Robustness to Spurious Correlations via Human Annotations

Megha Srivastava [1] **Tatsunori Hashimoto** [1] **Percy Liang** [1]

Abstract

The reliability of machine learning systems critically assumes that the associations between features and labels remain similar between training and test distributions. However, *unmeasured variables*, such as confounders, break this assumption—useful correlations between features and labels at training time can become useless or even harmful at test time. For example, high obesity is generally predictive for heart disease, but this relation may not hold for smokers who generally have lower rates of obesity and higher rates of heart disease. We present a framework for making models robust to spurious correlations by leveraging humans' common sense knowledge of causality. Specifically, we use human annotation to augment each training example with a potential unmeasured variable (i.e. an underweight patient with heart disease may be a smoker), reducing the problem to a covariate shift problem. We then introduce a new distributionally robust optimization objective over unmeasured variables (UV-DRO) to control the worst-case loss over possible test-time shifts. Empirically, we show improvements of 5–10% on a digit recognition task confounded by rotation, and 1.5–5% on the task of analyzing NYPD Police Stops confounded by location.

1. Introduction

The increasing use of machine learning in socioeconomic problems as well as high-stakes decision-making emphasizes the importance of designing models that can perform well over a wide range of users and conditions (Barocas & Selbst, 2016; Blodgett et al., 2016; Hovy & Sgaard, 2015; Tatman, 2017). In some cases, the set of target users and test distributions are known—for example, research in the fair machine learning community has largely been motivated by case studies such as face-recognition systems performing poorly on populations with dark skin color (Buolamwini & Gebru, 2018). However, there exist many more distributional shifts that the designer of a machine learning system may have been unaware of when collecting data, or may be impossible to measure. Existing approaches such as distributional robustness (Ben-Tal et al., 2013; Lam & Zhou, 2015) and domain adaptation (Mansour et al., 2009b; Blitzer et al., 2011; Gong et al., 2013) require either a priori specifying the distribution shifts, or sampling from the target test distributions. How can we ensure that a model performs reliably at test time without explicitly specifying the domain shifts?

Existing research on human-in-the-loop systems and crowdsourcing have shown that humans have a rich understanding of the plausible domain shifts in our world, as well as how these changes affect the prediction task (Talmor et al., 2019; Mostafazadeh et al., 2016). Can we leverage humans' strong prior knowledge to understand the possible distribution shifts for a specific machine learning task? The key idea of our paper is to use human commonsense reasoning as a source of information about potential test-time shifts, and effectively use this information to learn robust models.

To see how human annotations may help, consider the task of creating large-scale diagnostic models for medicine. Although these models are trained on large amounts of data, they almost invariably lack features for important risk factors that were either hidden for legal reasons (e.g. health insurance providers cannot collect genetic information), privacy concerns (e.g. collection of ethnic information (Ploeg et al., 2004)), or simply unobserved (e.g. drug use). For example, a diagnostic model for heart disease trained on the general population may learn to predict heart disease based upon obesity. However, when used in a drug rehabilitation facility with former smokers, this model may perform poorly, as smokers are often underweight and have high heart disease risk (Jarvik, 1991). In this case, smoking is an unmeasured confounder which degrades the model's robustness. A human expert could help with such confounded shifts by annotating the data and identifying that examples with low obesity but significant heart problems may be due to smoking. We would then be able to train our model to be robust to distribution shifts over these unmeasured variables (i.e. if our test set consists primarily of smokers).

[1]Computer Science Department, Stanford University. Correspondence to: Megha Srivastava <megha@cs.stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

The setting we consider is a prediction task where given features (x) and labels (y) from a training distribution, our goal is to perform well at predicting y given x on an a priori unknown test distribution. To make this task possible, we hypothesize that the distribution shift occurs solely over the features x and a set of unmeasured variables c—which can encode obvious confounding factors such as the location of the collected data, as well as more complex factors such as time or demographic information of individuals. Although this assumption reduces the problem to the well-studied covariate shift case, we cannot apply any of these algorithms directly as c is unobserved.

The key insight of our work is that if we design our model to only depend on the features x (and not the unmeasured variables c), we *do not* need to recover the true value of c. Instead, our procedure only requires samples from the conditional distribution $c \mid x, y$ during training. This property allows us to augment the training data with c using crowdsourcing, and leverage human commonsense reasoning to define the potential test-time shifts over unmeasured variables. We will first augment the dataset with approximate $\bar{c}$ by asking humans for natural language descriptions of additional reasons why features x would lead to a label y. Eliciting c in natural language means that we do not have to specify the set of potential unmeasured variables, and allows annotators to easily express a diverse and rich class of c's. We then use these annotations to learn a model from features x to labels y that is robust to distribution shifts over the features and unmeasured variables (x, c).

2. Problem Statement

Formally, consider a prediction problem where we observe features x, and predict a label y. A model θ suffers loss $\ell((x, y), \theta)$, and we train this model using samples $(x, y) \sim p_{\text{train}}$. While standard practice minimizes risk with respect to the training distribution,

$$\mathbb{E}_{p_{\text{train}}}[\ell((x, y); \theta)], \tag{1}$$

this approach can fail when $p_{\text{test}} \neq p_{\text{train}}$, as is common in real-world tasks that involve domain adaptation. For example, the training distribution may be affected by annotation biases (Geva et al., 2019) or underrepresentation of minority groups (Oren et al., 2019; Hashimoto et al., 2018) compared to the test distribution. In this situation we would like the model to perform well over the set of potential test distributions $\mathcal{P}$ by minimizing

$$\mathcal{R}(\theta, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell((x, y); \theta)]. \tag{2}$$

The minimax objective captures many existing settings of interest, such as domain adaptation (where $\mathcal{P}$ is a small-number of target domains (Mansour et al., 2009b)), uniform subgroup guarantees (where $\mathcal{P}$ are minority subgroups of the training distribution (Hashimoto et al., 2018; Duchi & Namkoong, 2018)), and distributionally robust optimization (where $\mathcal{P}$ is a divergence ball centered around the training distribution (Ben-Tal et al., 2013)). We will focus on *uniform subgroup guarantees* which define the set of potential test distributions as subpopulations with size at least α^*,

$$\mathcal{P}^{\alpha}_{x,y} := \{\mathcal{Q}_0 : p_{\text{train}}(x, y) = \alpha^* \mathcal{Q}_0(x, y) + (1 - \alpha^*)\mathcal{Q}_1(x, y) \text{ for some } \mathcal{Q}_1 \text{ with } \alpha^* > \alpha\}. \tag{3}$$

Prior work has shown that such shifts over groups can be used to capture a wide range of test time distributions including label shifts (Hu et al., 2018), topics within a corpus (Oren et al., 2019) , and demographic groups (Hashimoto et al., 2018). However, the set $\mathcal{P}^{\alpha}_{x,y}$ includes *all* possible subpopulations which can be too pessimistic since this allows the conditional distribution $p_{\text{test}}(y \mid x)$ to change arbitrarily and adversarially subject to the α-overlap constraint. For example, minimizing $\mathcal{R}(\theta, \mathcal{P}^{\alpha}_{x,y})$ with the zero-one loss results in a degenerate worst-case group that simply groups all the misclassified examples (up to a α fraction) into an adversarial worst-case group (Hu et al., 2018). This drawback will lead us to consider restricted forms of the subpopulation guarantee in $\mathcal{P}^{\alpha}_{x,y}$.

A common approach for avoiding such degeneracy is to make a *covariate shift* assumption (Shimodaira, 2000; Quiñonero-Candela et al., 2009) which asserts that

$$p_{\text{train}}(y \mid x) = p_{\text{test}}(y \mid x). \tag{4}$$

This resolves the earlier issues by restricting the subpopulation shift (3) to the covariate x. We will define this uncertainty set as $\mathcal{P}^{\alpha}_{x}$ analogously to $\mathcal{P}^{\alpha}_{x,y}$. One particularly appealing property of covariate shift is that the Bayes-optimal classifier on p_{train} will be Bayes-optimal on any $p_{\text{test}} \in \mathcal{P}^{\alpha}_{x}$, making it possible to simultaneously perform well on both the average *and* worst case. Unfortunately, this assumption is usually violated, as many distributional shifts involve unmeasured variables c and even if $y \mid x, c$ remains fixed across train and test, the same may not hold for $y \mid x$.

2.1. Covariate shifts over unobserved variables

Recall our earlier example of a model trained on the general population to predict heart disease (y) from features such as obesity (x) and tested on recent smokers in a rehabilitation center. The covariate shift assumption does not hold, as the conditional distribution $y \mid x$ differs substantially from training to test. This example of *omitted variable bias* arises whenever we fail to account for confounders, mediators, and effect modifiers (Angrist & Pischke, 2009; VanderWeele, 2015).

Using a general purpose uncertainty set such as $\mathcal{P}^{\alpha}_{x,y}$ to capture these types of shifts would also allow for nearly

arbitrary shifts in the predictive distribution and would prevent us from making any predictions at all. However, the situation changes drastically if we observed whether individuals were smokers. If smoking is the only unmeasured variable which changes between train and test, $y \mid x, c$ remains fixed and this allows us to make predictions based only on correlations between y and x which remain reliable under distributional shifts on c.

Making this intuition precise, we will require that c make the train and test distributions differ by a covariate shift in (x, c),

$$p_{\text{train}}(y \mid x, c) = p_{\text{test}}(y \mid x, c). \tag{5}$$

This criterion (known as exogeneity in Pearl (2000)) defines our desired set c; however, this definition neither guarantees the existence of c nor allows us to find a valid c for a given generative mechanism. We will now show how to identify valid unmeasured variables c under a given graphical model.

2.2. Conditions given a graphical model

Suppose that the features and labels x, y are associated with a probabilistic graphical model that captures the generative process of x and y. Now define a selector variable z which determines whether a sample is included in the train or test distribution, with edges in the graph consistent with the covariate shifts (i.e. $p_{\text{train}}(x, c, y) = p(x, c, y, z = 0)$ and $p_{\text{test}}(x, c, y) = p(x, c, y, z = 1)$).

We now state a necessary and sufficient condition for c to fulfill (5), which is that c consists of all variables such that y is d-separated from z by (x, c),

Proposition 1. *A set of variables c in a causal graph fulfills the exogeneity condition (5) whenever z and y are d-separated by (x, c).*

This follows from the definitions of exogeneity and d-separation, which imply $p(y \mid x, c, z) = p(y \mid x, c)$. When the graph has a causal interpretation, and z has no children[1] (Figure 1), z acts as a treatment indicator and c is the set of confounders for the effect of z on y conditional x. This follows from the fact that d-separation and blocking backdoor paths are equivalent if z has no children (VanderWeele & Shpitser, 2013).

2.3. Sampling Unmeasured Variables

Our main challenge is that even if we know unmeasured variables c exist, we cannot measure their value on our training data and use c to constrain the test time conditionals $p_{\text{test}}(y \mid x)$. However, if we could sample from the distribution $p(c \mid x, y)$, and combine this with (x, y) samples in the training data, we can obtain samples from the full joint distribution of (c, x, y). This distribution would, in turn, allow us to understand the set of potential p_{test} that can occur when we shift the marginal distribution of (x, c).

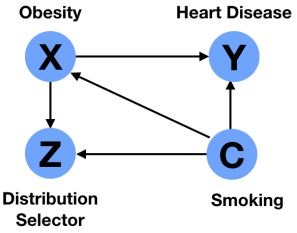


Figure 1. Smoking (c) d-separates whether an example is in the test or train set (z) from the heart disease label (y) conditioned on obesity (x). In this example z has no children since we select rather than generate examples, and assuming this is the true causal graph, c is a confounder for the effect of z on y.

This key point will allow us to reduce the domain adaptation problem over (x, y) to a covariate shift problem over features (x, c). Robustness under covariate shift is still challenging, but we can now apply existing techniques from the covariate shift literature, such as likelihood re-weighting (Shimodaira, 2000) or distributionally robust optimization (DRO) (Duchi et al., 2019).

The main insight of this paper is that we can approximate this conditional distribution $(c \mid x, y)$ through *human annotation*: we ask human annotators for "additional reasons" why feature x would lead to y, and record the natural language explanations as approximate unmeasured variables $\bar{c}$. A key property of our human annotation procedure is that it is only used to augment *training* data. We cannot sample from this same conditional distribution at *test time*, since we do not have y, and sampling from $c \mid x$ does not provide any additional information beyond x. Our proposed procedure therefore does not rely on any annotation of unmeasured variables c at test time. We use shifts over the elicited $\bar{c}$ at training time to determine potential shifts in $(x, \bar{c})$, and learn a model over the original features x that is robust to these shifts.

Conceptually, our approach has three parts: we first elicit $\bar{c} \mid x, y$ from human annotators over our training data. We then use $(x, \bar{c}, y)$ to define the potential test-time shifts $p_{\text{test}}(x, \bar{c}) \in \mathcal{P}$. Finally, we learn a model θ that predicts $x \to y$ such that $\ell(x, y; \theta)$ is small over the entirety of $\mathcal{P}$.

3. Estimation and Optimization

We now discuss the challenges of learning a robust model over unmeasured variables c. We first define our estimator in terms of the true unmeasured variable c and later discuss the challenges associated with using elicited $\bar{c}$ in Section 4.

[1] This common situation occurs whenever the data already exists, and the training and test distributions are constructed by sampling and selecting examples from a population.

Given samples from $p(c \mid x, y)$, we can consider the covariate shift problem over (x, c). In principle, our proposal can utilize any covariate shift approach. However, to illustrate concrete performance improvements for our proposal, we will focus on the uniform subpopulation setting with distributionally robust optimization.

Adapting the earlier covariate subpopulation uncertainty set $\mathcal{P}_x^\alpha$ to this case, we obtain uncertainty sets defined over (x, c),

$$\mathcal{P}_{x,c}^\alpha := \{\mathcal{Q}_0 : p_{\text{train}}(x, c) = \alpha^* \mathcal{Q}_0(x, c) + (1 - \alpha^*)\mathcal{Q}_1(x, c) \text{ for some } \mathcal{Q}_1 \text{ with } \alpha^* > \alpha\}. \quad (6)$$

The resulting distributionally robust objective is now

$$\inf_{\theta \in \theta} \sup_{Q_0 \in \mathcal{P}_{x,c}^\alpha} \mathbb{E}_{x,c \sim Q_0}[\mathbb{E}[\ell(\theta; (x, y))|x, c]]. \quad (7)$$

We refer to the distributionally robust optimization problem over this uncertainty set ($\mathcal{R}(\theta, \mathcal{P}_{x,c}^\alpha)$) as distributionally robust optimization over shifts in unmeasured variables, or (UV-DRO). Although there exist many techniques for efficient distributionally robust optimization (Ben-Tal et al., 2013; Namkoong & Duchi, 2016; Duchi & Namkoong, 2018), the UV-DRO objective is challenging to estimate from finite samples, as the outer supremum depends on the *conditional* risk $\mathbb{E}[\ell(\theta; (x, y))|x, c]$ rather than the loss $\ell(\theta; (x, y))$.

Finite Sample Estimation Having defined the UV-DRO objective in terms of the population expectations, we now turn to the question of estimating this objective from finite samples. As we have mentioned earlier, the empirical plug-in estimator fails to provide tight bounds for UV-DRO, and a Jensen's inequality argument shows that a naive plug-in UV-DRO estimator ignores the covariate shift structure, resulting in an estimator that is equivalent to the worst-case subpopulation objective over $\mathcal{P}_{x,y}^\alpha$.

One straightforward way to sidestep this challenge is to make smoothness assumptions on $\mathbb{E}[\ell(\theta; (x, y)) \mid x, c]$. Let the L_2-normalized conditional risk $\frac{(\mathbb{E}[\ell(\theta;(x,y))|x,c]-\eta)_+}{\|(\mathbb{E}[\ell(\theta;(x,y))|x,c]-\eta)_+]\|_2}$ be L-Lipschitz,[2] and $\mathcal{H}_L$ be the set of L-Lipschitz positive functions with L_2 norm less than one. Standard variational arguments for distributionally robust optimization in Duchi et al. (2019) give the following variational upper bound:

$$\begin{aligned} R(\theta, \mathcal{P}_{x,c}^\alpha) &= \inf_\eta \frac{1}{\alpha}\mathbb{E}[(\mathbb{E}[\ell(\theta; (x, y))|x, c] - \eta)_+] + \eta \\ &\le \inf_\eta \sup_{h \in \mathcal{H}_L} \frac{1}{\alpha}\mathbb{E}[h(x, c)\,(\mathbb{E}[\ell(\theta; (x, y))|x, c] - \eta)] + \eta \\ &=: R_L(\theta) \end{aligned}$$

where the expectations are taken with respect to p_{train}. The first step follows from standard convex duality for distributional robustness, while the second follows from the variational form of the L_2 norm. The dual form of R_L has a simple empirical plug-in estimator,

$$\inf_{B, \eta \ge 0} \frac{1}{\alpha}\left(\frac{1}{n}\sum_{i=1}^n \left(\ell(\theta; (x_i, y_i)) - \sum_{j=1}^n (B_{ij} - B_{ji}) - \eta\right)_+^2\right)^{1/2} + \frac{L}{n}\sum_{i,j=1}^n \left(\|x_i - x_j\| + \|c_i - c_j\|\right) B_{ij} + \eta. \quad (8)$$

For detailed derivation, see the supplement. This is a special case of the family of L_p norm variational DRO estimators proposed and studied by Duchi et al. (2019) and is known to converge to its population counterpart at rate $O(n^{-1/d})$.

This estimator intuitively captures both the smoothness assumption and worst-case structure of our objective. The dual variable η serves as a cutoff: all losses ℓ below η within the sum are set to zero, forcing the model to focus on the worst losses incurred by the model. The dual variable B is a transport matrix, where the entry B_{ij} transports loss from example i to j in exchange for a cost of $L(\|x_i - x_j\| + \|c_i - c_j\|)$. This smoothing ensures that the model focuses its attention on neighborhoods of the input (x, c) that systematically have high losses.

4. Approximation with Crowdsourcing

Effect of Approximating Unmeasured Variables Minimizing the UV-DRO objective (8) with c provides a model which is robust to test time shifts that potentially change $y \mid x$. Unfortunately, we do not have access to the *unmeasured* variables c, and instead only observe noisy and approximate samples $\overline{c} \mid x, y$ (e.g. natural language explanation from human crowdworkers, equipped with a semantic similarity metric).

We now characterize the conditions under which a model estimated using approximate unmeasured variables ($\overline{R}_L$) performs well on the true risk (R_L) under the unmeasured variable c. A major challenge in comparing approximate and true unmeasured variables is that $\overline{c} \in \overline{\mathcal{C}}$ and $c \in \mathcal{C}$ are unlikely to even exist in the same metric space.

We overcome this difficulty by characterizing the risk in terms of an alignment. If there exists smooth functions f and g which align the space of approximate unmeasured

[2] Incorporating smoothness acknowledges the fact that real world domain adaptation tasks are not arbitrary, and similar examples suffer similar conditional risk. This prior knowledge can help reduce the effective dimensionality of our inputs $\overline{c}$.

variables $\overline{\mathcal{C}}$ with the space of true unmeasured variables $\mathcal{C}$, then the optimal model under the approximate $\overline{c}$ performs well on the true risk R_L.

Proposition 2. *Let $f : \mathcal{C} \to \overline{\mathcal{C}}$ and $g : \overline{\mathcal{C}} \to \mathcal{C}$ be any K_f and K_g Lipschitz-continuous functions. For positive losses bounded above by M, the minimizer for the approximate risk ($\overline{R}_L$) given by $\overline{\theta}^* := \arg\min_\theta \overline{R}_L(\theta)$ fulfills*

$$R_L(\overline{\theta}^*) - \inf_\theta R_L(\theta)$$
$$\leq \inf_\theta R_L(\theta)\,(K_f K_g - 1) + \frac{LM}{\alpha}(A_f K_g + A_g)$$

where

$$A_f = \mathbb{E}W_1(\overline{c}|xy, f(c)|xy) \;\; \textit{and} \;\; A_g = \mathbb{E}W_1(c|xy, g(\overline{c})|xy)$$

and $W_1(\overline{c}, f(c))$ is the Wasserstein distance between the distribution of $\overline{c}$ and the pushforward measure of c under f.

See the supplement for proofs and additional bounds.

The $K_f K_g$ distortion term captures the fact that a Lipschitz continuity assumption under $\overline{c}$ differs from one under c. The additive terms A_f, A_g captures the distributional differences between $c|xy$ and $\overline{c}|xy$. Note that if $\overline{c}$ and c share the same metric space, the relative error term $(K_f K_g - 1)$ is zero, and the model approximation quality depends on the average Wasserstein distance between $c \mid xy$ and $\overline{c} \mid xy$.

Crowdsourcing for Elicitation To better understand the approximation bound, consider the example of a digit recognition task confounded by rotation, where we are asked to classify images (x) of digits which have undergone an unobserved rotation (c). The unmeasured variable is a real-valued angle, but $\overline{c}$ is a natural language annotation with the metric defined by a semantic similarity metric over sentences (Figure 4).

In this example, the distortion term $K_f K_g$ captures whether the distance between natural language description of two images whose rotations differ by d degrees is close to d. The Wasserstein term captures the fact that natural language descriptions can be noisy, and we can sometimes get annotations that do not correspond to any c.

We attempt to mitigate the effect of two terms through the crowdsourcing design:

1. Each user annotated many examples to reduce phrasing variation (a person using "turn" instead of "rotate" likely always uses "turn" for applicable examples).
2. We selected a vector sentence representation (Sent2Vec) whose distances have been shown to correlate with semantic similarity.
3. We collected and averaged multiple annotations per training example to reduce crowdsourcing noise.

This data collection procedure allows us to capture the well-studied ability for humans to identify unobserved causes (Schulz et al., 2008; Saxe et al., 2007) and allows us to obtain more robust models under several types of bias from unmeasured variables.

5. Experimental Results

We now demonstrate that distributional robustness over unmeasured variables (UV-DRO) results in more robust models that rely less upon spurious correlations. Across all experiments, we show that UV-DRO achieves more robust models than baselines as well as other DRO objectives, including that of Duchi et al. (2019) ("Covariate Shift DRO") and Hashimoto et al. (2018) ("Baseline DRO").

Experimental Procedures. Both the linear regression (Section 5.1) and logistic regression models (Sections 5.2 and 5.3) were optimized using batch gradient descent with AdaGrad. We tuned hyperparameters such as the learning rate, regularization, and DRO parameters using a held-out validation set, which we describe in the supplement.

Human annotations were performed by crowdworkers on Amazon Mechanical Turk, and the specific prompts for each task are included in the relevant sections. In both tasks, we define the distance between two annotations c by embedding each sentence into vector space with the FastText Sent2Vec library, and measuring the average cosine distance between the two vectors across two replicate annotations.

5.1. Simulated Medical Diagnosis Task

We begin with a simple simulated medical diagnosis dataset. One source of bias in medical datasets is that patients can sometimes lie about symptoms to their doctors. It has been well-studied that adolescents have a substantially higher chance of lying to physicians about sexual activity (Zhao et al., 2016), which complicates medical diagnosis and the ability to prescribe teratogenic drugs such as the acne drug Accutane (Honein et al., 2001). In this example, age is an effect modifier which for simplicity we assume is the only unmeasured variable (VanderWeele, 2012).

We consider a simplified scenario of using the patient's self-reported pregnancy symptoms x_1 and clinical measurements x_2 to predict pregnancy y via a least-squares linear regression model ($y = \beta^\top x + b$). We demonstrate that a model trained with empirical risk minimization (ERM) learns the unreliable correlation between self-reporting and pregnancy, while UV-DRO using an imputed age learns to use noisier but more reliable clinical measurements.

We will define the data generating distribution for our observations as $x_1 = cy$ where $c \sim 1 - 2\,\text{Bernoulli}(q)$ is the patient's truthfulness, and the clinical measure-

ments follow $x_2 = y + \epsilon$ where $\epsilon \sim N(0, 4)$ is a measurement noise term. We evaluate the models on a series of training distributions with a mix of adults and adolescents where the probability of lying ranges over $q_{train} = \{.05, .1, .2, .3, .4, .5, .6, .7, .8\}$. At test time, our model is applied to adolescents who lie with probability $q_{test} = 0.8$. These datasets fulfill the subpopulation condition (i.e. $\mathcal{P}^{\alpha}_{x,c}$) where the train-test overlap varies over $\alpha^* = \{.0625, .125, .250, .375, .5, .625, .750, .875, 1.0\}$.

From our data generating distribution, we can see that c is an unmeasured variable which affects the correlation between y and x_1. During training time, where the patient set largely consists of adults that are less likely to lie, c is often 1, and a model optimized on this data will predict y using primarily x_1, as shown by the low relative weight of x_2 in Figure 2. However, at test time when there are many adolescent patients who have high likelihood of lying, the correlation between x_1 and y is reversed, making this model perform poorly on the test set (Figure 3).

On the other hand, if we apply UV-DRO with c sampled according to the true conditional distribution $c \mid x_1, x_2, y$, then our loss will account for the fact that the correlation between x_1 and y may flip at test time, and our learned model uses x_2 (Figure 2)—which reliably measures y. This results in substantial gains in test performance that are stable across a wide range of α^*s (Figure 3). Finally, we observe that conditioning on y is critical when generating c, and using $c \mid x$ instead, as would be the case if we sampled c at test time, fails to improve robustness (Figure 3).

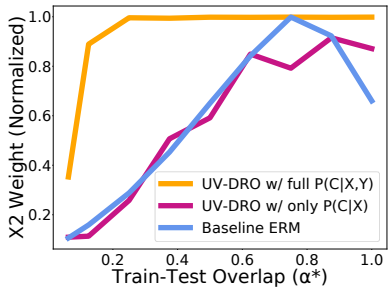


Figure 2. UV-DRO consistently places higher relative weight on more reliable feature x_2 over x_1, unlike both ERM and a baseline UV-DRO with c drawn without access to label y.

5.2. Digit Classification Under Transformations

We evaluate the efficacy of UV-DRO on synthetic domain shifts on the MNIST digit classification task. Specifically, we apply random rotations or occlusions to the images and treat the identity of these transformations as an unmeasured variable. We show that if the distribution of such unmeasured variables shifts from training to test sets, classification accuracy for a simple logistic regression model degrades rapidly for both ERM and existing DRO approaches, but this performance loss is mitigated when using UV-DRO.

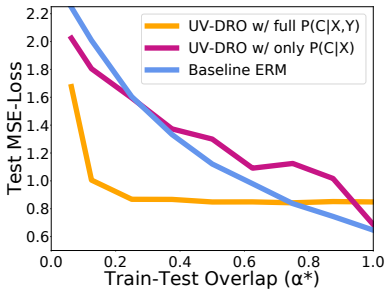


Figure 3. UV-DRO achieves lower loss than both ERM and a baseline UV-DRO with c drawn without access to label y.

Examples of these image transformations, as well as crowdsourced annotations, are shown in Figure 4. We can see that digits such as rotated 6s and 9s can become difficult or impossible to distinguish without knowledge of the rotation angle, and our logistic regression model's performance rapidly degrades as the fraction of rotated digits changes between train and test sets.

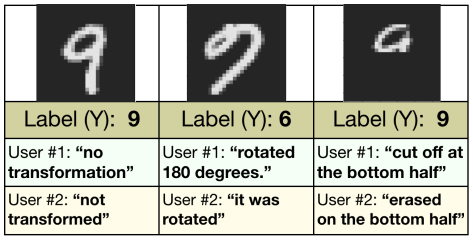
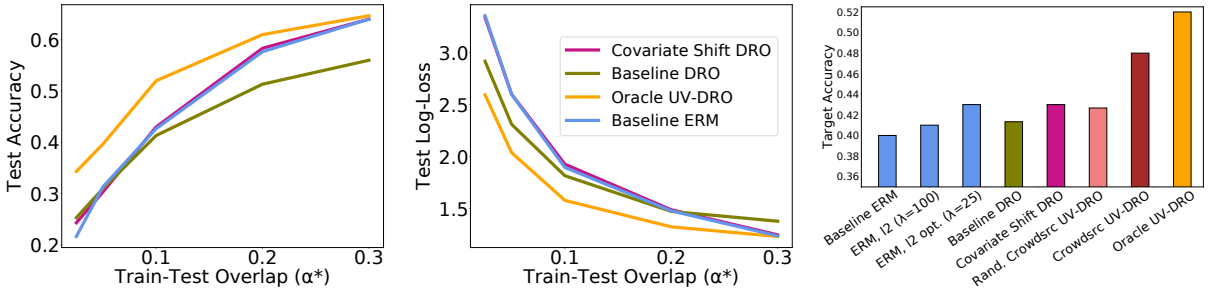


Figure 4. Examples from our confounded training dataset as well as the user-provided annotations from our crowdsourcing task. Human crowdworkers are able to directly recover the unmeasured variables, leading to nearly oracle level performance for UV-DRO.

Estimation with an oracle In our first experiment, we consider the oracle setting, where the unmeasured variable c is a rotation angle and we obtain samples from the joint distribution (x, y, c) to use with UV-DRO for predicting y given x. The training distribution consists of images of which a proportion $\alpha^* = \{0.05, 0.1, 0.2, 0.4, 0.6\}$ are transformed, while the test set consists only of images that are rotated by 180 degrees.

Unsurprisingly, we find low performance for baselines which do not use c at small values of α^* (Figure 5a). This includes ERM (with and without L_2 regularization), DRO over $\mathcal{P}^{\alpha}_{x,y}$ (baseline DRO), and DRO over just the features $\mathcal{P}^{\alpha}_{x}$ (covariate shift DRO). UV-DRO substantially improves performance over these baselines, with a 15% *absolute* accuracy gain for the most extreme shift of $\alpha^* = 0.05$, as well as substantial accuracy gains that persist until $\alpha^* = 0.6$ (Figure 5a). The same trend holds for the log-losses incurred by each model (Figure 5b).



(a) Accuracy vs. minority group size (b) Log-Loss vs. minority group size (c) Accuracy with crowdsourced data

Figure 5. For the MNIST Digit Classification task, our UV-DRO approach using oracle unmeasured variable values results in substantial improvements in both accuracy and loss under large train test shifts (α^*). Using crowdsourced annotations with our UV-DRO approach successfully provides an accuracy gain (48% UV-DRO vs. 43% ERM) over the Baseline ERM models, as well as other DRO approaches.

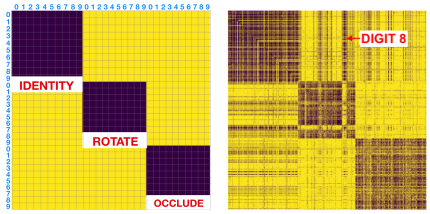


Figure 6. Cosine-distances between natural language annotations (right) strongly correlates with ground truth (left). The distance matrix is ordered by the transformation type and then digit. Note the confusion between "rotation" and "identity" transform for digit 8, which is invariant to 180° rotations.

Estimation with crowdsourced unmeasured variables We next demonstrate that UV-DRO performs well even with crowdsourcing the unmeasured variables, and show this performance approaches that of the oracle. Specifically, we consider a training distribution where images are either rotated with probability 0.1, occluded with probability 0.1, or not manipulated. The test distribution is the same as before, with all images rotated.

In order to obtain samples from $p(c|x, y)$, we performed an Amazon Mechanical Turk task where crowdworkers were shown (x, y) pairs of confounded images (x) and their label (y) from our training dataset of 4000 images. Each user was prompted to answer a free-text question, "What transformation do you think happened to the image?". Importantly, we *did not* inform the users what types of unmeasured variables are present or possible in the dataset. We processed these natural language descriptions into a distance metric suitable for UV-DRO using our embedding procedure described earlier. We find that the resulting distances closely match the true unmeasured variable structure (Figure 6).

Training the UV-DRO model on this dataset, Figure 5c shows substantial accuracy improvement from crowdsourced unmeasured variables (48%) compared to the ERM (43%) and two DRO (41–43%) baselines. Surprisingly, we also observe that crowdsourcing unmeasured variables results in only a 4% accuracy drop relative to using oracle c's (52%), showing that we substantially close the gap to the optimal robust model. Finally, a randomly shuffled permutation of the annotation distances causes a significant drop in accuracy (42%), showing that it is the crowdsourced information—rather than loss or hyperparameter changes—that results in performance gains. Further analysis on the effect of crowdsourcing quality is in the supplement.

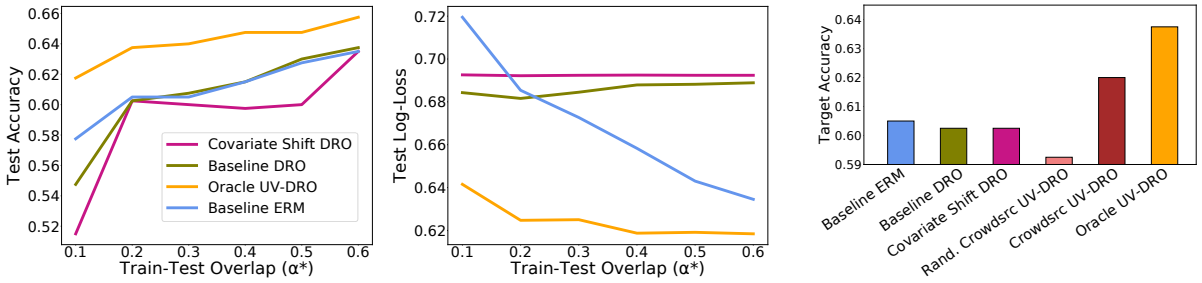
5.3. Analyzing Policing Under Location Shifts

Having demonstrated gains on the semi-synthetic MNIST task, we evaluate UV-DRO on a more complex real-world distribution shift. Stop-and-frisk is a controversial program of temporarily stopping, questioning, or searching civilians by the police, and has been well-studied for amplifying racial biases. We consider the task of trying to detect false positives—or police stops that do not result in arrests—by training classifiers on data from police stops spanning 2003-2014 in New York City (NYCLU, 2019).

For our observed features, we consider a set of 27 possible observations reported by police officers as reasons for stops, such as "furtive movements" and "outline of weapon". Examples from this dataset, as well as crowdsourced annotations, can be seen in Figure 8. Previous work on this dataset has shown that racial minorities are stopped more frequently for less serious observations (Goel et al., 2016; Gelmand et al., 2007).

Our unmeasured variable in this setting is the location of the police stop. We consider training data where the majority of police stops are from Manhattan, and measure the model performance on stops in Brooklyn. This is a natural form of domain shift, where our goal is for a model to make predictions that are reliable regardless of location.

Estimation with an oracle We first consider UV-DRO with an oracle, where the distribution of unmeasured variables matches the ground truth. The training distributions



(a) Accuracy vs. minority group size (b) Log-Loss vs. minority group size (c) Accuracy on crowdsourced data

Figure 7. For the police stop analysis, our UV-DRO approach with oracle location variables provides consistent improvements in accuracy and loss. Using crowdsourced annotations with UV-DRO also improves accuracy (62% UV-DRO vs. 60.5% ERM) over the Baseline ERM model and other DRO approaches. Only the unregularized ERM model is shown, as we found $\lambda = 0$ to be optimal.

*A **white man** was stopped by police, and was then arrested. He:* **1. was acting as a lookout,** **2. was associating with criminals** **3. was seen late at night,** **4. appeared to be involved in drug activity,** and **5. was in a high crime area.**	*A **black man** was stopped by police, but was NOT arrested. He:* **1. was reported by another person to be suspicious,** and **2. fit the crime description.**
Label (Y): **Arrested**	Label (Y): **Not Arrested**
User #1: **"he was involved with drug dealing."**	User #1: **"because the person reporting him was racist"**
User #2: **"due to he was involved with drug activities."**	User #2: **"because someone called the police on him, likely because of his race."**

Figure 8. Examples from the stop-and-frisk task identify perceived reliable features (left) as well as potential confounding (right).

consists of police stops which occurred in Manhattan, except for a proportion $\alpha^* = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ which occurred in Brooklyn. The test distribution consists solely of police stops in Brooklyn.

Similar to our previous experiments, we see that this task is very challenging under large train-test shifts, with most models performing near 50–60% accuracy. For Logistic Regression, using UV-DRO with oracle provides substantial gains on both accuracy and loss over a wide range of α^* (Figure 7). For accuracy, UV-DRO provides just under a 4% accuracy gain across most values of α^*, while DRO baselines surprisingly do *worse* than naively using ERM.

Estimation with crowdsourced unmeasured variables Unlike the previous digit classification task, police stops and observations can be affected by an incredibly large set of confounders. While we hypothesize that it is unlikely for crowdworkers to actually predict location as an unmeasured confounder, Proposition 2 suggests that this is not necessary. We show that crowdworkers capture a variety of social and demographic factors, which are sufficiently indicative location to provide robustness gains.

For this task, we fix the proportion of training examples from Brooklyn at $\alpha^* = 0.2$. We present Amazon Mechanical Turk users (x, y) pairs of police stop descriptions—including the full feature set of race, gender, and officer observations (x)—and the label of whether the individual was arrested or not (y). Each user is then asked, "What factors do you think led to the individual being stopped and [arrested/not arrested]?", for which they provide free-text responses. We use the same procedure as the MNIST experiment to process these responses into a distance matrix.

Figure 8 includes example annotations we elicited from users. Many free-text annotations showed strong ability to recover social factors (i.e. racism), as well as filter features to identify the relevant factors for an arrest decision.

Training our UV-DRO model, we find that crowdsourced UV-DRO gives a 1.5% accuracy gain, which is once again nearly half of the gap between the ERM baseline and oracle DRO (Figure 7). Similar to our previous experiment, we find that existing DRO baselines, as well as shuffling our crowdsourcing data, result in no gains over ERM.

Finally, to understand how the crowdsourced annotations capture unmeasured variables, we trained a logistic regression model to predict the police stop location from (1) only observed features (61.3% test accuracy), (2) observed features and annotation unigrams (64.8%), and (3) observed features and arrest labels (65.9%), compared with a majority class baseline of 51%. These results confirm that the annotations indeed help provide more information about the location than the observed features can capture. Further exploratory analysis on predicting the location from annotation unigrams results in assigned weights that are consistent in showing that race, police judgement, and individual circumstances were more predictive of Brooklyn while unigrams associated with violent crime were more predictive of Manhattan (See supplement). This suggests that UV-DRO can not only be used to improve model robustness, but also has the potential to improve interpretability by highlighting unmeasured variables that may result in spurious correlations.

6. Related Works and Discussion

Although our work draws on ideas from domain adaptation (Mansour et al., 2009a;b; Ben-David et al., 2006) causal invariance (Peters et al., 2016; Meinshausen & Bühlmann, 2015; Rothenhäusler et al., 2018), and robust optimization (Ben-Tal et al., 2013; Duchi & Namkoong, 2018; Bertsimas et al., 2018; Lam & Zhou, 2015), few prior works seek to elicit information on the possible shifts in unmeasured variables. For example, Heinze-Deml & Meinshausen (2017) improve robustness under unobserved style shifts in images by relying on multiple images which vary only by style. Similarly, Landeiro & Culotta (2016) develop a back-door adjustment which controls for confounding in text classification tasks when the confounder is known. Recent work by Kaushik et al. (2019) use crowdsourcing to revise document text given a specific counterfactual label. This approach is orthogonal to our work, as it does not attempt to address shifts in unmeasured variables and is intended to improve models under observed covariate shifts.

The crowdsourcing aspect of our work builds on existing work on eliciting human commonsense understanding and counterfactual reasoning. Roemmele et al. (2011) showed that humans achieve high performance on commonsense causal reasoning tasks, while Sap et al. (2019) has used crowdsourcing to build an "if-then" commonsense reasoning dataset. These works support our results which show crowdsourcing can successfully capture how humans reason about unmeasured variables.

Discussion We have demonstrated that domain adaptation problems with unmeasured variables can be recast as covariate shift problems once we obtain samples from $c \mid x, y$ at train time. Notably, rather than expanding the training dataset, our work accounts for variables that may be inaccessible in an already existing dataset. Our UV-DRO approach and experiments show that crowdsourcing can be an effective way of eliciting these unmeasured variables, and we often obtain results close to an oracle model which uses the true c distribution. This work is the first step towards explicitly incorporating human knowledge of potential unmeasured variables via natural language annotations, and opens the possibility of methods that make use of counterfactual explanations from domain experts to learn reliable models in high-stakes situations.

Reproducibility We provide all source code, data, and experiments as part of a worksheet on the CodaLab platform: `https://bit.ly/uvdro-codalab`.

Acknowledgements We thank all anonymous reviewers for their valuable feedback. Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. MS was additionally supported by the NSF Graduate Research Fellowship Program under Grant No. DGE-1656518.

References

Angrist, J. D. and Pischke, J. *Mostly Harmless Econometrics: An Empiricist's Companian*. Princeton University Press, 2009.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *104 California Law Review*, 3:671–732, 2016.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 137–144, 2006.

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

Bertsimas, D., Gupta, V., and Kallus, N. Data-driven robust optimization. *Mathematical Programming Series A*, 167, 2018.

Blitzer, J., Kakade, S., and Foster, D. P. Domain adaptation with coupled subspaces. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 173–181, 2011.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of African-American English. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1119–1130, 2016.

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

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses against mixture covariate shifts. `https://cs.stanford.edu/~thashim/assets/publications/condrisk.pdf`, 2019.

Gelmand, A., Fagan, J., and Kiss, A. An analysis of the new york city police departments stop-and-frisk policy in the context of claims of racial bias. In *Journal of the American Statistical Association*, 2007.

Geva, M., Goldberg, Y., and Berant, J. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Goel, S., Rao, J. M., and Shroff, R. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. In *The Annals of Applied Statistics*, 2016.

Gong, B., Grauman, K., and Sha, F. Reshaping visual datasets for domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

Heinze-Deml, C. and Meinshausen, N. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

Honein, M., Paulozzi, L., and Erickson, J. Continued occurrence of accutane exposed pregnancies. In *Teratology*, 2001.

Hovy, D. and Sgaard, A. Tagging performance correlates with age. In *Association for Computational Linguistics (ACL)*, pp. 483–488, 2015.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, 2018.

Jarvik, M. E. Beneficial effects of nicotine. In *British Journal of Addiction*, 1991.

Kaushik, D., Hovy, E., and Lipton, Z. C. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Lam, H. and Zhou, E. Quantifying input uncertainty in stochastic optimization. In *2015 Winter Simulation Conference*, 2015.

Landeiro, V. and Culotta, A. Robust text classification in the presence of confounding bias. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1041–1048, 2009a.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory (COLT)*, 2009b.

Meinshausen, N. and Bühlmann, P. Maximin effects in inhomogeneous large-scale data. *Annals of Statistics*, 43, 2015.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *North American Association for Computational Linguistics (NAACL)*, 2016.

Namkoong, H. and Duchi, J. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

NYCLU. Stop-and-frisk data. https://www.nyclu.org/en/stop-and-frisk-data, 2019.

Oren, Y., Sagawa, S., Hashimoto, T., and Liang, P. Distributionally robust language modeling. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Pearl, J. *Causality: Models, Reasoning and Inference*, volume 29. Springer, 2000.

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

Ploeg, M. V., Perrin, E., on D. C. of Race, P., and Data, E. *Eliminating Health Disparities: Measurement and Data Needs*. National Academies Press, 2004.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Roemmele, M., Bejan, C. A., and Gordon, A. S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2011.

Rothenhäusler, D., Bühlmann, P., Meinshausen, N., and Peters, J. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

Sap, M., LeBras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., and Choi, Y. Atomic: An atlas of machine commonsense for if-then reasoning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

Saxe, R., Tzelnic, T., and Carey, S. Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. In *Developmental Psychology*, 2007.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., and Jenkins, A. C. Going beyond the evidence: Abstract laws and preschoolers responses to anomalous data. In *Cognition*, 2008.

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

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *North American Association for Computational Linguistics (NAACL)*, 2019.

Tatman, R. Gender and dialect bias in youtubes automatic captions. In *Workshop on Ethics in Natural Langauge Processing*, volume 1, pp. 53–59, 2017.

VanderWeele, T. J. Confounding and effect modification: distribution and measure. In *Epidemiologic Methods*, 2012.

VanderWeele, T. J. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

VanderWeele, T. J. and Shpitser, I. On the definition of a confounder. In *Annals of Statistics*, 2013.

Zhao, J., Lau, M., Vermette, D., Liang, D., and Flores, G. Communication between asian american adolescents and health care providers about sexual activity, sexually transmitted infections, and pregnancy prevention. In *Journal of Adolescent Research*, 2016.