A PROOFS OF THEORETICAL DISCUSSIONS

A.1 LEMMA A.1 AND PROOF

Lemma A.1. *With sample* $(\mathbf{x}, y)$ *and two labeling functions* $f_1(\mathbf{x}) = f_2(\mathbf{x}) = y$*, for an estimated* $\theta \in \Theta$*, if* $\theta(\mathbf{x}) = y$*, then with* ***A3*** *, we have*

$$d(\theta, f_1, \mathbf{x}) = 1 \implies r(\theta, \mathcal{A}(f_2, \mathbf{x})) = 1. \tag{20}$$

Proof. If $\theta(\mathbf{x}) = y$ and $d(\theta, f_1, \mathbf{x}) = 1$, according to **A3**, we have $d(\theta, f_2, \mathbf{x}) = 0$.

We prove this by contradiction.

If the conclusion does not hold, $r(\theta, \mathcal{A}(f_2, \mathbf{x})) = 0$, which means

$$\max_{\mathbf{x}_{\mathcal{A}(f_2,\mathbf{x})} \in \mathcal{X}_{\mathcal{A}(f_2,\mathbf{x})}} |\theta(\mathbf{x}) - y| = 0 \tag{21}$$

Together with $d(\theta, f_2, \mathbf{x}) = 0$, which means

$$\max_{\mathbf{z} \in \mathcal{X}: \mathbf{z}_{\mathcal{A}(f_2,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f_2,\mathbf{x})}} |\theta(\mathbf{z}) - y| = 0, \tag{22}$$

we will have

$$\max_{\mathbf{x} \in \mathcal{X}} |\theta(\mathbf{x}) - y| = 0, \tag{23}$$

which is $\theta(\mathbf{x}) = y$ for any $\mathbf{x} \in \mathbf{P}$.

This contradicts with the premises in **A3** (θ is not a constant function).

□

A.2 THEOREM 3.1 AND PROOF

Theorem. *With Assumptions* ***A1****-****A3****, with probability at least* $1 - \delta$*, we have*

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta) + \phi(|\Theta|, n, \delta) \tag{24}$$

where $c(\theta) = \frac{1}{n} \sum_{(\mathbf{x},y) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta, \mathcal{A}(f_m, \mathbf{x}))$.

Proof.

$$\widehat{\epsilon}_{\mathbf{P}_s}(\theta) = \frac{1}{n} \sum_{(\mathbf{x},y) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} |\theta(\mathbf{x}) - f(\mathbf{x})| \tag{25}$$

$$= 1 - \frac{1}{n} \sum_{(\mathbf{x},y) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})]\big) \tag{26}$$

$$= 1 - \frac{1}{n} \sum_{(\mathbf{x},y) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})] \mathbb{I}[d(\theta, f_h, \mathbf{x}) = 0] + \mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})] \mathbb{I}[d(\theta, f_h, \mathbf{x}) = 1]\big) \tag{27}$$

$$= 1 - \frac{1}{n} \sum_{(\mathbf{x},y) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})] \mathbb{I}[d(\theta, f_h, \mathbf{x}) = 0]\big) - \frac{1}{n} \sum_{(\mathbf{x},y) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})] \mathbb{I}[d(\theta, f_h, \mathbf{x}) = 1] \tag{28}$$

$$\geq \widehat{\epsilon}_h(\theta) - \frac{1}{n} \sum_{(\mathbf{x},y) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})] r(\theta, \mathcal{A}(f_m, \mathbf{x})), \tag{29}$$

where the last line used Lemma A.1.

Thus, we have

$$\widehat{\epsilon}_h(\theta) \leq \widehat{\epsilon}(\theta) + \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})] r(\theta, \mathcal{A}(f_m, \mathbf{x})) \tag{30}$$

where

$$\widehat{\epsilon}_h(\theta) = 1 - \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})]\mathbb{I}[d(\theta, f_h, \mathbf{x}) = 0]\big), \tag{31}$$

which describes the correctly predicted terms that θ functions the same as f_h and all the wrongly predicted terms. Therefore, conventional generalization analysis through uniform convergence applies, and we have

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_h(\theta) + \phi(|\Theta|, n, \delta) \tag{32}$$

Thus, we have:

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_{\mathbf{P}_s}(\theta) + \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta, \mathcal{A}(f_m, \mathbf{x})) + \phi(|\Theta|, n, \delta) \tag{33}$$

□

A.3 THEOREM 3.2 AND PROOF

Theorem. *With Assumptions **A2**-**A4**, and if* $1 - f_h \in \Theta$*, we have*

$$c(\theta) \leq D_\Theta(\mathbf{P}_s, \mathbf{P}_t) + \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta, \mathcal{A}(f_m, \mathbf{x})) \tag{34}$$

where $c(\theta) = \frac{1}{n}\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta, \mathcal{A}(f_m, \mathbf{x}))$ *and* $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$ *is defined as in* (8).

Proof. By definition, $g(\mathbf{x}) \in \Theta\Delta\Theta \iff g(\mathbf{x}) = \theta(\mathbf{x}) \oplus \theta'(\mathbf{x})$ for some $\theta, \theta' \in \Theta$, together with Lemma 2 and Lemma 3 of [Ben-David et al., 2010], we have

$$\begin{aligned}
D_\Theta(\mathbf{P}_s, \mathbf{P}_t) =& \frac{1}{n} \max_{\theta,\theta'\in\Theta} \Big| \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} |\theta(\mathbf{x}) - \theta'(\mathbf{x})| - \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} |\theta(\mathbf{x}) - \theta'(\mathbf{x})| \Big| && (35)\\
\geq& \frac{1}{n} \Big| \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} |\theta(\mathbf{x}) - f_z(\mathbf{x})| - \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} |\theta(\mathbf{x}) - f_z(\mathbf{x})| \Big| && (36)\\
=& \frac{1}{n} \Big| \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y] - \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y] \Big| && (37)\\
=& \frac{1}{n} \Big| \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 1] - \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 1] && (38)\\
&+ \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 0] - \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 0] \Big| && (39)\\
=& \frac{1}{n} \Big| \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta, \mathcal{A}(f_m, \mathbf{x})) - \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta, \mathcal{A}(f_m, \mathbf{x})) \Big| && (40)\\
\geq& c(\theta) - \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta, \mathcal{A}(f_m, \mathbf{x})) && (41)
\end{aligned}$$

First line: see Lemma 2 and Lemma 3 of [Ben-David et al., 2010].

Second line: if $1 - f_h \in \Theta$, and we use f_z to denote $1 - f_h$.

Fifth line is a result of using that fact that

$$\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 0] = \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 0] \tag{42}$$

as a result of our assumptions. Now we present the details of this argument:

According to **A3**, if $\theta(\mathbf{x}) = y$, $d(\theta, f_h, \mathbf{x})d(\theta, f_m, \mathbf{x}) = 0$. Since $r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 0$, $d(\theta, f_m, \mathbf{x})$ cannot be 0 unless θ is a constant mapping that maps every sample to 0 (which will contradicts **A3**). Thus, we have $d(\theta, f_h, \mathbf{x}) = 0$.

Therefore, we can rewrite the left-hand term following

$$\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 0] = \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[d(\theta, f_h, \mathbf{x}) = 0] \tag{43}$$

and similarly

$$\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[r(\theta, \mathcal{A}(f_m, \mathbf{x})) = 0] = \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[d(\theta, f_h, \mathbf{x}) = 0] \tag{44}$$

We recap the definition of $d(\cdot, \cdot, \mathbf{x})$, thus $d(\theta, f_h, \mathbf{x}) = 0$ means

$$d(\theta, f_h, \mathbf{x}) = \max_{\mathbf{z}\in\mathcal{X}:\mathbf{z}_{\mathcal{A}(f,\mathbf{x})}=\mathbf{x}_{\mathcal{A}(f_h,\mathbf{x})}} |\theta(\mathbf{z}) - f_h(\mathbf{z})| = 0 \tag{45}$$

Therefore $d(\theta, f_h, \mathbf{x}) = 0$ implies $\mathbb{I}(\theta(\mathbf{x}) = y)$, and

$$|\theta(\mathbf{z}) - f_h(\mathbf{z})| = 0 \quad \forall \quad \mathbf{z}_{\mathcal{A}(f_h,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f_h,\mathbf{x})} \tag{46}$$

Therefore, we can continue to rewrite the left-hand term following

$$\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[d(\theta, f_h, \mathbf{x}) = 0] = \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{z}) - f_h(\mathbf{z})] = \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) - f_h(\mathbf{x})] \tag{47}$$

and similarly

$$\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = y]\mathbb{I}[d(\theta, f_h, \mathbf{x}) = 0] = \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{z}) - f_h(\mathbf{z})] \tag{48}$$

where $\mathbf{z}$ denotes any $\mathbf{z} \in \mathcal{X}$ and $\mathbf{z}_{\mathcal{A}(f_h,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f_h,\mathbf{x})}$.

Further, because of **A4**, we have

$$\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{z}) - f_h(\mathbf{z})] = \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) - f_h(\mathbf{x})]. \tag{49}$$

Thus, we show the (42) holds and conclude our proof. □

B ADDITIONAL DISCUSSION TO CONNECT TO ROBUST MACHINE LEARNING METHODS

B.1 WORST-CASE DATA AUGMENTATION IN PRACTICE

In practice, when we use data augmentation to learn human-aligned models, we need either of the two following assumptions to hold:

A4-1: **Labeling Functions Separability of Features** For any $\mathbf{x} \in \mathcal{X}$, $\mathcal{A}(f_h, \mathbf{x}) \cap \mathcal{A}(f_m, \mathbf{x}) = \emptyset$

A4-2: **Labeling Functions Separability of Input Space** We redefine $f_m : \text{dom}(f_m) \to \mathcal{Y}$ and $\text{dom}(f_m) \subsetneq \mathcal{X}$. For any $\mathbf{x} \in \mathcal{X}$, $\max_{\mathbf{z} \in \text{dom}(f_m) \cap \text{dom}(f_h)} |f_h(\mathbf{z}) - f_h(\mathbf{z})| = 0$

While both of these assumptions appear strong, we believe a general discussion of human-aligned models may not be able to built without these assumptions. In particular, **A4-1** describes the situations that f'_h do not use the same set of features as f'_m. One example of this situation could be that the background of an image in dog vs. cat classification is considered features for f'_m, and the foreground of an image is considered as features for f'_h. **A4-2** describes the situations that while f'_m can uses the features that are considered by f'_h, the perturbation of the features within the domain of f'_m will not change the output of f'_h. One example of this situation could be that the texture of dog or cat in the dog vs. cat classification, while the texture can be perturbed, the perturbation cannot be allowed to an arbitrary scale of pixels (otherwise the perturbation is not a perturbation of texture). If neither of these assumptions holds, then the perturbation will be allowed to replace a dog's body with the one of a dolphin, and even human may not be able to confidently decide the resulting image is a dog, thus human-aligned learning will not be worth discussion.

B.2 DERIVATION OF WEIGHTED RISK MINIMIZATION.

B.2.1 Connections to Distributionally Robust Optimization (DRO)

Recall that we generalize the above analysis of worst-case data augmentation to a DRO problem [Ben-Tal et al., 2013, Duchi et al., 2021]. Given n data points, consider a perturbation set $\mathcal{Q} := \{\mathbf{x}_{\mathcal{A}(f_m, \mathbf{x}_i)} \in \text{dom}(f)_{\mathcal{A}(f_m, \mathbf{x}_i)}\}_{i=1}^n$ encoding the features of $\mathbf{x}$ indexed by $\mathcal{A}(f, \mathbf{x})$ over input space $\text{dom}(f_m)$. Denote $q(\mathbf{x}, y)$ and $p(\mathbf{x}, y)$ are densities from the $\mathcal{Q}$ and training distribution $\mathcal{X} \times \mathcal{Y}$, respectively. Then (12) can be rewritten as a DRO problem over a new distribution $\mathcal{Q}$.

$$c(\theta) \leq \min_{\theta \in \Theta} \max_{\mathbf{z} \in \mathcal{Q}(\mathbf{x})} \frac{1}{n} \sum_{(\mathbf{x}, y) \in (\mathbf{X}, \mathbf{Y})} \ell(\theta(\mathbf{z}), y) \tag{50}$$

To transform DRO into WRM, we introduce the following assumptions about perturbation set $\mathcal{Q}$:

A2-1: $q \ll p$. $p(x, y) = 0 \implies q(x, y) = 0$

A2-2: **f-Divergence.** Given a function ξ is convex and $\xi(1) = 0$ and $\delta > 0$ as a radius to control the degree of the distribution shift, $D_\xi\left(q\left(\mathbf{x}, \mathbf{y}\right) \| p\left(\mathbf{x}, \mathbf{y}\right)\right) \leq \delta$ holds.

$\mathcal{Q}$ encodes the priors about feature perturbation that model should be robust to. Therefore, choosing f-divergence as the distance metric where ξ is convex with $\xi(1) = 0$, $\delta > 0$ as a radius to control the degree of the distribution shift, adversarial robustness in Section 4.1 can be viewed as an example of DRO on an infinite family of distributions with implicit assumptions that samples in $\mathcal{Q}$ are visually indistinguishable from original ones. For p and q that $p(\mathbf{x}, y) = 0$ implies $q(\mathbf{x}, y) = 0$, we arrive at a generic weighted risk minimization (WRM) formulation [Namkoong and Duchi, 2016, Duchi et al., 2021] when weights (by default as density ratios) $\lambda = q(\mathbf{x}, y)/p(\mathbf{x}, y)$ in (51) derived from misaligned functions for

$$c(\theta) \leq \min_{\theta \in \Theta} \max_{\mathbf{z} \in \mathcal{Q}_\xi(\mathbf{x})} \frac{1}{n} \sum_{(\mathbf{x}, y) \in (\mathbf{X}, \mathbf{Y})} \lambda\left(\mathbf{z}\right) \cdot \ell\left(\theta\left(\mathbf{z}\right), y\right) \tag{51}$$

where the uncertainty set $\mathcal{Q}_\xi$ is reformulated as

$$\mathcal{Q}_\xi := \{\lambda(\mathbf{z}_i)|D_\xi(q||p) \leq \delta, \tag{52}$$

$$\sum_{i=1}^{n} \lambda(\mathbf{z}_i) = 1, \tag{53}$$

$$\forall \lambda(\mathbf{z}_i) \geq 0\} \tag{54}$$

When $\lambda(\cdot) = q(\mathbf{x}, \mathbf{y})/p(\mathbf{x}, \mathbf{y})$ is the density ratio, we use change of measure technique to show the equivalence of DRO and WRM by transoforming the optimization problem on q to an optimization problem $\lambda(\cdot)$. And the inner optimization problem are equivalent to

$$\mathbb{E}_q[\ell(\theta, \mathbf{x})] = \int \ell(\theta, \mathbf{x}) q(\mathbf{z}) d\mathbf{z} = \int \ell(\theta, \mathbf{x}) \frac{q(\mathbf{z})}{p(\mathbf{z})} p(\mathbf{z}) d\mathbf{z} = \mathbb{E}_p[\lambda(\mathbf{x})\ell(\theta, \mathbf{x})] \tag{55}$$

Moreover, choosing $f = x\log(x)$, f-divergence becomes KL-divergence and then the constraint can be converted to

$$D_\xi(p\|q) = \int_{q>0} \frac{p(\mathbf{x})}{q(\mathbf{x})} \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) q(\mathbf{x}) d\mathbf{x} = \mathbb{E}_q[\lambda(\mathbf{x}) \log \lambda(\mathbf{x})] \leq \delta \tag{56}$$

Next we prove the equivalence of DRO and WRM for general λ under additional assumptions below.

A2-3: **Finite perturbation set.** $\mathcal{Q}$ is a finite set.

A2-4: **Convexity.** Loss function ℓ is convex in θ and concave in λ. $\mathcal{Q}$ and Θ are convex sets.

A2-5: **Continuity.** Loss function ℓ and its weighted sum $\sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})} \lambda(\mathbf{z})\ell\left(\theta\left(\mathbf{z}\right), \mathbf{y}\right)$ are continuous.

A2-6: **Compactness.** $\mathcal{Q}$ and Θ are compact.

Given n data points, we introduce slack variable ξ and consider a constrained optimization formulation of (50) as

$$\min_{\theta\in\Theta,\xi} \xi \qquad s.t. \quad \sum_{i=1}^{n} \lambda(\mathbf{z}_i)\ell(\theta(\mathbf{z}_i), y_i) - \xi \leq 0 \quad \forall \mathbf{z} \in \mathcal{Q} \tag{57}$$

By the strong convex duality, we have the Lagrangian $L(\theta, \alpha, \lambda(\mathbf{z}_1), \ldots, \lambda(\mathbf{z}_n)) = \alpha + \sum_{i=1}^{n} \lambda(\mathbf{z}_i)(\ell(\theta(\mathbf{z}_i), y_i) - \alpha)$ and the dual problem as

$$\max_{\forall\lambda_i\succeq 0} \min_{\theta,\alpha} L(\theta, \alpha, \lambda(\mathbf{z}_1), \ldots, \lambda(\mathbf{z}_n)) \quad s.t. \sum_{i=1}^{n} \lambda(\mathbf{z}_i) = 1, \quad i = 1, \ldots, n \tag{58}$$

Which can be expressed as

$$\max_{\forall\lambda\succeq 0} \min_{\theta\in\Theta} \mathcal{L}(\lambda, \theta) = \max_{\forall\lambda_i\succeq 0} \min_{\theta\in\Theta} \sum_{i=1}^{n} \lambda_i \ell(\theta(\mathbf{z}_i), y_i) \quad s.t. \sum_{i=1}^{n} \lambda(\mathbf{z}_i) = 1, i = 1, \ldots, n \tag{59}$$

By the minimax equality, we have

$$\max_{\forall\lambda\succeq 0} \min_{\theta\in\Theta} \mathcal{L}(\lambda, \theta) = \min_{\theta\in\Theta} \max_{\forall\lambda\succeq 0} \mathcal{L}(\lambda, \theta) \tag{60}$$

Denote the optimality of $\max_{\forall\lambda\succeq 0} \min_{\theta\in\Theta} \mathcal{L}(\lambda, \theta)$ and $\min_{\theta\in\Theta} \max_{\forall\lambda\succeq 0} \mathcal{L}(\lambda, \theta)$ as λ^* and $\theta^* \in \Theta$, respectively. Then we have (λ^*, θ^*) form a saddle point that

$$\max_{\lambda\succeq 0} \mathcal{L}(\theta^*, \lambda) = \mathcal{L}(\theta^*, \lambda^*) = \min_{\theta\in\Theta} \mathcal{L}(\theta, \lambda^*) \tag{61}$$

which means that λ^* exists in the WRM such that $\theta^* \in \arg\min_\theta \mathcal{L}(\lambda, \theta)$ is optimal for DRO.

Intuitively, learner θ and adversary ϕ are playing a minimax game where ϕ finds worst-case weights and computationally-identifiable regions of errors to improve the robustness of the learner θ. In this scenario, we unify a line of WRM approaches where weights λ are mainly determined by misaligned features $\mathcal{A}(f_m, \mathbf{x})$, either parameterized by a biased model or derived from some heuristic statistics.

B.3 DETAILS TO CONNECT METHODS TO REGULARIZE THE HYPOTHESIS SPACE

First, we need to formally introduce the properties regarding f'_m, as a correspondence to those of f_m.

Notations and Background with Encoder/Decoder Structure With the same binary classification problem from feature space $\mathcal{X}$ to label space $\mathcal{Y}$. We consider the encoder $\theta_e : \mathcal{X} \to \mathcal{E}$ and decoder $\theta_d : \mathcal{E} \to \mathcal{Y}$, $f' : \mathcal{E} \to \mathcal{Y}$ is the function that maps the embedding to the label.

Similarly, we introduce the assumptions on the $\mathcal{E}$ space.

A2': **Existence of Superficial Features:** For any $\mathbf{x} \in \mathcal{X}$ and an oracle encoder θ_e that $\mathbf{e} = \theta_e(\mathbf{x})$, $y := f'_h(\mathbf{e})$. We also have a f'_m that is different from f'_h, and for $\mathbf{x} \sim \mathbf{P}_s$ and $\mathbf{e} = \theta_e(\mathbf{x})$, $f'_h(\mathbf{e}) = f'_m(\mathbf{e})$.

A3': **Realized Hypothesis:** Given a large enough hypothesis space Θ_d for decoders, for any sample $(\mathbf{x}, y)$ and an encoder θ_e that $\mathbf{e} = \theta_e(\mathbf{x})$, for any $\theta_d \in \Theta_d$, which is not a constant mapping, if $\theta_d(\mathbf{e}) = y$, then $d(\theta_d, f'_h, \mathbf{e})d(\theta_d, f'_m, \mathbf{e}) = 0$

With the above assumptions, following the same logic, we can derive the theorem corresponding to Theorem 3.1, with the only difference that how $c(\theta)$ is now derived.

Lemma B.1. *With Assumptions **A1**, **A2'**, **A3'**, $l(\cdot, \cdot)$ is a zero-one loss, with probability as least $1 - \delta$, we have*

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta) + \phi(|\Theta|, n, \delta) \tag{62}$$

where $c(\theta) = \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = y] r(\theta_d, \mathcal{A}(f'_m, \theta_e(\mathbf{x})))$.

Now, we continue to show that how training for small $c(\theta)$ amounts to solving (17). To proceed, we need either of the two following assumptions to hold:

A4-1': **Labeling Functions Separability of Features** For any $\mathbf{x} \in \mathcal{X}$ and an encoder θ_e that $\mathbf{e} = \theta_e(\mathbf{x})$, $\mathcal{A}(f'_h, \mathbf{e}) \cap \mathcal{A}(f'_m, \mathbf{e}) = \emptyset$

A4-2': **Labeling Functions Separability of Input Space** We redefine $f'_m : \text{dom}(f'_m) \to \mathcal{Y}$ and $\text{dom}(f'_m) \subsetneq \mathcal{E}$. For any $\mathbf{x} \in \mathcal{X}$ and an encoder θ_e that $\mathbf{e} = \theta_e(\mathbf{x})$, $\max_{\mathbf{z}\in\text{dom}(f'_m)\cap\text{dom}(f'_h)} |f'_h(\mathbf{z}) - f'_h(\mathbf{z})| = 0$

Also, notice that, assumptions **A4-1'** and **A4-2'** also regulates the encoder to be reasonably good. In other words, these assumptions will not hold for arbitrary encoders.

Now, we continue to derive (17) from Lemma B.1 as the following:

$$\begin{aligned}
c(\theta) &= \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})} \mathbb{I}[\theta_d(\theta_e(\mathbf{x})) = y] r(\theta_d, \mathcal{A}(f'_m, \mathbf{x})) \\
&= \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})} \mathbb{I}[\theta_d(\theta_e(\mathbf{x})) = y] \max_{\theta_e(\mathbf{x})_{\mathcal{A}(f'_m,\mathbf{x})}\in\text{dom}(\theta_d)_{\mathcal{A}(f'_m,\mathbf{x})}} |\theta_d(\theta_e(\mathbf{x})) - y| \\
&= \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})} \max_{\theta_e(\mathbf{x})_{\mathcal{A}(f'_m,\mathbf{x})}\in\text{dom}(\theta_d)_{\mathcal{A}(f'_m,\mathbf{x})}} |f'_m(\theta_e(\mathbf{x})) - y| \\
&\leq \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})} \max_{\theta_e(\mathbf{x})\in\text{dom}(\theta_d)} |f'_m(\theta_e(\mathbf{x})) - y|
\end{aligned}$$

The third line is because of the definition of $\mathbb{I}[\theta_d(\theta_e(\mathbf{x})) = y] r(\theta_d, \mathcal{A}(f'_m, \mathbf{x}))$ and assumptions of A3' and either **A4-1'** or **A4-2'**. Therefore, optimizing the empirical loss and $c(\theta)$ leads to

$$\min_{\theta_d,\theta_e} \frac{1}{n} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{Y})} l(\theta_d(\theta_e(\mathbf{x})), y) - l(f'_m(\theta_e(\mathbf{x})), y)$$

C THEORY-SUPPORTING EXPERIMENTS

Synthetic Data with Spurious Correlation We extend the setup in Figure 1 to generate the synthetic dataset to test our methods. We study a binary classification problem over the data with n samples and p features, denoted as $\mathbf{X} \in \mathcal{R}^{n \times p}$. For every training and validation sample i, we generate feature j as following:

$$\mathbf{X}_j^{(i)} \sim \begin{cases} N(0,1) & \text{if } 1 \le j \le 3p/4 \\ N(1,1) & \text{if } 3p/4 < j \le p, \text{ and } y^{(i)} = 1, \quad \text{w.p. } \rho \\ N(-1,1) & \text{if } 3p/4 < j \le p, \text{ and } y^{(i)} = 0, \quad \text{w.p. } \rho \\ N(0,1) & \text{if } 3p/4 < j \le p, \quad \text{w.p. } 1-\rho \end{cases},$$

In contrast, testing data are simply sampled with $\mathbf{x}_j^{(i)} \sim N(0,1)$.

To generate the label for training, validation, and test data, we sample two effect size vectors $\beta_1 \in \mathcal{R}^{p/4}$ and $\beta_2 \in \mathcal{R}^{p/4}$ whose each coefficient is sampled from a Normal distribution. We then generate two intermediate variables:

$$\mathbf{c}_1^{(i)} = \mathbf{X}_{1,2,\ldots,p/4}^{(i)} \beta_1 \quad \text{and} \quad \mathbf{c}_2^{(i)} = \mathbf{X}_{1,2,\ldots,p/4}^{(i)} \beta_2$$

Then we transform these continuous intermediate variables into binary intermediate variables via Bernoulli sampling with the outcome of the inverse logit function ($g^{-1}(\cdot)$) over current responses, *i.e.*,

$$\mathbf{r}_1^{(i)} = \text{Ber}(g^{-1}(\mathbf{c}_1^{(i)})) \quad \text{and} \quad \mathbf{r}_2^{(i)} = \text{Ber}(g^{-1}(\mathbf{c}_2^{(i)}))$$

Finally, the label for sample i is determined as $y^{(i)} = \mathbb{I}(\mathbf{r}_1^{(i)} = \mathbf{r}_2^{(i)})$, where $\mathbb{I}$ is the function that returns 1 if the condition holds and 0 otherwise.

Intuitively, we create a dataset of p features, half of the features are generalizable across train, validation and test datasets through a non-linear decision boundary, one-forth of the features are independent of the label, and the remaining features are spuriously correlated features: these features are correlated with the labels in train and validation set, but independent with the label in test dataset. There are about ρn train and validation samples have the correlated features.

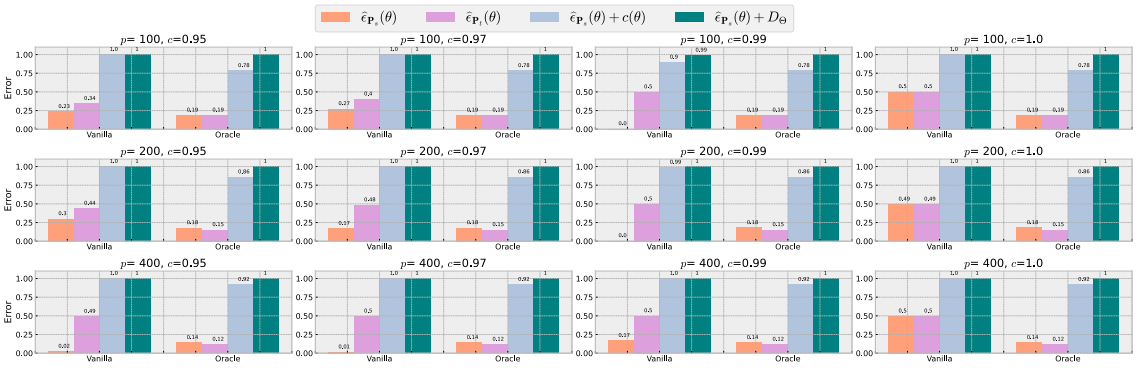

Figure 2: Results of Synthetic Data with Spurious Correlation. Each panel represents one setting. Five methods are reported in each panel. For each method, four bars are plotted: from left to right, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta)$, $\widehat{\epsilon}_{\mathbf{P}_t}(\theta)$, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta)$, and $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + D_\Theta$.

We train a vanilla ERM method, and in comparison, we also train an oracle method which that uses data augmentation to randomized the previously known spurious features. We report training error (*i.e.*, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta)$), test error (*i.e.*, $\widehat{\epsilon}_{\mathbf{P}_t}(\theta)$), $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta)$, and $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + D_\Theta$ so that we can directly compare the bars to evaluate whether $c(\theta)$ can quantify the expected test error. Our results suggest that $c(\theta)$ is often a tighter estimation of the test error than $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$, which aligns well with our analysis in Section 3.

Binary Digit Classification over Transferable Adversarial Examples For the second one, we consider a binary digit classification task, where the train and validation sets are digits 0 and 1 from MNIST train and validation sets. To create the test set, we first estimate a model, and perform adversarial attacks over this model to generate the test samples with

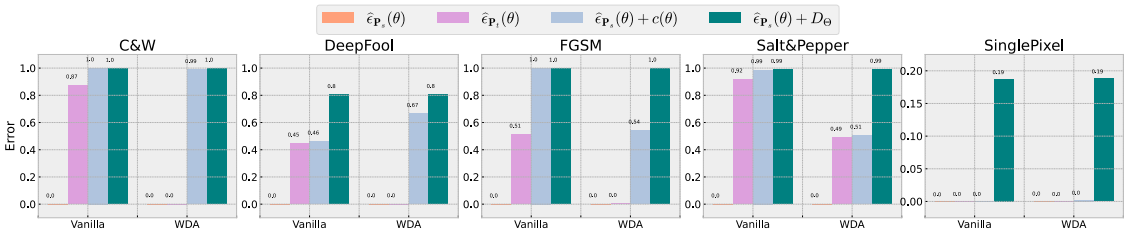

Figure 3: Binary MNIST classification error and estimated bounds. Each panel represents one out-of-domain data generated through an attack method. Four methods are reported in each panel. For each method, four bars are plotted: from left to right, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta)$, $\widehat{\epsilon}_{\mathbf{P}_t}(\theta)$, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta)$, and $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + D_\Theta$. Some bars are not visible because the values are small.

five adversarial attack methods (C&W, DeepFool, FGSM, Salt&Pepper, and SinglePixel). These adversarially generated examples are considered as the test set from another distribution.

An advantage of this setup is that we can have f_m well defined as $1 - f_{adv}$, where the f_{adv} is the function each adversarial attack relies on. Thus, according to our discussion on the estimation of $c(\theta)$ in Section 3, we can directly use the corresponding adversarial attack methods to estimate $c(\theta)$ in our case. Therefore, we can assess our analysis on image classification.

We train the models with the vanilla method, and worst-case data augmentation (WDA, *i.e.*, adversarial training). In addition to the training error (*i.e.*, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta)$) and test error (*i.e.*, $\widehat{\epsilon}_{\mathbf{P}_t}(\theta)$), we also report $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta)$ and $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + D_\Theta$ so that we can directly compare the bars to evaluate whether $c(\theta)$ can quantify the expected test error. By comparing the four different bars within every panel for every method, we notice that $c(\theta)$ is often a tighter estimation of the test error than $D_\Theta(\mathbf{P}_s, \mathbf{P}_t)$, which aligns well with our analysis in Section 3.