Supplementary Materials

S1. Effect of Crowdsourcing Quality

![Graphs showing the effect of crowdsourcing quality.](image)

Figure S1. Decreasing crowdsourcing quality by randomly shuffling results in a highly correlated decrease in accuracy over both MNIST (left) and stop-and-frisk (right) datasets.

We empirically evaluate the role of crowdsourcing data quality on UV-DRO performance to complement our theoretical bound in Section 4. We previously showed a significant performance gap when we shuffle 100% of the crowdsourced unmeasured variables, causing random associations that impact the crowdsourcing quality. We further investigate this gap by shuffling [0, 2, 5, 10, 20, 50, 75]% of the crowdsourced unmeasured variables, and find a highly correlated accuracy drop for both MNIST ($R^2 = .89$) and stop-and-frisk datasets ($R^2 = .91$), as seen in Figure S1. This demonstrates a linear relationship between crowdsourcing quality and robust performance.

S2. Annotation Unigrams Analysis Table

Table S1. Exploratory analysis on the annotations collected over stop-and-frisk data by training a logistic regression model to predict location from a selection of annotation unigrams.

<table>
<thead>
<tr>
<th>UNIGRAM</th>
<th>BROOKLYN</th>
<th>MANHATTAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISCRIMINATION</td>
<td>-1.22</td>
<td>0.82</td>
</tr>
<tr>
<td>RACIST</td>
<td>-0.29</td>
<td>0.21</td>
</tr>
<tr>
<td>RACIAL</td>
<td>-0.19</td>
<td>0.89</td>
</tr>
<tr>
<td>HOMELESS</td>
<td>-0.84</td>
<td>0.43</td>
</tr>
<tr>
<td>UNRELATED</td>
<td>-1.68</td>
<td>1.03</td>
</tr>
<tr>
<td>CLEARED</td>
<td>-0.98</td>
<td>0.79</td>
</tr>
<tr>
<td>EVIDENCE</td>
<td>-0.12</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table S1. Exploratory analysis on the annotations collected over stop-and-frisk data by training a logistic regression model to predict location from a selection of annotation unigrams.
S3. Derivation of the empirical dual estimator

The arguments given here are a simplification of the class of duality arguments from Duchi et al. (2019). Recall that the inner maximization sup_{h \in \mathcal{H}_L} \mathbb{E}[h(x, c)(\mathbb{E}[\ell(\theta; (x, y)) | x, c] - \eta)] admits a plug-in estimator which can be written as a linear objective with Lipschitz smoothness and L_2 norm constraints,

$$\max_{h \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} h_i(\ell(\theta; (x_i, y_i)) - \eta)$$

s.t. $h_i \geq 0$ for all $i \in [n]$, $\frac{1}{n} \sum_{i=1}^{n} h_i^2 \leq 1$,

$$h_i - h_j \leq L(\|x_i - x_j\| + \|c_i - c_j\|)$$

for all $i, j \in [n]$.

Now taking the dual with $\gamma \in \mathbb{R}^n_+$, $\lambda \geq 0$, and $B \in \mathbb{R}^{n \times n}$, the associated Lagrangian is

$$\mathcal{L}(h, \gamma, \lambda, B) := \frac{1}{n} \sum_{i=1}^{n} h_i(\ell(\theta; (x_i, y_i)) - \eta) + \frac{1}{n} \gamma^\top h + \frac{\lambda}{2} \left( 1 - \frac{1}{n} \sum_{i=1}^{n} h_i^2 \right)$$

$$+ \frac{1}{n} \left( L \text{tr}(B^\top D) - h^\top (B1 - B^\top 1) \right)$$

where $D \in \mathbb{R}^{n \times n}$ is a matrix with entries $D_{ij} = \|x_i - x_j\| + \|c_i - c_j\|$. From strong duality, the primal optimal value (1) is $\inf_{\gamma \in \mathbb{R}^n_+, \lambda \geq 0, B \in \mathbb{R}^{n \times n}} \sup_h \mathcal{L}(h, \gamma, \lambda, B)$.

The first order conditions for the inner supremum give

$$h_i^* := \frac{1}{\lambda} \left( \ell(\theta; (x_i, y_i)) - \eta + \gamma - (B1 - B^\top 1)_i \right).$$

Substituting these values and taking the infimum over $\lambda, \gamma \geq 0$, we obtain

$$\inf_{\lambda \geq 0} \sup_{\gamma \in \mathbb{R}^n_+} \sup_h \mathcal{L}(h, \gamma, \lambda, B) = \left( \frac{1}{n} \sum_{i=1}^{n} \left( \ell(\theta; (x_i, y_i)) - \sum_{j=1}^{n} (B_{ij} - B_{ji}) - \eta \right)^2 \right)^{1/2}$$

$$+ \frac{L}{n} \sum_{i, j=1}^{n} (\|x_i - x_j\| + \|c_i - c_j\|) B_{ij}.$$

Taking the infimum over $B, \eta$ and substituting this expression into the inner supremum of $R_L$ gives the desired estimator.

S4. Distortion Proof

Terminology in this section generally follows that of the main text. We will use $c$ to describe some true set of unmeasured variables, and $\tau$ to describe the elicited set. All notation with overhead lines are defined in this space of elicited unmeasured variables (e.g. $h, \mathcal{H}_L$).

Additionally we will define a forward map from true unmeasured variables to elicited ones, $f : \mathcal{C} \rightarrow \mathcal{C}$ and a reverse map from elicited unmeasured variables to true ones $g : \mathcal{C} \rightarrow \mathcal{C}$.

For convenience, define the following risk functionals for the DRO problem under the true unmeasured variables

$$R_L(\theta) := \inf_{\eta} \sup_{h \in \mathcal{H}_L} \frac{1}{\alpha} \mathbb{E}_{x, y, c} [h(x, c)\ell(x, y) - \eta] + \eta,$$

and under the estimated ones

$$\overline{R}_L(\theta) := \inf_{\eta} \sup_{\overline{h} \in \overline{\mathcal{H}}_L} \frac{1}{\alpha} \mathbb{E}_{x, y, \tau} [\overline{h}(x, \tau)\ell(x, y) - \eta] + \eta.$$  

(2)
We can define the upper bound for the Lipschitz case,

**Proposition.** Let \( f : \mathcal{C} \to \overline{\mathcal{C}} \) define \( \hat{h}(x, c) := \mathcal{H}(x, f(c)) \) such that \( \frac{1}{K_f} \hat{h} \in \mathcal{H}_L \) for all \( \mathcal{H} \in \mathcal{H}_L \). Then,

\[
\mathcal{R}_L(\theta) \leq K_f R_L(\theta) + \frac{LME_{xy}W_1(f(c|xy), \mathcal{C}|xy)}{\alpha}
\]

where \( f(c|xy) \) is the pushforward measure of \( c|xy \) under \( f \).

**Proof.** Let \( \mathcal{H}^\ast \) be the \( \mathcal{H} \in \mathcal{H}_L \) which is the maximizer to Eq (2). For convenience define

\[
\Delta f_y := E_{c|xy}[\hat{h}^*(x, c)] - E_{\mathcal{C}|xy}[\mathcal{H}^*(x, \mathcal{C})]
\]

\[
= E_{\mathcal{C}|xy}[\mathcal{H}^*(x, \mathcal{C})] - E_{\mathcal{C}|xy}[\mathcal{H}^*(x, \mathcal{C})]
\]

The equality follows the change of variables property of pushforward measures. Now rewriting the risk measure in terms of \( \Delta \),

\[
\mathcal{R}_L(\theta) = \inf_{\eta} \frac{1}{\alpha} E_{xy} \left[ \left( E_{c|xy}[\hat{h}^*(x, c)] - \Delta f_y \right) \ell(x, y) - \eta \right] + \eta
\]

\[
\leq \inf_{\eta} \frac{1}{\alpha} E_{xy} \left[ E_{c|xy}[\hat{h}^*(x, c)]\ell(x, y) - \eta \right] + \eta
\]

\[
+ E_{xy}[|\Delta f_y|] M
\]

\[
\leq \inf_{\eta} K_f \sup_{h \in \mathcal{H}_L} \frac{1}{\alpha} E_{xy} \left[ E_{c|xy}[h(x, c)]\ell(x, y) - \eta \right] + \eta
\]

\[
+ E_{xy}[|\Delta f_y|] M
\]

\[
= K_f R_L(\theta) + \frac{E_{xy}[|\Delta f_y|] M}{\alpha}
\]

\[
\leq K_f R_L(\theta) + \frac{LMW_1(f(c|xy), \mathcal{C}|xy)}{\alpha}
\]

First inequality follows from Hölder’s inequality, and the fact that \( 0 \leq \ell(x, y) \leq M \). The second one follows from the assertion that \( \frac{1}{K_f} \hat{h} \in \mathcal{H}_L \), and the last inequality follows from the fact that \( \mathcal{H} \) is \( L \)-Lipschitz, and utilizing the pushforward measure form of \( \Delta \).

A analogous argument shows the other side of this bound given by,

\[
R_L(\theta) \leq K_g \mathcal{R}_L(\theta) + \frac{LM_{XY}W_1(c|xy, g(\mathcal{C}|xy))}{\alpha}
\]

This shows that our DRO estimator achieves multiplicative error scaling with \( K_f, K_g \) and additive error scaling with the Wasserstein distance between the true and the estimated unmeasured variables.

Our assumptions on \( K_f \) and \( K_g \) are easily fulfilled in the case where there is a single bi-Lipschitz bijection \( f : \mathcal{C} \to \overline{\mathcal{C}} \). In this case, \( g = f^{-1} \) and \( K_f = K_g = K \).

We can interpret this bound as capturing two sources of error: our metric can be inappropriate and our estimates of \( \overline{\mathcal{C}} \) can be inherently noisy. For the first term, note that a map with higher metric distortion (e.g. bi-Lipschitz maps with large constants) results in a looser bound. This is because the Lipschitz function assumption in the original space \( \mathcal{C} \) does not correspond closely to Lipschitz functions in \( \overline{\mathcal{C}} \).

For the second term, we incur error whenever \( W_1(c|xy, g(\mathcal{C}|xy)) \) is large. The alignment map \( g \) takes our elicited unmeasured variables and approximates the true ones. However, if \( \mathcal{C} \) does not contain enough information to reconstruct \( c \) then no function \( g \) can exactly map \( \mathcal{C} \) to \( c \), and we incur an approximation error that scales as the transport distance between the two.

We can now provide a simple lemma that bounds the quality of the model estimate under the approximation \( \mathcal{C} \) compared to the minimizer of the exact unmeasured variables \( c \).
For convenience we will use the following shorthand for the additive error terms,

\[
A_f = \frac{\mathbb{E}_{X,Y} W_1(\tau|xy, f(c|xy))}{\alpha}
\]

\[
A_g = \frac{\mathbb{E}_{X,Y} W_1(c|xy, g(\tau|xy))}{\alpha}.
\]

**Corollary.** Let \( \theta^* := \arg\min_\theta \mathcal{R}_L(\theta) \), then

\[
\mathcal{R}_L(\theta^*) - \inf_\theta \mathcal{R}_L(\theta) \\
\leq \inf_\theta \mathcal{R}_L(\theta) (K_f K_g - 1) + K_g A_f + A_g
\]

**Proof.** By Proposition \( S4 \), we have both

\[
\inf_\theta \mathcal{R}_L(\theta) \leq \inf_\theta K_f \mathcal{R}_L(\theta) + A_f
\]

\[
\mathcal{R}_L(\theta^*) \leq K_g \mathcal{R}_L(\theta^*) + A_g.
\]

By definition of \( \theta^* \) as the minimizer of \( \mathcal{R}_L \), we obtain

\[
\mathcal{R}_L(\theta^*) \leq K_f K_g \inf_\theta \mathcal{R}_L(\theta) + K_g A_f + A_g
\]

which gives the stated result. 

The corollary shows that the best model under the estimated unmeasured variables \( \tau \) performs well under the true DRO risk measure \( \mathcal{R}_L \) as long as \( K_f K_g \approx 1 \) and \( A_f, A_g \) are small. There are two sources of error: the metric distortion results in a relative error that scales as \( K_f K_g \), and the noise in estimation \( (A_f, A_g) \) results in additive error. The \( K_g A_f \) scaling term arises from the fact that error is measured with respect to the metric over \( c \), not over \( \tau \).

Importantly, these bounds show that we need not directly estimate the true unmeasured variables \( c \) using \( \tau \) - our estimated unmeasured variables can live in an entirely different space, and as long as there exists some low-distortion alignment functions \( f, g \) that align the two spaces, the implied risk functions are similar.

**S5. Reproducibility & Experiment Details**

All experiments and data described below are available on CodaLab: \texttt{https://bit.ly/uvdro-codalab}.

**S5.1. Simulated Medical Diagnosis Task**

We simulate our data \( (n=1,000) \) using the following generation procedure:

1. \( q_{\text{train}} = .05, .1, .2, .3, .4, .5, .6, .7, .8 \) and \( q_{\text{test}} = 0.8 \).
2. \( c \) is sampled from the \( c \sim 1 - 2 \text{Bernoulli}(q) \).
3. \( y \) is sampled from \( y \sim \mathcal{N}(0, 2) \), independent from from train or test.
4. For each \((c, y)\) sample, set \( x_1 = c * y \) and \( x_2 = y + \epsilon \) where \( \epsilon \sim \mathcal{N}(0, 4) \).

For both ERM and UV-DRO, we trained a linear regression model over \( p(y|x_1, x_2) \), optimized using batch gradient descent over 3k steps with AdaGrad with an optimal learning rate of .0001. We set UV-DRO parameter \( \alpha = 0.2 \), and tune \( \eta \) via grid-search for each \( q_{\text{train}} \) value. We present results (Mean Squared Error) on the same held-out test set for all models.
S5.2. MNIST Digit Classification with Confounding Transformations

We use the popular MNIST dataset (http://yann.lecun.com/exdb/mnist/). We train on only a subset (n=4000) of the training data due to the cost of collecting annotations, and tune parameters on a separate validation set. For all data points, we treat the pixels of a (possibly transformed) image as the features \( x \), the fact of whether a transformation occurred as the unmeasured variable \( c \), and the MNIST digit as label \( y \). We simulate a shift in an unmeasured rotation confounding variable using the following procedure:

1. \( q_{\text{train}} = .05, .1, .2, .4, .6 \) and \( q_{\text{test}} = 1.0 \).
2. \( c \) is sampled from the \( c \sim \text{Bernoulli}(q) \), where \( c = 1 \) means the image was rotated.
3. For each \((x, y)\) pair in the dataset, we rotate the original MNIST image \( x \) by 180 degrees if \( c = 1 \).

For all ERM, DRO, and UV-DRO models, we trained a logistic regression model, optimized with batch gradient descent using AdaGrad and an optimal learning rate of .001. The optimal \( l_2 \) penalty found for ERM models was 25. Optimal UV-DRO parameters (tuned on 20% of data as valid) include \( l_2 \) penalty of 50, a Lipschitz constant \( L \) of 1, \( \alpha = 0.2 \), and we explicitly solve for the minimizer of \( \eta \) with regards to the empirical distribution at each gradient step. We present results (Log-Loss, Accuracy) on the same held-out test set for all models.

S5.3. Police Stop Analysis with Confounding Locations

We use a dataset of NYPD police stops (https://www.nyclu.org/en/stop-and-frisk-data). We train on only a subset (n=2000) of the training data due to the cost of collecting annotations, and tune parameters on a separate validation set. For all data points, we filter out all variables except for 26 police stop observation as features \( x \) (i.e. “in a high crime area”), the NYC borough as the unmeasured location variable \( c \), and the label for arrest \( y \). We simulate a shift in the location variable \( c \) using the following procedure:

1. \( q_{\text{train}} = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 \) and \( q_{\text{test}} = 1.0 \).
2. \( c \) is sampled from the \( c \sim \text{Bernoulli}(q) \), where \( c = 1 \) means the location is Brooklyn.
3. We build the dataset by drawing from the entire dataset a \((x, y, c = c')\) example for each \( c' \) sampled.

For all ERM, DRO, and UV-DRO models, we trained a logistic regression model optimized with batch gradient descent using AdaGrad and an optimal learning rate of .005. The optimal \( l_2 \) penalty found for ERM models was 0. Optimal UV-DRO (tuned on 20% of data as valid) parameters include \( l_2 \) penalty of 50, a Lipschitz constant \( L \) of 1, \( \alpha = 0.2 \), and we explicitly solve for the minimizer of \( \eta \) with regards to the empirical distribution at each gradient step. We present results (Log-Loss, Accuracy) on the same held-out test set for all models.

References