts [60, 87, 101]. Complementary to our work, other works focus on collecting data across distribution shifts. However, there is not much research on attributing changes in model performance to different types of distribution shifts. Concurrent to our work, Zhang et al. [112] use Shapley value to attribute changes in performance to various types of shift, but do not address differences in support or density between distributions as we do with the shared distribution in DISDE (Section 2).

In the machine learning community, domain adaptation methods train models using data from a training distribution, and aim to perform well on a specified target distribution. Standard approaches assume the $Y \mid X$ distribution is fixed and reweight training data to resemble the target distribution [91, 43, 11, 96, 102]. When we expect shifts in the $Y \mid X$ distribution, Meinshausen and Bühlmann [61], Rothenhäusler et al. [82] propose learning models that have good performance against causal interventions that affect $Y \mid X$. More recently, several authors have studied approaches that

aim to learn feature representations $\phi(X)$ such that $Y \mid \phi(X)$, or functionals thereof, is invariant across multiple environments [73, 83, 2, 81].

There is a large body of work on distributionally robust optimization (DRO) methods that optimize performance over distributions that are "close" to the training distribution. While the majority of these works consider shifts in the joint distribution (X, Y) [32, 25, 107, 8, 55, 65, 9, 54, 93, 9, 104, 29], some notions of closeness (e.g., Wasserstein distance) can flexibly handle both joint and covariate shifts [88, 12, 37, 13, 53]. Despite the extensive literature on DRO, limited work study particular types of distribution shifts. Focusing on the case of covariate shifts, Duchi et al. [28] develop tailored methods that optimize worst-case subpopulation shifts over a set of covariates. Similarly, Sahoo et al. [85] propose a related worst-case subpopulation formulation over $Y \mid X$-shifts, holding the marginal distribution of X fixed.

The causal inference community takes a more nuanced approach to distribution shift. There is extensive work on sensitivity analysis (e.g., [79, 80, 48, 35, 49, 110]), which studies the amount of unobserved confounding—analogous to the magnitude of $Y \mid X$ shift in predictive scenarios—to call into question the conclusion of observational studies. When a known target population differs from the study population, a large body of research adjusts observations to resemble the target [95, 98, 51, 58]. These approaches are analogous to domain adaptation under covariate shift; analogous to distributionally robust optimization, Jeong and Namkoong [47] recently formulate worst-case sensitivity approaches for covariate shift. More generally, the external validity of a study can be called into question due to different distribution shifts. Egami and Hartman [31] formalize and contextualize the type of shifts that naturally arise in social sciences.

Our approach complements design-based and benchmarking research in statistics and machine learning. Collecting maximally heterogeneous data is a classical idea in experimental design, with a recent focus on multi-site designs [22, 5, 38, 24, 100, 99]. In machine learning, building on older literature on distribution shifts [40, 76], there is renewed interest in benchmarking model performance on unseen test distributions different from training [78, 52, 41, 62]. For example, Recht et al. [78], Taori et al. [97] observe a trend in which models trained on the same data universally suffer a relative performance degradation on new distributions, regardless of the training method and model capacity. By evaluating model robustness over multiple distributions, these works spur empirical progress by simulating typical ML deployment processes where models encounter a priori unknown distributions under operation. While out-of-distribution benchmarks and industrial monitoring systems primarily focus on detecting performance degradation [44], we build a diagnostic that generates qualitative insights on the cause of the observed failure. We hope this work provides a starting point for building a principled language for understanding performance degradation over distribution shifts.

6.2 Limitations and future work

Our work has several limitations, some of which are by design. For example, by design, DISDE does not directly propose new interventions; it is a diagnostic that can help decide between potential interventions, or to help assess existing interventions. In addition, DISDE only applies in certain settings: for now, it only concerns shifts between pairs of distributions, and not among more distributions. It also only applies to losses that can be written as $\mathbb{E}[\ell(f(X), Y)]$. Therefore, it does not directly apply to an F1-score or a worst-subgroup loss, for example. Lastly, our work assumes the existence of a shared distribution S_X (Section 2). If there is no shared support between P_X and Q_X, then our decomposition will not be valid.

Other limitations of our approach concern the limitations of the work that it builds on. In particular, our methodology involves an auxiliary domain classifier, which is analogous to the propensity score in causal inference. How best to learn a propensity score is a field of active research. For example, there are questions on the role of balancing covariates between treatment and control, versus modeling treatment assignment [45, 113, 15]. In our work we use a very simple way to learn a propensity score and leave extensions to future work.

Broadly speaking, our work highlights the potential benefits of an operations engineering approach to the training, deployment, and use of machine learning models. Classical quality control and reliability engineering methods have had major impacts across multiple industries (e.g., car manufacturing [75], electronics manufacturing [64], software development processes [71]). As the use of ML models becomes increasingly prevalent and complex, we will need ways to diagnose model failures and to direct resources to the appropriate modeling interventions. However, currently, even the most advanced industrial monitoring systems can only detect performance degradation, without attribution to its cause [44]. We hope this paper spurs further work to build rigorous and scalable tools for the quality management of ML applications. For example, to analyze the local sensitivity of a simulation output, several authors have used conditional variances to attribute variance due to different subsets of the input [42, 86, 69, 92] and related approaches may yield fruit in the context of ML models.

7 Acknowledgements

Tiffany Cai is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2036197. Hongseok Namkoong was partially supported by the Amazon Research Award.

References

[1] E. Amorim, M. Cançado, and A. Veloso. Automated essay scoring in the presence of biased ratings. In *Association for Computational Linguistics (ACL)*, pages 229–237, 2018.

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893 [stat.ML]*, 2019.

[3] S. Asmussen and P. W. Glynn. *Stochsatic Simulation: Algorithms and Analysis*. Springer, 2007.

[4] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[5] A. Banerjee, D. Karlan, and J. Zinman. Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, 7(1):1–21, 2015.

[6] S. Beery, E. Cole, and A. Gjoka. The iwildcam 2020 competition dataset. *arXiv:2004.10340 [cs.CV]*, 2020.

[7] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 20*, pages 137–144, 2007.

[8] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013.

[9] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming, Series A*, 167(2):235–292, 2018.

[10] P. Bickel, C. A. J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer Verlag, 1998.

[11] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[12] J. Blanchet, Y. Kang, F. Zhang, and K. Murthy. Data-driven optimal transport cost selection for distributionally robust optimizatio. *arXiv:1705.07152 [stat.ML]*, 2017.

[13] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[14] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[15] A. Chattopadhyay, C. H. Hase, and J. R. Zubizarreta. Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, 39(24), 2020. doi: https://doi.org/10.1002/sim.8659. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8659.

[16] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi. Ethical machine learning in health care. *arXiv:2009.10576 [cs.SY]*, 2020.

[17] M. S. Chen, P. N. Lara, J. H. Dang, D. A. Paterniti, and K. Kelly. Twenty years post-NIH revitalization act: enhancing minority participation in clinical trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014.

[18] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[19] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[20] E. Chung and J. P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484 – 507, 2013.

[21] M. Colombo, J. Gondzio, and A. Grothey. A warm-start approach for large-scale stochastic linear programs. *Mathematical Programming*, 127(2):371–397, 2011.

[22] G. Cruces and S. Galiani. Fertility and female labor supply in latin america: New causal evidence. *Labour Economics*, 14(3):565–573, 2007.

[23] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research, 2006.

[24] R. Dehejia, C. Pop-Eleches, and C. Samii. From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics*, 39(1):217–243, 2021.

[25] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[26] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[27] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems 32*, 34, 2021.

[28] J. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 2022.

[29] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406, 2021.

[30] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, 1982. doi: 10.1137/1.9781611970319.

[31] N. Egami and E. Hartman. Elements of external validity: Framework, design, and analysis. *American Political Science Review*, page 1–19, 2022.

[32] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski. Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. *IEEE Transactions on Signal Processing*, 52(8): 2177–2188, 2004.

[33] C. Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[34] A. Fang, G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, and L. Schmidt. Data determines distributional robustness in contrastive language-image pre-training (clip). In *icml22*, 2022.

[35] C. B. Fogarty. Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, pages 1–13, 2019.

[36] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[37] R. Gao, X. Chen, and A. Kleywegt. Wasserstein distributional robustness and regularization

in statistical learning. *arXiv:1712.06050 [cs.LG]*, 2017.

[38] P. Gertler, M. Shah, M. L. Alzua, L. Cameron, S. Martinez, and S. Patil. How does health promotion work? evidence from the dirty business of eliminating open defecation. Technical report, National Bureau of Economic Research, 2015.

[39] R. Gutman, E. Karavani, and Y. Shimoni. Propensity score models are better when post-calibrated. *arXiv:2211.01221 [stat.ML]*, 2022. URL https://arxiv.org/abs/2211.01221.

[40] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

[41] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, S. Dawn, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[42] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, 1996.

[43] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 20*, pages 601–608, 2007.

[44] C. Huyen. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly, 2022.

[45] K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[46] G. Imbens and D. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[47] S. Jeong and H. Namkoong. Assessing external validity over worst-case subpopulations. *arXiv:2007.02411 [stat.ML]*, 2020.

[48] N. Kallus and A. Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems 31*, pages 9269–9279, 2018.

[49] N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[50] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv:2203.06469 [stat.ME]*, 2022.

[51] H. L. Kern, E. A. Stuart, J. Hill, and D. P. Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1):103–127, 2016.

[52] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, S. Beery, et al. Wilds: A benchmark of in-the-wild distribution

shifts. *arXiv:2012.07421 [cs.LG]*, 2020.

[53] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[54] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[55] H. Lam and E. Zhou. Quantifying input uncertainty in stochastic optimization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

[56] A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. Gimenez, C. de Masson d'Autume, T. Kocisky, S. Ruder, D. Yogatama, K. Cao, S. Young, and P. Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *nips21*, 2021.

[57] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[58] C. R. Lesko, A. L. Buchanan, D. Westreich, J. K. Edwards, M. G. Hudgens, and S. R. Cole. Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass.)*, 28(4):553, 2017.

[59] F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[60] Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *ICML18*, 2018.

[61] N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

[62] J. Miller, K. Krauth, B. Recht, and L. Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR, 2020.

[63] J. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[64] M. A. Miner. Cumulative Damage in Fatigue. *Journal of Applied Mechanics*, 12(3):A159–A164, 1945.

[65] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing with virtual adversarial training. *arXiv:1507.00677 [stat.ML]*, 2015.

[66] W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382, 1994.

[67] W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.

[68] W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

[69] A. B. Owen. Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.

[70] A. B. Owen. *Monte Carlo Theory, Methods, and Examples.* 2015. Online at http://statweb.stanford.edu/~owen/mc/.

[71] M. Paulk, B. Curtis, M. Chrissis, and C. Weber. Capability maturity model, version 1.1. *IEEE Software*, 10(4):18–27, 1993.

[72] F. Pedregosa, 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[73] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78 (5):947–1012, 2016.

[74] J. Praestgaard and J. A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability*, 21(4):2053–2086, 1993.

[75] T. Pyzdek and P. A. Keller. *Quality engineering handbook.* CRC Press, 2003.

[76] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning.* MIT Press, 2008.

[77] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[78] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[79] P. R. Rosenbaum. *Design of Observational Studies.* Springer Series in Statistics. Springer, 2010.

[80] P. R. Rosenbaum. A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics*, 67(3):1017–1027, 2011.

[81] E. Rosenfeld, P. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *Proceedings of the Ninth International Conference on Learning Representations*, 2021.

[82] D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv:1801.06229 [stat.ME]*, 2018.

[83] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new

domains. In *Proceedings of the European Conference on Computer Vision*, pages 213–226. Springer, 2010.

[84] S. Sagawa, P. W. Koh, T. Lee, I. Gao, S. M. Xie, K. Shen, A. Kumar, W. Hu, M. Yasunaga, H. Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. In *Advances in Neural Information Processing Systems 21*, 2021.

[85] R. Sahoo, L. Lei, and S. Wager. Learning from a biased sample. *arXiv:2209.01754 [stat.ME]*, 2022.

[86] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[87] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1255–1262, 2012.

[88] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584, 2015.

[89] V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt. Do image classifiers generalize across time? *arXiv:1906.02168 [cs.LG]*, 2019.

[90] V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt. Evaluating machine accuracy on ImageNet. In *icml20*, 2020.

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

[92] E. Song, B. L. Nelson, and J. Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016.

[93] M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9131–9141, 2019.

[94] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982.

[95] E. A. Stuart, S. R. Cole, C. P. Bradshaw, and P. J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

[96] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[97] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems 20*, 2020.

[98] E. Tipton. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38 (3):239–266, 2013.

[99] E. Tipton and R. B. Olsen. A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, 47(8):516–524, 2018.

[100] E. Tipton and L. R. Peck. A design-based approach to improve external validity in welfare policy evaluations. *Evaluation review*, 41(4):326–356, 2017.

[101] D. Tran, J. Liu, M. W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu, N. Band, T. G. J. Rudner, K. Singhal, Z. Nado, J. van Amersfoort, A. Kirsch, R. Jenatton, N. Thain, H. Yuan, K. Buchanan, K. Murphy, D. Sculley, Y. Gal, Z. Ghahramani, J. Snoek, and B. Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions. *arxiv:2207.07411 [stat.ML]*, 2022.

[102] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

[103] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[104] B. P. Van Parys, P. M. Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.

[105] A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021.

[106] M. Wortsman, G. Ilharco, M. Li, J. W. Kim, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt. Robust fine-tuning of zero-shot models. *arXiv:2109.01903 [cs.CV]*, 2021.

[107] H. Xu, C. Caramanis, and S. Mannor. A distributional interpretation of robust optimization. *Mathematics of Operations Research*, 37(1):95–110, 2012.

[108] S. Yadlowsky. On cross-fitting with plug-in estimators, Oct 2022. URL https://www.syadlowsky.com/blog/semiparametric/2022/10/24/on-cross-fitting-with-plug-in-estimators.html.

[109] S. Yadlowsky, S. Fleming, N. Shah, E. Brunskill, and S. Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv:2111.07966 [stat.ML]*, 2021.

[110] S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, and L. Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *Annals of Statistics*, 50(5): 2587–2615, 2022.

[111] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[112] H. Zhang, H. Singh, M. Ghassemi, and S. Joshi. "why did the model fail?": Attributing model performance changes to distribution shifts, 2022. URL https://arxiv.org/abs/2210.10769.

[113] Q. Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of*

Statistics, 47(2):965–993, 2019.

[114] W. Zheng and M. J. van der Laan. *Cross-Validated Targeted Minimum-Loss-Based Estimation*, pages 459–474. Springer New York, New York, NY, 2011. ISBN 978-1-4419-9782-1. doi: 10.1007/978-1-4419-9782-1_27. URL https://doi.org/10.1007/978-1-4419-9782-1_27.

A Proofs for asymptotic properties for the estimator

In this section, we will prove Theorem 2, to show the asymptotically linear expansion of $\hat{\theta}_Q$. To do so, we will first use the following lemma to show that it suffices to find the asymptotically linear expansion of the numerator and denominator of $\hat{\theta}_Q$ separately.

Lemma 1. *Assume that* $A_n \xrightarrow{p} A$, $B_n \xrightarrow{p} B$, $\sqrt{n}(A_n - A) = \frac{1}{\sqrt{n}}\sum_{i=1}^n a_i + o_P(1)$ *and* $\sqrt{n}(B_n - B) = \frac{1}{\sqrt{n}}\sum_{i=1}^n b_i + o_P(1)$. *Additionally, assume that* $B > 0$. *Then,*

$$\sqrt{n}\left(\frac{A_n}{B_n} - \frac{A}{B}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{1}{B}a_i - \frac{A}{B^2}b_i\right) + o_P(1).$$

Proof of this lemma is deferred to Section B.1.

With this lemma in mind, observe that the numerator and denominator of θ_Q (and also $\hat{\theta}_Q$) are nearly identical: we can think of the denominator D_Q (and also $\hat{D}_Q Q$) as being a special case of the numerator N_Q (and also $\hat{N}_Q$) but with $L := 1$:

$$D_Q = \mathbb{E}\left[\frac{T}{\pi(X)}\lambda(\pi(X), \alpha)\right] \quad \text{and} \quad N_Q = \mathbb{E}\left[\frac{T}{\pi(X)}\lambda(\pi(X), \alpha)L\right].$$

Because of Lemma 1, it suffices to show the following proposition, which we will show in the rest of this section:

Proposition 3. $\sqrt{n}(\hat{N}_Q - N_Q) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{N_Q}(w_i) + o_P(1)$, *for*

$$N_Q := \mathbb{E}\left[L\frac{T}{\pi^*(X)}\lambda(\pi^*(X), \alpha^*)\right] \quad \textit{and} \quad \hat{N}_Q := \sum_{i=1}^n l_i \frac{t_i}{\hat{\pi}(x_i)}\lambda(\hat{\pi}(x_i), \hat{\alpha}) \tag{A.1}$$

for the choice $\psi_{N_Q}(w) = g(w) + h(w) + \delta(w)$, *where*

$$\begin{aligned}
g(w) &= l\frac{t}{\pi^*(x)}\lambda(\pi^*(x), \alpha^*) - N_Q \\
h(w) &= \mathbb{E}\left[L\frac{T}{\pi^*(X)}\frac{\partial}{\partial\alpha}\lambda(\pi^*(X), \alpha)\Big|_{\alpha=\alpha^*}\right](t - \alpha^*) \\
\delta(w) &= \mu_Q(x)\pi^*(x)(t - \pi^*(x))\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi, \alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}.
\end{aligned}$$

To begin, recall that $\pi(x) = \gamma_2(x)/\gamma_1(x)$ and define

$$g(w, \gamma, \alpha) = l\frac{t}{\pi(x)}\lambda(\pi(x), \alpha) - N_Q \tag{A.2}$$

so that

$$\hat{N}_Q - N_Q = n^{-1}\sum_{i=1}^n g(w_i, \hat{\gamma}, \hat{\alpha}).$$

Since g is continuously differentiable with respect to α, expand $\hat{\alpha}$ around α^* as

$$g(w_i, \hat{\gamma}, \hat{\alpha}) = \nabla_\alpha g(w_i, \hat{\gamma}, \overline{\alpha})(\hat{\alpha} - \alpha^*) + g(w_i, \hat{\gamma}, \alpha^*)$$

where $\overline{\alpha}$ is a mean value. By Lemma 2 below, $n^{-1}\sum_{i=1}^n \nabla_\alpha g(w, \hat{\gamma}, \alpha) \xrightarrow{p} \mathbb{E}[\nabla_\alpha g(w, \gamma^*, \alpha)]$.

Lemma 2. *Under the assumptions and definitions of Theorem 2,*

$$n^{-1}\sum_{i=1}^{n}\nabla_\alpha g(w,\hat{\gamma},\overline{\alpha}) \xrightarrow{p} \mathbb{E}[\nabla_\alpha g(w,\gamma^*,\alpha^*)] =: G_\alpha$$

Proof of this lemma is deferred to Section B.2. Noting that $\hat{\alpha} = n^{-1}\sum_{i=1}^{n} t_i$ and letting

$$\begin{aligned} h(w) &:= \mathbb{E}[\nabla_\alpha g(w,\gamma^*,\alpha^*)](t-\alpha^*) \\ &= \mathbb{E}\left[L\frac{T}{\pi^*(X)}\nabla_\alpha\lambda(\pi^*(X),\alpha^*)\right](t-\alpha^*) \end{aligned}$$

gives

$$\sum_{i=1}^{n} g(w_i,\hat{\gamma},\hat{\alpha})/\sqrt{n} = \sum_{i=1}^{n}[h(w_i)+g(w_i,\hat{\gamma},\alpha^*)]/\sqrt{n} + o_P(1). \tag{A.3}$$

To find an asymptotically linear representation of $\sum_{i=1}^{n} g(w_i,\hat{\gamma},\alpha^*)/\sqrt{n}$, we will show that under the assumptions made in Theorem 2, $\sum_{i=1}^{n} g(w_i,\hat{\gamma},\alpha^*)/\sqrt{n}$ satisfies the assumptions of Newey and McFadden [68, Theorem 8.11], which we restate now for convenience. Note that while the theorem as stated in Newey and McFadden [68] only shows asymptotic normality, in the proof they show asymptotic linearity, so we have modified the result to state the asymptotic linearity result, instead. We have also modified the result for the case where g is linear in the estimated parameter, in which case consistency of the estimated parameter does not need to be shown [68].

Assumption D (Newey and McFadden [68, Assumptions 8.1-8.3 and assumption from Theorem 8.11])**.** *Let d be the dimension of X.*

1. *$K(u)$ is differentiable of order p, the derivatives of order p are bounded, $K(u)$ is zero outside of a bounded set, $\int K(u)du = 1$, there is a positive integer m such that for all $j < m$, $\int K(u)[\bigotimes_{l=1}^{j} u]du = 0$.*
2. *There is a version of $\gamma^*(x)$ that is continuously differentiable to order p with bounded derivatives on an open set containing $\mathcal{X}$.*
3. *There is $r \geq 4$ such that $\mathbb{E}[\|A\|^r] < \infty$ and $\mathbb{E}[\|A\|^r|X=x]m_X(x)$ is bounded.*
4. *The bandwidth $\sigma = \sigma(n)$ satisfies $n\sigma^{2d+4p}/(\ln n)^2 \to \infty$ and $n\sigma^{2m} \to 0$.*

Assumption E (Newey and McFadden [68, assumptions from Theorem 8.11])**.** *Let β be the parameter of interest, and $\hat{\beta}$ its estimator where $\hat{\beta} - \beta = \frac{1}{n}\sum_{i=1}^{n} g(W_i,\hat{\gamma})$.*

1. $\mathbb{E}[g(W,\gamma^*)] = 0$
2. $\mathbb{E}[\|g(W,\gamma^*)\|^2] < \infty$
3. *$\mathcal{X}$ is a compact set.*

Assumption F (Newey and McFadden [68, enumerated assumptions from Theorem 8.11])**.** *There is a vector of functionals $G(w,\gamma)$ that is linear in γ such that*

(i) *for* $\|\gamma - \gamma^*\|$ *small where the norm is the Sobolev norm*[4], $\|g(w,\gamma) - g(w,\gamma^*) - G(w,\gamma-\gamma^*)\| \leq b(w)\|\gamma-\gamma^*\|^2$, *and* $\mathbb{E}[b(W)] < \infty$;

(ii) $\|G(w,\gamma)\| \leq c(w)\|\gamma\|$ *and* $\mathbb{E}[c(W)^2] < \infty$;

(iii) *there is* $v(x)$ *with* $\int G(w,\gamma)dM(w) = \int v(x)\gamma(x)dx$ *for all* $\|\gamma\| < \infty$;

(iv) $v(x)$ *is continuous almost everywhere,* $\int \|v(x)\|dx < \infty$, *and there is* $\epsilon > 0$ *such that* $\mathbb{E}[\sup_{\|\nu\|\leq\epsilon} \|v(X+\nu)\|^4] < \infty$.

Theorem 4 (N+M Theorem 8.11)**.** *Let* γ^* *be the nuisance parameter and* $\hat{\gamma}$ *its kernel density estimate satisfying Assumption D. Let* β *be the parameter of interest, and* $\hat{\beta}$ *its estimator satisfying Assumption E. Assume there is a vector of functionals* $G(w,\gamma)$ *satisfying Assumption F. Then for* $\delta(w) = v(x)a - \mathbb{E}[v(x)a]$,

$$\sum_{i=1}^n g(w_i,\hat{\gamma})/\sqrt{n} = \sum_{i=1}^n [g(w_i,\gamma^*) + \delta(w_i)]/\sqrt{n} + o_P(1).$$

We now proceed to use this result to prove the asymptotically linear representation

$$\sum_{i=1}^n g(w_i,\hat{\gamma},\alpha^*)/\sqrt{n} = \sum_{i=1}^n [g(w_i,\gamma^*,\alpha^*) + \delta(w_i)]/\sqrt{n} + o_P(1) \tag{A.4}$$

for our choice of $g(w,\gamma,\alpha)$. Then, the desired result follows from combining Equations (A.4) and (A.3), and then applying Lemma 1. What remains is to check the conditions of Theorem 4.

Verifying Assumption D:

1. Assumed in Assumption A
2. Assumed in Assumption B
3. There is $r \geq 4$ such that $\mathbb{E}[\|A\|^r] < \infty$ and $\mathbb{E}[\|A\|^r | X = x]m_X(x)$ is bounded: recall that $A = [1,T]$ and T takes values in $\{0,1\}$. Then let $r = 4$, $\|A\|^r \leq 2$, so that $\mathbb{E}[\|A\|^r] \leq 2 < \infty$, and and $\mathbb{E}[\|A\|^r \mid X = x]m_X(x) \leq 2B_{mU} < \infty$ by Assumption B.
4. Assumed in Assumption A

Verifying Assumption E:

1. $\mathbb{E}[g(W,\gamma^*,\alpha^*)] = 0$: this holds by definition of g (Equation (A.2)) and N_Q (Equation (A.1)).
2. $\mathbb{E}[|g(W,\gamma^*,\alpha^*)|^2] < \infty$:

$$\mathbb{E}[|g(W,\gamma^*,\alpha^*)|^2] = \mathbb{E}\left[\left|L\frac{T}{\pi^*(X)}\lambda(\pi^*(X),\alpha^*) - N_Q\right|^2\right].$$

This is finite since $N_Q < \infty$, and then by Cauchy-Schwarz since $\mathbb{E}[L^4] < B_{L^4}$ by Assumption C, and $\left|\frac{T}{\pi^*(X)}\lambda(\pi^*(X),\alpha^*)\right|^4$ is bounded, since $T \in \{0,1\}$ and

$$\frac{\lambda(\pi^*(X),\alpha^*)}{\pi^*(X)} = \frac{1-\pi^*(X)}{(1-\alpha^*)\pi^*(X) + \alpha^*(1-\pi^*(X))} \leq \frac{1-\delta_\pi}{\delta_\pi}$$

[4] $\|\gamma\| := \max_{\ell\leq p}\sup_{x\in\mathcal{X}}\|\partial^\ell\gamma(x)/\partial^\ell x\|$

with δ_π from Assumption B.

3. Assumed in Assumption B.

Verifying Assumption F:

(i) For $\|\gamma-\gamma^*\|$ small[5], $\left|g(w,\gamma,\alpha^*) - g(w,\gamma^*,\alpha^*) - G(w,\gamma-\gamma^*)\right| \le b(w)\|\gamma-\gamma^*\|^2$, and $\mathbb{E}[b(W)] < \infty$: By Taylor-expanding g around $\gamma^*(x)$ (where in the following equations we abuse notation and also use g to mean $g(w,\gamma(x),\alpha)$, where the second argument is the value of the function γ evaluated at x, rather than the function γ. We also write $\nabla_\gamma g(\cdot)$ to denote the derivative of this new g with respect to its second argument, the value $\gamma(x)$, rather than to the function $\gamma(\cdot)$, and similarly for $\nabla^2_\gamma g(\cdot)$),

$$\begin{aligned} g(w,\gamma,\alpha^*) = g(w,\gamma^*(x),\alpha^*) + \nabla_\gamma g(w,\gamma^*(x),\alpha^*)^\top(\gamma(x)-\gamma^*(x)) \\ + \frac{1}{2}(\gamma(x)-\gamma^*(x))^\top \nabla^2_\gamma g(w,\overline{m},\alpha^*)(\gamma(x)-\gamma^*(x)) \end{aligned}$$

where $\overline{m}$ is a mean value on the line between $\gamma(x)$ and $\gamma^*(x)$. Note that g is twice differentiable with respect to $\gamma(x)$ in an open set containing this line (since $\|\gamma-\gamma^*\|$ small, so that π is bounded away from 0 and 1) so that the expansion holds. Thus let $G(w,\gamma-\gamma^*)$ be the first-order term in the expansion above:

$$\begin{aligned} G(w,\gamma) &:= \nabla_\gamma g(w,\gamma(x),\alpha^*)^\top \gamma(x) \\ &= lt \frac{\partial}{\partial \pi}\left(\frac{\lambda(\pi,\alpha^*)}{\pi}\right)\Big|_{\pi=\pi^*(x)} \nabla_\gamma \pi(\gamma^*(x))^\top \gamma(x) \\ &= lt \frac{\partial}{\partial \pi}\left(\frac{\lambda(\pi,\alpha^*)}{\pi}\right)\Big|_{\pi=\pi^*(x)} m_X(x)^{-1}[-\pi^*(x),1]\gamma(x) \end{aligned}$$

where we used $\pi(\gamma(x)) = \gamma_2(x)/\gamma_1(x)$, $\gamma_1(x) = m_X(x)$ is the marginal density of X, and

$$\frac{\partial}{\partial \pi}\left(\frac{\lambda(\pi,\alpha^*)}{\pi}\right)\Big|_{\pi=\pi^*(x)} = -\frac{1-\alpha^*}{((1-\alpha^*)\pi^*(x)+\alpha^*(1-\pi^*(x)))^2}.$$

Then to verify the assumption,

$$\begin{aligned} &\left|g(w,\gamma,\alpha^*) - g(w,\gamma^*,\alpha^*) - G(w,\gamma-\gamma^*)\right| \\ &= \frac{1}{2}\left|(\gamma(x)-\gamma^*(x))^\top \nabla^2_\gamma g(w,\overline{m},\alpha^*)(\gamma(x)-\gamma^*(x))\right| \\ &\le \frac{1}{2}\left\|\nabla^2_\gamma g(w,\overline{m},\alpha^*)\right\|_F \|\gamma-\gamma^*\|^2. \end{aligned}$$

Thus to verify the assumption, let $b(w) = \frac{1}{2}\left\|\nabla^2_\gamma g(w,\overline{m},\alpha^*)\right\|_F$. Some simple calculus shows that $\nabla^2_\gamma g(w,\overline{m},\alpha^*) = lt\phi(\overline{m})$ where

$$\phi(\overline{m}) = \frac{\partial^2}{\partial \pi^2}\left(\frac{\lambda(\pi,\alpha^*)}{\pi}\right)\Big|_{\pi=\overline{\pi}} \nabla_\gamma \pi(\overline{m})\nabla_\gamma \pi(\overline{m})^\top + \frac{\partial}{\partial \pi}\left(\frac{\lambda(\pi,\alpha^*)}{\pi}\right)\Big|_{\pi=\overline{\pi}} \nabla^2_\gamma \pi(\overline{m}).$$

Then $\mathbb{E}[\|\phi(\overline{m})\|_F^2]$ is bounded: since $\|\gamma-\gamma^*\|$ is small, $\overline{m}$ is close to $\gamma^*(x)$, so that $\overline{m}_1$ is bounded away from 0, and $\overline{\pi}$ is bounded away from both 0 and 1, so that each term in $\phi(\overline{m})$

[5] the norm on $\|\gamma-\gamma^*\|$ is the Sobolev norm

is bounded, so that $\mathbb{E}[\|\phi(\overline{m})\|_F^2]$ is bounded. Then, applying Cauchy-Schwarz, $\mathbb{E}[b(W)] \leq \frac{1}{2}\sqrt{\mathbb{E}[(LT)^2]\mathbb{E}[\|\phi(\overline{m})\|_F^2]} < \infty$ as $\mathbb{E}[(LT)^2] \leq \mathbb{E}[L^2] \leq \mathbb{E}[L^4] < B_{L^4}$ by Assumption C.

(ii) $|G(w,\gamma)| \leq c(w)\|\gamma\|$ and $\mathbb{E}[c(w)^2] < \infty$:

$$\begin{aligned}
|G(w,\gamma)| &= lt \left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi=\pi^*(x)} m_X(x)^{-1}[-\pi^*(x), 1]\gamma(x) \right| \\
&\leq lt \left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi=\pi^*(x)} \right| m_X(x)^{-1} \|[-\pi^*(x), 1]\| \|\gamma(x)\| \\
&\leq lt \left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi=\pi^*(x)} \right| m_X(x)^{-1} \sup_{x \in \mathcal{X}} \left\{ \|[-\pi^*(x), 1]\| \right\} \|\gamma\| \\
&\leq ltC\|\gamma\|
\end{aligned}$$

for some constant C since $\left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi=\pi^*(x)} \right|$ is bounded as in the previous part, and $m_X(x)^{-1}, \pi^*(x)$ are bounded by Assumption B. Here, the norm $\|\gamma\|$ is the Sobolev norm while $\|\gamma(x)\|$ is the Euclidean norm. Thus let $c(w) = ltC$, and $\mathbb{E}[c(W)^2] = C^2\mathbb{E}[L^2T^2] \leq C^2\sqrt{\mathbb{E}[(LT)^4]} \leq C^2\sqrt{\mathbb{E}[L^4]} \leq C^2\sqrt{B_{L^4}} < \infty$ by Assumption C.

(iii) There is $v(x)$ with $\int G(w,\gamma)dM(w) = \int v(x)\gamma(x)dx$:

Define $v(x)$ by rewriting $\int G(w,\gamma)dM(w)$:

$$\begin{aligned}
\int G(w,\gamma)dM(w) &= \int lt \frac{\partial}{\partial \pi} \left[\frac{\lambda(\pi, \alpha^*)}{\pi} \right] \Big|_{\pi=\pi^*(x)} m_X(x)^{-1}[-\pi^*(x), 1]\gamma(x)dM(w) \\
&= \int \mu_Q(x)\pi^*(x) \frac{\partial}{\partial \pi} \left[\frac{\lambda(\pi, \alpha^*)}{\pi} \right] \Big|_{\pi=\pi^*(x)} [-\pi^*(x), 1]\gamma(x)dx
\end{aligned}$$

where the last equality is essentially obtained by using iterated expectations to rewrite

$$\mathbb{E}[LT\xi(X)] = \mathbb{E}[\mathbb{E}[L \mid T = 1, X]\mathbb{E}[T \mid X]\xi(X)]$$

since $T \in \{0, 1\}$, for $\xi(x) = \frac{\partial}{\partial \pi} \left[\frac{\lambda(\pi, \alpha^*)}{\pi} \right] \Big|_{\pi=\pi^*(x)} [-\pi^*(x), 1]\gamma(x)$ so that

$$v(x) = \mu_Q(x)\pi^*(x) \frac{\partial}{\partial \pi} \left[\frac{\lambda(\pi, \alpha^*)}{\pi} \right] \Big|_{\pi=\pi^*(x)} [-\pi^*(x), 1].$$

(iv) $v(x)$ is continuous almost everywhere, $\int \|v(x)\|dx < \infty$, and there is $\epsilon > 0$ such that $\mathbb{E}[\sup_{\|\nu\| \leq \epsilon} \|v(X + \nu)\|^4] < \infty$:

$v(x)$ is continuous almost everywhere since it is the product of functions that are continuous almost everywhere: $\pi^*(x)$ and $\mu_Q(x)$ are continuous almost everywhere by Assumptions B and C, and $\frac{\partial}{\partial \pi} \left[\frac{\lambda(\pi, \alpha^*)}{\pi} \right] \Big|_{\pi=\pi^*(x)}$ is also continuous in x.

$\int \|v(x)\|dx < \infty$: $\|v(x)\|$ is bounded on $\mathcal{X}$ since it is the product of several terms that are each bounded on $\mathcal{X}$: $\mu_Q(x)$ is bounded by Assumption C, $\pi^*(x)$ is bounded by Assumption B, and $\frac{\partial}{\partial \pi} \left[\frac{\lambda(\pi, \alpha^*)}{\pi} \right] \Big|_{\pi=\pi^*(x)}$ is bounded since $\delta_\pi < \pi^*(x) < 1 - \delta_\pi$ for $\delta_\pi > 0$ from in Assumption B. The integral is finite since $\mathcal{X}$ is compact.

The sup condition is satisfied since $\|v(x)\|$ is bounded in $\mathcal{X}$.

Now that we have satisfied the conditions of Theorem 4 and we have $v(x)$, let $a = [1, t]^\top$ and define

$$\begin{aligned}\delta(w) &= v(x)a - \mathbb{E}[v(X)A] \\ &= \mu_Q(x)(t - \pi^*(x))\pi^*(x)\frac{\partial}{\partial \pi}\left[\frac{\lambda(\pi, \alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}.\end{aligned}$$

Then by Theorem 4,

$$\sum_{i=1}^{n} g(w_i, \hat{\gamma}, \alpha^*)/\sqrt{n} = \sum_{i=1}^{n}[g(w_i, \gamma^*, \alpha^*) + \delta(w_i)]/\sqrt{n} + o_P(1)$$

as desired.

B More proofs for asymptotic properties of the estimator

B.1 Proof of Lemma 1

Define $r(a, b) = a/b$. It is continuously differentiable, so we can apply the mean value theorem to its derivative, so that for some choices of values $\overline{A}, \overline{B}$ between A_n and A, and B_n and B, respectively,

$$\begin{aligned}\sqrt{n}\left(\frac{A_n}{B_n} - \frac{A}{B}\right) &= \sqrt{n}(r(A_n, B_n) - r(A, B)) \\ &= \sqrt{n}\nabla r(\overline{A}, \overline{B})^\top \begin{bmatrix} A_n - A \\ B_n - B \end{bmatrix} \\ &= [1/\overline{B}, -\overline{A}/\overline{B}^2] \begin{bmatrix} \sum_{i=1}^n a_i/\sqrt{n} + o_P(1) \\ \sum_{i=1}^n b_i/\sqrt{n} + o_P(1) \end{bmatrix}.\end{aligned}$$

Since $A_n \xrightarrow{p} A$ and $B_n \xrightarrow{p} B$, we have that $\overline{A} \xrightarrow{p} A$ and $\overline{B} \xrightarrow{p} B$, so

$$\sqrt{n}\left(\frac{A_n}{B_n} - \frac{A}{B}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{B}a_i - \frac{A}{B^2}b_i\right) + o_P(1).$$

B.2 Proof of Lemma 2

We will use the following from Newey and McFadden [68]:

Lemma 3 (Direct consequence of Newey and McFadden [68, Lemma 8.10])**.** *If Assumption D is satisfied, then* $\sqrt{n}\|\hat{\gamma} - \gamma^*\|^2 \xrightarrow{p} 0$.

Since we are assuming Assumption D for showing Theorem 2, the conclusion holds.

Now, since $\frac{1}{n}\sum_{i=1}^n \nabla_\alpha g(w, \gamma^*, \alpha^*) \xrightarrow{p} \mathbb{E}[\nabla_\alpha g(w, \gamma^*, \alpha^*)]$ by law of large numbers, it suffices to show

$$\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_\alpha g(w_i, \hat{\gamma}, \overline{\alpha}) - \nabla_\alpha g(w_i, \gamma^*, \alpha^*)\right] \xrightarrow{p} 0.$$

Using the Markov inequality, it suffices to show that for $\|\hat{\gamma} - \gamma^*\|$ small enough,

$$\left|\nabla_\alpha g(w, \hat{\gamma}, \overline{\alpha}) - \nabla_\alpha g(w, \gamma^*, \alpha^*)\right| \leq b(w) \left[\|\hat{\gamma} - \gamma^*\| + \|\hat{\alpha} - \alpha^*\|\right]$$

for some $b(w)$ such that $\mathbb{E}[b(W)] < \infty$, since $\|\hat{\gamma} - \gamma^*\| \xrightarrow{p} 0$ and $\|\hat{\alpha} - \alpha^*\| \xrightarrow{p} 0$. Similar to verifying Assumption F (where again we abuse notation to write g on the RHS to take $\gamma(x)$ as the second argument, rather than $\gamma(\cdot)$, and we write ∇_γ to denote derivatives with respect to $\gamma(x)$ instead of $\gamma(\cdot)$),

$$\nabla_\alpha g(w, \hat{\gamma}, \overline{\alpha}) - \nabla_\alpha g(w, \gamma^*, \alpha^*) = \nabla_{\gamma,\alpha} \nabla_\alpha g(w, \overline{m}, \overline{\overline{\alpha}})^\top \begin{bmatrix} \hat{\gamma}(x) - \gamma^*(x) \\ \overline{\alpha} - \alpha^* \end{bmatrix}$$

where $(\overline{m}, \overline{\overline{\alpha}})$ is a mean value on the line between $(\hat{\gamma}(x), \overline{\alpha})$ and $(\gamma^*(x), \alpha^*)$. Then, note that $\hat{\gamma}(x) - \gamma^*(x) \leq \|\hat{\gamma} - \gamma^*\|$. Let

$$b(w) := \sup_{\overline{m}, \overline{\overline{\alpha}}} \|\nabla_{\gamma,\alpha} \nabla_\alpha g(w, \overline{m}, \overline{\overline{\alpha}})\|_\infty,$$

so that

$$\left|\nabla_\alpha g(w, \hat{\gamma}, \overline{\alpha}) - \nabla_\alpha g(w, \gamma^*, \alpha^*)\right| \leq b(w) \left[\|\hat{\gamma} - \gamma^*\| + \|\hat{\alpha} - \alpha^*\|\right].$$

Finally, $\mathbb{E}[b(W)] < \infty$ by an argument similar to the verification of Assumption F.

C Proofs for efficiency

We verify that our estimator in Theorem 2 attains the nonparametric efficiency bound, i.e. that our estimator has the best possible asymptotic variance. Instead of rigorously deriving the nonparametric efficiency bound, for brevity, we use heuristics from [50] to obtain the efficient influence function of the estimand. To show the desired result, we then show the efficient influence function of the estimand is the same as the influence function of the estimator in Theorem 2.

C.1 Efficient influence function

We calculate the efficient influence function for the estimand using heuristics from [50]. To simplify notation, we first deal with only the numerator of θ_Q, which we denoted as

$$N_Q = \mathbb{E}_M \left[\ell(f(X), Y) \frac{T}{\pi^*(X)} \lambda(\pi^*(X), \alpha^*)\right] = \mathbb{E}_M \left[\mu_Q(X) \lambda(\pi^*(X), \alpha^*)\right].$$

To calculate the efficient influence function, as in [50], we treat $\mathcal{X}$ as discrete, and use derivative rules with simple influence functions as building blocks. For notational simplicity, we omit *'s for π^*, α^* for the calculation of efficient influence functions.

$$\begin{aligned}
\mathbb{IF}\{\alpha\} &= T - \alpha \\
\mathbb{IF}\{p(x)\} &= \mathbf{1}\{X = x\} - p(x) \\
\mathbb{IF}\{p(t)\} &= \mathbf{1}\{T = t\} - p(t) \\
\mathbb{IF}\{\pi(x)\} &= \frac{\mathbf{1}\{X = x\}}{p(x)} (T - \pi(x)) \\
\mathbb{IF}\{\mu_Q(x)\} &= \frac{\mathbf{1}\{X = x, T = 1\}}{\mathbb{P}(X = x, T = 1)} (L - \mu_Q(x))
\end{aligned}$$

$$
\begin{aligned}
&= \frac{T\mathbf{1}\{X=x\}}{p(x)\pi(x)}(L-\mu_Q(x)) \\
\mathbb{IF}\{\lambda(\pi(x),\alpha)\} &= \nabla_\pi\lambda(\pi(x),\alpha)\mathbb{IF}\{\pi(x)\} + \nabla_\alpha\lambda(\pi(x),\alpha)\mathbb{IF}\{\alpha\} \\
&= \nabla_\pi\lambda(\pi(x),\alpha)\frac{\mathbf{1}\{X=x\}}{p(x)}(T-\pi(x)) + \nabla_\alpha\lambda(\pi(x),\alpha)(T-\alpha)
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathbb{IF}\{N_Q\} &= \sum_{x\in\mathcal{X}} \mu_Q(x)\lambda(\pi(x),\alpha)p(x) \\
&= \sum_{x\in\mathcal{X}} \left(\mathbb{IF}\{\mu_Q(x)\}\lambda(\pi(x),\alpha)p(x) + \mu_Q(x)\mathbb{IF}\{\lambda(\pi(x),\alpha)\}p(x) + \mu_Q(x)\lambda(\pi(x),\alpha)\mathbb{IF}\{p(x)\}\right) \\
&= \sum_{x\in\mathcal{X}} \frac{T\mathbf{1}\{X=x\}}{p(x)\pi(x)}(L-\mu_Q(x))\lambda(\pi(x),\alpha)p(x) \\
&\quad + \sum_{x\in\mathcal{X}} \mu_Q(x)\nabla_\pi\lambda(\pi,\alpha)\frac{\mathbf{1}\{X=x\}}{p(x)}(T-\pi(x))p(x) \\
&\quad + \sum_{x\in\mathcal{X}} \mu_Q(x)\nabla_\alpha\lambda(\pi(x),\alpha)(T-\alpha)p(x) \\
&\quad + \sum_{x\in\mathcal{X}} \mu_Q(x)\lambda(\pi(x),\alpha)(\mathbf{1}\{X=x\}-p(x)) \\
&= (L-\mu_Q(X))\frac{T}{\pi(X)}\lambda(\pi(X),\alpha) \\
&\quad + \mu_Q(X)(T-\pi(X))\nabla_\pi\lambda(\pi,\alpha) \\
&\quad + (T-\alpha)\mathbb{E}\left[\mu_Q(x)\nabla_\alpha\lambda(\pi(x),\alpha)\right] \\
&\quad + \mu_Q(X)\lambda(\pi(X),\alpha) - N_Q.
\end{aligned}
$$

C.2 Comparison

Now we show the efficient influence function from the previous section is the same as the influence function of the estimator in Theorem 2. We do so by first comparing the efficient influence function of the numerator of θ_Q, $\mathbb{IF}(N_Q)$, with the influence function of the estimator for the numerator of θ_Q, $\psi_{N1}(w)$. Then we do the same for the denominator, then θ_Q, then θ_P, then the terms in the decomposition (2.1). We start with the numerator of θ_Q, N_Q. Recall that from Proposition 3,

$$
\sqrt{n}(\hat{N}_Q - N_Q) = \sum_{i=1}^{n} \psi_{N1}(w_i)/\sqrt{n} + o_P(1)
$$

where $\psi_{N1}(w) = g(w) + h(w) + \delta(w)$ with

$$
\begin{aligned}
g(w) &= l\frac{t}{\pi^*(x)}\lambda(\pi^*(x),\alpha^*) - N_Q \\
h(w) &= \mathbb{E}\left[L\frac{T}{\pi^*(X)}\nabla_\alpha\lambda(\pi^*(X),\alpha^*)\right](t-\alpha^*) \\
\delta(w) &= \mu_Q(x)\pi^*(x)(t-\pi^*(x))\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}.
\end{aligned}
$$

Note that we can also express $\delta(w)$ as

$$\begin{aligned}\delta(w) &= \mu_Q(x)(t - \pi^*(x)) \left(\nabla_\pi \lambda(\pi^*(x), \alpha^*) - \frac{\lambda(\pi^*(x), \alpha^*)}{\pi^*(x)} \right) \\ &= \mu_Q(x) \left(\lambda(\pi^*(x), \alpha^*) - \frac{t}{\pi^*(x)} \lambda(\pi^*(x), \alpha^*) + (t - \pi^*(x)) \nabla_\pi \lambda(\pi^*(x), \alpha^*) \right).\end{aligned}$$

It is clear that $\psi_{N1}(W)$ is the same as $\mathbb{IF}\{N_Q\}$ from the previous section. Then

$$\sqrt{n}(\hat{N}_Q - N_Q) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{N1}(w_i) + o_P(1) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\mathbb{IF}\{N_Q\}))$$

where $\psi_{N1}(w) := g(w) + h(w) + \delta(w)$ as before. An analogous argument applies to the denominator D_Q, as D_Q is the same as N_Q but with L replaced by 1. Then similar to before,

$$\sqrt{n}(\hat{D}_Q - D_Q) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{D1}(w_i) + o_P(1) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\mathbb{IF}\{D_Q\})).$$

Then to see that

$$\sqrt{n} \left(\frac{\hat{N}_Q}{\hat{D}_Q} - \frac{N_Q}{D_Q} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{1}{D_Q} \psi_{N1}(w_i) - \frac{N_Q}{D_Q^2} \psi_{D1}(w_i) \right) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\mathbb{IF}\{N_Q/D_Q\})),$$

note that $\mathbb{IF}\{N_Q/D_Q\}$ is composed of $\mathbb{IF}\{N_Q\}$ and $\mathbb{IF}\{D_Q\}$ the same way that ψ_1 is composed of ψ_{N1} and ψ_{D_Q}:

$$\mathbb{IF}\{N_Q/D_Q\} = \frac{D_Q \mathbb{IF}\{N_Q\} - N_Q \mathbb{IF}\{D_Q\}}{D_Q^2} = [1/D_Q, -N_Q/D_Q^2] \begin{bmatrix} \mathbb{IF}\{N_Q\} \\ \mathbb{IF}\{D_Q\} \end{bmatrix}$$

which is the same as for $\hat{\theta}_Q = \hat{N}_Q/\hat{D}_Q$ as in Lemma 1:

$$\sqrt{n}(\hat{N}_Q/\hat{D}_Q - N_Q/D_Q) = [1/D_Q, -N_Q/D_Q^2] \begin{bmatrix} \sum_{i=1}^n \psi_{N1}(w_i)/\sqrt{n} + o_P(1) \\ \sum_{i=1}^n \psi_{D1}(w_i)/\sqrt{n} + o_P(1) \end{bmatrix}.$$

The results for the decomposition terms in Equation (2.1) follow similarly.

D Additional details for estimation algorithm

D.1 Data splits and cross-fitting

As is standard practice, performance for the model $f(\cdot)$ is not evaluated on training data so that we measure loss degradation only from distribution shift, rather than also from overfitting. Thus, the data from training distribution P has separate training and validation splits, the evaluated model f is trained on the training split, and evaluated on the validation split. The evaluated model is also evaluated on the data from target distribution Q (which consists of only one split). We measure the change in performance on Q vs on the validation split of P.

The practice of evaluating models on data separate from the data they were trained on also applies to the domain classifier classifier $\hat{\pi}(x)$. One natural option that uses all of the available data is called cross-fitting [114, 18, 108] in which the data are first split into K folds. Then, for each fold k, the domain classifier is learned on the union of all the other folds, and finally evaluated on fold k.

There are various ways to set up the data splits. We use two different ways of splitting data in Sections 4.1 and 4.3 and describe them in Appendix E.1.2 and E.3.2, respectively.

D.2 Practical considerations for learning the domain classifier

To estimate $\hat{\pi}(x) = \mathbb{P}(T = 1|X = x)$, we train classifier on samples from P_X and Q_X using logistic loss, since the true $\pi(x)$ minimizes the expected logistic loss, and furthermore minimizers of the logistic loss produce values in $[0, 1]$, in contrast to squared loss.

To perform model checking, since the auxiliary domain classifier is analogous to a propensity score in causal inference, the usual checks on propensity scores can be used to check validity of the domain classifier. These include checks for balance, overlap, and calibration [46, 39]. It is also useful to check moments: for example, since $\mathbb{E}[\pi(X)] = \mathbb{P}(T = 1) = \mathbb{E}[T]$, one should confirm that the sample mean of $\hat{\pi}(X)$ is also close to the sample mean of T.

E Additional details for experiments

E.1 Adult dataset experiment details

E.1.1 Models

We use random forest classifiers using default settings from the `sklearn` python package to fit both the employment model and the domain classifier $\hat{\pi}(x)$. The employment model has parameter `max_depth` set to 2.

E.1.2 Data splits

Each of train and target datasets are split into an 80%-20% split. The model to be evaluated is trained on the 80% split of the train dataset. The domain classifier is trained and the decomposition is evaluated on the 20% split of the train dataset and all of the target dataset, using cross-fitting with 3 splits as described in Section D.1. This way, both the model to be evaluated and the domain classifier are evaluated on data on which they are not trained.

E.1.3 Additional data processing

The X features are all of those in the Adult dataset unless otherwise specified, but with some of them discretized and made into a binary encoding:

- SCHL (educational attainment) was made into the following binary outcomes:
  - Whether someone finished high school
  - Whether someone finished college
  - Whether someone finished a post-grad degree
- MIL (military) was made into the following:
  - Whether someone is active military
  - Whether someone is a veteran

- CIT (citizenship) was made into the following:
  - Whether someone is born in the US
  - Whether someone is born in a US territory
  - Whether someone has American parents
  - Whether someone is naturalized
  - Whether someone is not a citizen
- MIG (mobility) was made into the following:
  - Whether someone moved residence
- MAR (marriage) was made into the following:
  - Whether someone is married

E.2 ImageNet vs ImageNetV2 experiment details

E.2.1 Models

The zero-shot CLIP models evaluated are those from the CLIP python package. The domain classifiers are logistic regressions on top of the corresponding CLIP image feature embeddings, trained using `sklearn` [72].

E.2.2 Data and data splits

We use the validation split for ImageNet and ImageNetV2. We use cross-fitting with three splits, so that the domain classifier is evaluated on a different cross-fitting split from which it was trained.

Because the zero-shot CLIP models evaluated are zero-shot, i.e. not trained on either ImageNet or ImageNetV2, we do not have to worry about evaluating zero-shot CLIP models on data on which it was trained.

E.3 FMoW-wilds experiment details

E.3.1 Models

The ERM and DANN models being evaluated are the ones trained in the WILDS papers and downloaded from the WILDS leaderboard. The domain classifiers are logistic regressions on top of last layer features from ERM and DANN, trained using `sklearn` [72]. For both ERM and DANN, the WILDS leaderboard provides three models each with different random seed, with the best hyperparameter settings. We repeat the decomposition for each seed, and present the average of the results. Reported bootstrap standard errors are thus standard errors of this average.

E.3.2 Data splits

The ERM and DANN models being evaluated have been trained on the `training` data split for ERM and DANN, and also on `test-unlabeled` for DANN. The domain classifiers are cross-fitted (with 2 splits) as described in Section D.1 on `training` and `test`. Then our decomposition is

evaluated on `test` and the union of `id_test` and `id_val`, which are separate from but drawn from the same distribution as `training`. Because we used cross-fitting, we also evaluate the domain classifier on a different cross-fitting split from which it was trained.

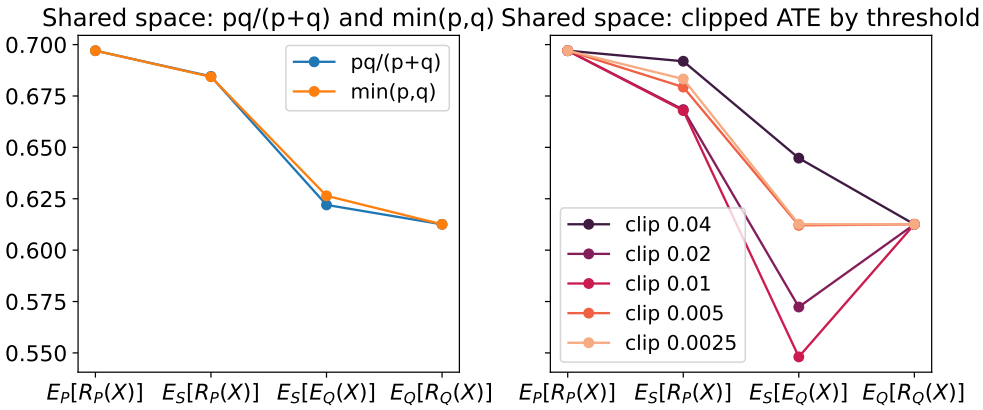
Similar to Appendix E.1.2, both the model to be evaluated and the domain classifier are evaluated on data on which they are not trained, even though the data splits are different from those in Appendix E.1.2.

F Alternative definitions of shared space

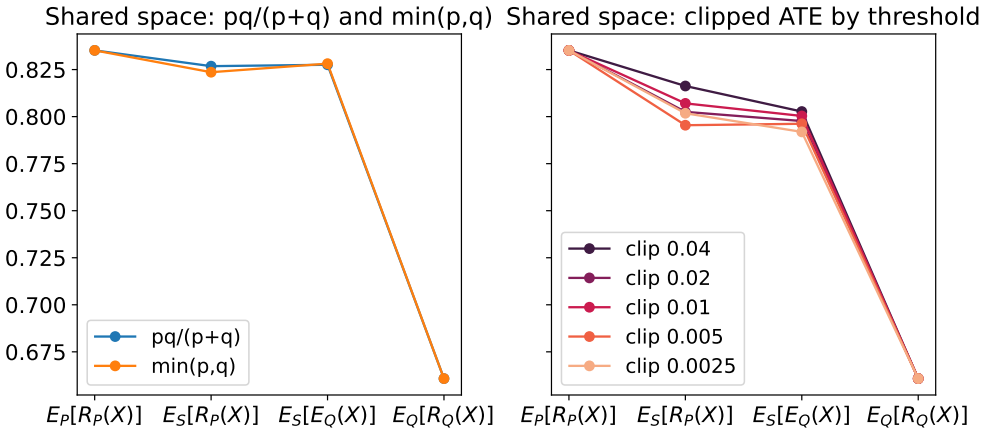
The shared space S_X we use in most of this work has density

$$s_X(x) \propto \frac{p_X(x)q_X(x)}{p_X(x) + q_X(x)}$$

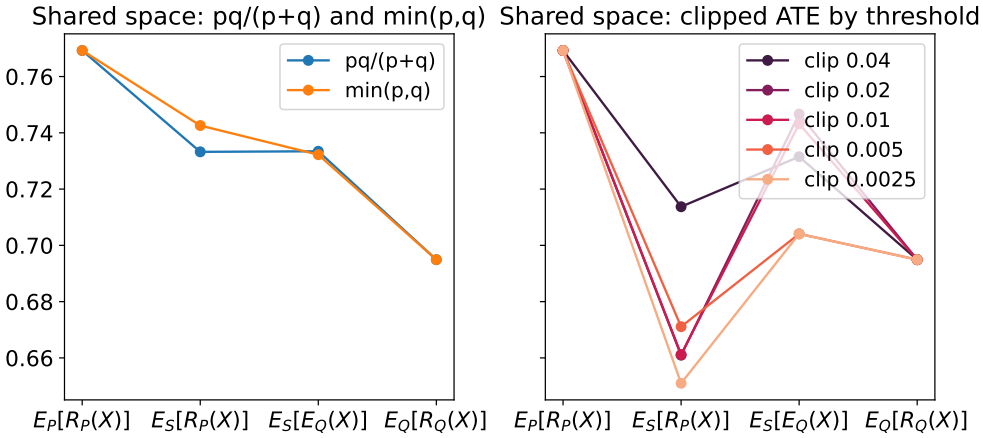
as first defined in Equation (2.2), so that it has support only where both p_X and q_X has support, and also higher (resp. lower) density when $p_X(x)$ and $q_X(x)$ have higher (resp. lower) density. As mentioned in Equation (2.3), there are other definitions of S_X that have similar properties. We repeat the experiments Section 4.1 with these alternative definitions of S_X and show that the decompositions are generally not too sensitive to the specific choice of S_X in Figure 7.



(a) $Y|X$ shift: original model trained on West Virginia and evaluated on Maryland



(b) X shift: model trained on only age $\leq$25 and evaluated on general population



(c) X shift: model trained on over-sampling age $\leq$ 25 and evaluated on general population

Figure 7. Comparison of different shared spaces. On the Y axis is accuracy; note that all results hold if we replace $\ell(\cdot)$ with accuracy. The leftmost plots correspond to Equations (2.2) and (2.3a), while the rightmost plots correspond to Equation (2.3b). Observe that the decompositions tend to be qualitatively similar across different shared spaces, with the exception of the decompositions corresponding to some truncation threshold values for Equation (2.3b). Because of the need to choose a threshold for Equation (2.3b) and the sensitivity of decompositoins to the threshold value, we do not generally recommend it.

G Zero-shot CLIP decompositions from ImageNet to ImageNetV2

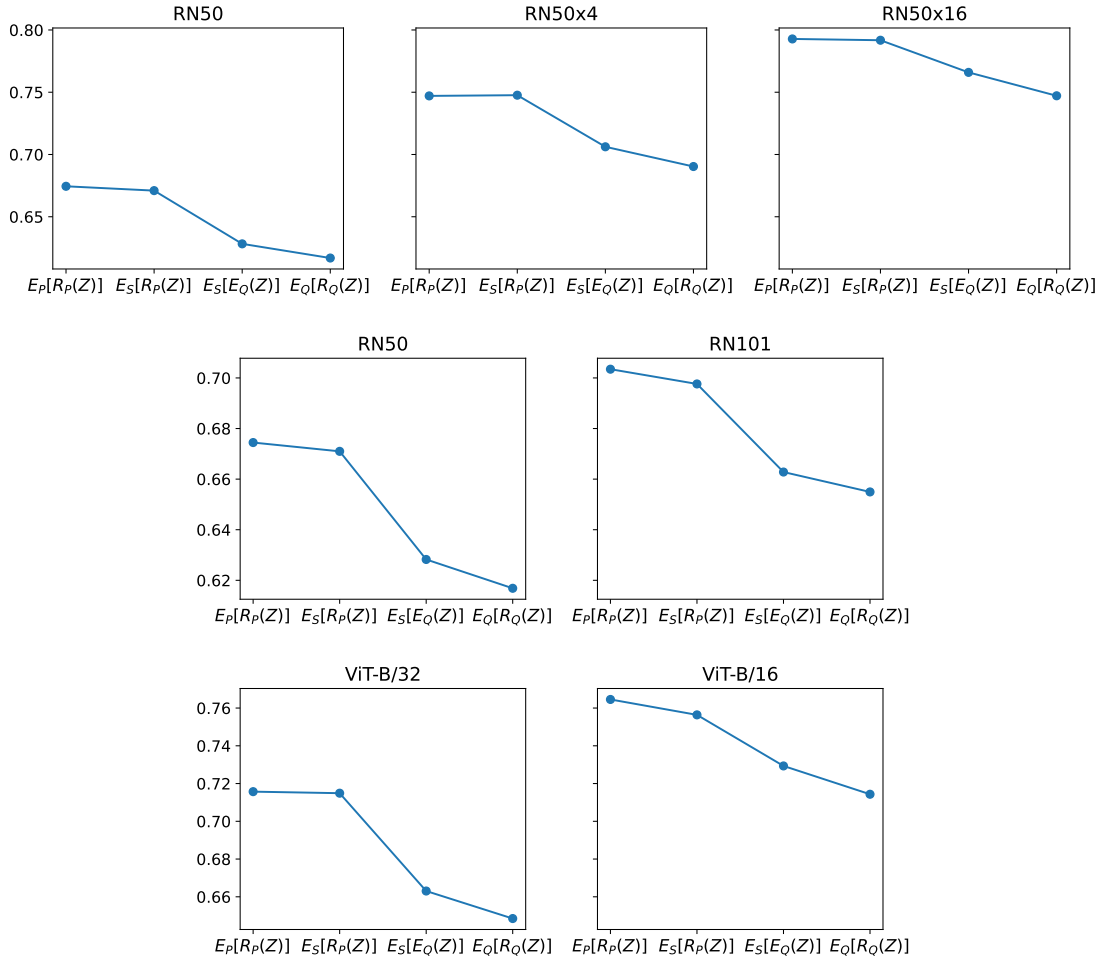


Figure 8. Decompositions for zero-shot CLIP for ImageNet vs ImageNetV2 with labels from [90]. Each row of plots contains a sequence of CLIP models of increasing size, from left to right. These decompositions correspond to the measurements in Figure 5a.