A RELATED WORKS

Domain generalization/Domain adaptation: In many real scenarios of machine learning, data in training phase is sampled from one or many source domains, while in the testing phase, data is sampled from an unseen target domain. Many works have been proposed to design robust ML models that can achieve good performances in deployment environment depending on whether they can access to the target data (domain adaptation) or not (domain generalization). However, most of these models focus only on transfering accuracy from source to target domains and can be categorized into five main approaches: (1) data manipulation (Volpi et al., 2018; Qiao et al., 2020; Zhou et al., 2020; Zhang et al., 2018; Shankar et al., 2018); (2) domain-invariant representation learning (Li et al., 2018b;a; Ganin & Lempitsky, 2015; Ganin et al., 2016; Phung et al., 2021; Nguyen et al., 2021); (3) distributional robustness (Krueger et al., 2021; Liu et al., 2021; Koh et al., 2021; Wang et al., 2021; Sagawa et al., 2019; Hu et al., 2018), (4) gradient operation (Huang et al., 2020; Shi et al., 2021; Rame et al., 2021; Tian et al., 2022), and (5) self-supervised learning (Carlucci et al., 2019; Kim et al., 2021; Jeon et al., 2021; Li et al., 2021).

Fairness in Machine Learning: Many fairness notions have been proposed to measure the unfairness in ML model, and they can be roughly classified into two classes: *Individual fairness* considers the equity at the individual-level and it requires that similar individuals should be treated similarly (Biega et al., 2018; Bechavod et al., 2020; Gupta & Kamble, 2021; Dwork et al., 2012). *Group fairness* attains a certain balance in the group-level, where the entire population is first partitioned into multiple groups and certain statistical measures are equalized across different groups (Hardt et al., 2016; Zhang et al., 2019; 2020). Various approaches have also been developed to satisfy these fairness notions, they roughly fall into three categories: (1) *Pre-processing*: modifying training dataset to remove bias before learning an ML model (Kamiran & Calders, 2012; Zemel et al., 2013). (2) *In-processing*: attain fairness during the training process by imposing certain fairness constraint or modifying loss function. (Zafar et al., 2019; Agarwal et al., 2018) (3) *Post-processing*: altering the output of an existing algorithm to satisfy a fairness constraint after training (Hardt et al., 2016). However, most of these methods assume the data distributions at training and testing are the same. In contrast, we study fairness problem under domain generalization in this paper.

Fairness under Domain Adaptation: There are some studies proposed to achieve good fairness when the testing environment changes but all of them focused on the domain adaptation setting. The most common adaptation setup is learning under the assumption of covariate shift. For example, Singh et al. (2021) leveraged a feature selection method in a causal graph describing data to mitigate fairness violation under covariate shift of distribution in testing data. Coston et al. (2019) proposed the weighting methods that can give fair prediction under covariate shift between source and target distribution when access to the sensitive attributes is prohibited. Rezaei et al. (2021) sought fair decisions by optimizing a worst-case testing performance. Besides convariate shift, there are some works proposed to handle other types of distribution shift including demographic shift and prior probability shift. Instead of learning fair model directly, Oneto et al. (2019) and Madras et al. (2018) find fair representation that can generalize to the new tasks. Aside from empirical studies, Schumann et al. (2019) and Yoon et al. (2020) developed theoretical frameworks to examine fairness transfer in domain adaptation setting and then offered modeling approaches to achieve good fairness in the target domain.

Comparison with existing bounds in the literature: We compare our bounds with most commons bound in the fields of domain adaptation and domain generalization as follows.

Accuracy bounds in domain adaptation.

- Bounds in Ben-David et al. (2010):

$$\epsilon^{\mathtt{Acc}}_{D^T}\left(\widehat{f}\right) \leq \epsilon^{\mathtt{Acc}}_{D^S}\left(\widehat{f}\right) + \mathcal{D}_{TV}\left(P^X_{D^S} \parallel P^X_{D^T}\right) + \min_{D\in\{D^S,D^T\}} \mathbb{E}_D\left[|f_{D^S}(X) - f_{D^T}(X)|\right]$$

  This bound is for binary classification problem under domain adaptation. The classification error in target domain is bounded by the error in source domain, the total variation distance of feature distribution between source and target domain, and the misalignment of the labeling function between source and target domain. The limitation of this bound is that (1) it's only applicable to settings with zero-one loss function and deterministic labeling function; (2) estimating the total variation distance is hard in practice and it doesn't relate the feature and representation spaces.

This paper also provides another accuracy bound based on $\mathcal{H}\Delta\mathcal{H}$ divergence:.

$$\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right) \leq \epsilon^{\texttt{Acc}}_{D^S}\left(\widehat{f}\right) + \mathcal{D}_{\mathcal{H}\Delta\mathcal{H}}\left(P^X_{D^S} \parallel P^X_{D^T}\right) + \inf_{\widehat{f}}\left[\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right) + \epsilon^{\texttt{Acc}}_{D^S}\left(\widehat{f}\right)\right]$$

where $\mathcal{D}_{\mathcal{H}\Delta\mathcal{H}}\left(P^X_{D^S} \parallel P^X_{D^T}\right) = \sup_{\widehat{f}_1, \widehat{f}_1}\left|P_{D^S}\left(\widehat{f}_1(X) \neq \widehat{f}_2(X)\right) - P_{D^T}\left(\widehat{f}_1(X) \neq \widehat{f}_2(X)\right)\right|$ is the $\mathcal{H}\Delta\mathcal{H}$ divergence. However, it has the same limitations as total variation distance mentioned above.

Accuracy bounds in domain generalization.

- Bounds in Albuquerque et al. (2019):

$$\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right) \leq \sum_{i=1}^{N}\pi_i\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{f}\right) + \max_{j,k\in[N]}\mathcal{D}_{\mathcal{H}}\left(P^X_{D^S_j} \parallel P^X_{D^S_k}\right) + \mathcal{D}_{\mathcal{H}}\left(P^X_{D^S_*} \parallel P^X_{D^T}\right)$$
$$+ \min_{D\in\{D^S_*, D^T\}}\mathbb{E}_D\left[\left|f_{D^S_*}(X) - f_{D^T}(X)\right|\right]$$

where $\mathcal{D}_{\mathcal{H}}\left(P^X_{D^S} \parallel P^X_{D^T}\right) = \sup_{\widehat{f}}\left|P_{D^S}\left(\widehat{f}(X) = 1\right) - P_{D^T}\left(\widehat{f}(X) = 1\right)\right|$ is the $\mathcal{H}$ divergence, $P^X_{D^S_*} = \arg\min_{\pi}\mathcal{D}_{\mathcal{H}}\left(\sum_{i=1}^{N}\pi_i P^X_{D^S_i} \parallel P^X_{D^T}\right)$ is the mixture of source domains that is closest to target domain with respect to $\mathcal{H}$ divergence. In this bound, the classification error in target domain is bounded by the convex combination of errors in source domains, the $\mathcal{H}$ divergence between source domains, the $\mathcal{H}$ divergence between target domain and its nearest mixture of source domains, and the misalignment of the labeling function between mixture source domains and target domain. Because this bound is constructed based on $\mathcal{H}$ divergence, it also has the limitations for the bounds in domain adaptation (Ben-David et al., 2010) as we mentioned. This bound can be transformed to the representation space $\mathcal{Z}$ by replacing X by Z in its formula. Then, this bound suggests enforcing invariant constraint of marginal distribution of representation Z across source domains, which has inherent trade-off as shown in Thm. 2. Because the target domain is unknown during training, the mixing weights $\{\pi_i\}_{i=1}^{N}$ are not useful for algorithmic design.

- Bounds in Phung et al. (2021):

$$\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right) \leq \sum_{i=1}^{N}\pi_i\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{f}\right) + C\max_{i\in[N]}\mathbb{E}_{D^S_i}\left[\left\|\left[\left|f_{D^T}(X)_y - f_{D^S_i}(X)_y\right|\right]_{y=1}^{|\mathcal{Y}|}\right\|_1\right]$$
$$+ \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{C\sqrt{2\pi_j}}{N}d_{1/2}\left(P^Z_{D^T}, P^Z_{D^S_i}\right) + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{C\sqrt{2\pi_j}}{N}d_{1/2}\left(P^Z_{D^S_i}, P^Z_{D^S_j}\right)$$

where $d_{1/2}\left(P^X_{D^S_i}, P^X_{D^S_j}\right) = \sqrt{\mathcal{D}_{1/2}\left(P^X_{D^S_i} \parallel P^X_{D^S_j}\right)}$ is Hellinger distance defined based on Hellinger divergence $\mathcal{D}_{1/2}\left(P^X_{D^S_i} \parallel P^X_{D^S_j}\right) = 2\int_{\mathcal{X}}\left(\sqrt{P^X_{D^S_i}} - \sqrt{P^X_{D^S_j}}\right)^2 dX$. This bound relates the feature and representation spaces that the classification error of target domain defined in feature space is bounded by classification errors of source domains defined in feature space, the misalignment of labeling function between target and source domains, and the Hellinger distances between source and target domains and between source domains of marginal distribution of representation Z. While this bound is not limited to zero-one loss and the labeling function can be stochastic, it suggests the alignment of marginal distribution of representation Z across source domains for generalization. Moreover, estimating Hellinger distance can be hard in practice.

The mismatch between existing bounds and adversarial learning approach for domain generalization.

All existing bounds mentioned above suggest minimizing the distances between representation distributions across source domains with respect to some discrepancy measures such as $\mathcal{H}$ divergence, total variation distance, and Hellinger distance. Based on these bounds, adversarial learning-based models are often proposed to minimize these distances. However, there is a misalignment between the objectives of adversarial learning and the bounds which results in the gap between theoretical findings and practical algorithms.

In particular, it has been shown that the objective of the minimax game between the representation mapping and the discriminator is equivalent to minimizing the JS divergence between representation distributions across source domains (Goodfellow et al., 2014). However, minimizing JS divergence does not guarantee the minimization of common distances used in the existing bounds. The details are as follows.

- $\mathcal{H}$ divergence: We show that JS divergence is not the upper bound of $\mathcal{H}$ divergence. Consider an example with two distributions $P(X)$ and $Q(X)$ where $\begin{cases} P(X)=0 & w.p\ \ 1/3 \\ P(X)=1 & w.p\ \ 2/3 \end{cases}$ and $\begin{cases} Q(X)=0 & w.p\ \ 1/3 \\ Q(X)=1 & w.p\ \ 2/3 \end{cases}$. By definition, $\mathcal{D}_{\mathcal{H}}(P \parallel Q) \sim 0.33 > \mathcal{D}_{JS}(P \parallel Q) \sim 0.08$.
- Total variation distance: We have $\mathcal{D}_{JS}(P \parallel Q) \leq \mathcal{D}_{TV}(P \parallel Q) \ \ \forall P, Q$ where $\mathcal{D}_{JS}$ and $\mathcal{D}_{TV}$ are JS divergence and total variation distance, respectively. Then, minimizing JS divergence does not guarantee the minimization of total variation distance.
- Hellinger distance: We have $\mathcal{D}_{JS}(P \parallel Q) \leq \sqrt{2} d_{1/2}(P, Q) \ \ \forall P, Q$ where $d_{1/2}$ is Hellinger distance and total variation distance, respectively. Then, minimizing JS divergence does not guarantee the minimization of Hellinger distance.

Different from the existing bounds, our bounds are based on JS divergence/distance. Then they align with the adversarial learning approach for domain generalization in general, and with our proposed method `FATDM` in particular.

Advantages of our proposed bounds in domain generalization.

In summary, our proposed bounds has several advantages in terms of the following:

- Most existing bounds (Ben-David et al., 2010; Albuquerque et al., 2019) do not relates feature and representation spaces so it is not clear how performance in input space is affected by the representations. In contrast, our bounds connect the representation and input spaces; this further guides us to find representations that lead to good performances in input space.
- Most prior studies adopt $\mathcal{H}$ divergence to measure the dissimilarity between domains, which is limited to deterministic labeling functions and zero-one loss (Ben-David et al., 2010; Albuquerque et al., 2019). In contrast, our bound is more general and is applicable to settings where domains are specified by stochastic labeling functions and general loss functions.
- Distant metrics (i.e., total variation distance, $\mathcal{H}$ divergence, Hellinger divergence, etc.) used in existing bounds (Ben-David et al., 2010; Albuquerque et al., 2019; Phung et al., 2021) are hard to compute in practice. In contrast, our bounds use JS divergence which is aligned with training objective for discriminator in adversarial learning Goodfellow et al. (2014).
- Existing bounds for domain generalization only imply the alignment of marginal distribution of feature across source domains (Albuquerque et al., 2019; Phung et al., 2021). As shown in Thm. 2, methods that learn invariance of marginal distribution have an inherent trade-off and may increase the lower bound of expected loss. In contrast, our bounds suggest the alignment of label-conditional distribution of feature across source domains which has been verified to be more effective in empirical studies (Li et al., 2018b;c; Zhao et al., 2020; Nguyen et al., 2021).
- Regarding the fairness, our work is the first that bounds the unfairness in domain generalization. In particular, our bounds suggest enforcing the invariant constraint of feature distribution given label and sensitive attribute across source domains to transfer fairness to the unseen target domain.

B DETAILS OF ALGORITHM `FATDM`

`FATDM` consists of density mapping functions $m^{y}_{i,j}$ and $m^{y,a}_{i,j}$, $\forall y \in \mathcal{Y}, a \in \mathcal{A}, i, j \in [N]$ (learned by two `DensityMatch` models), feature mapping function g (ResNet18 model), and the classifier $\widehat{h}$. In our study, we experiment with two different `DensityMatch` architectures: StarGAN (i.e., in `FATDM-StarGAN`) and CycleGAN (in `FATDM-CycleGAN`). We show the details of `FATDM-StarGAN` below. For `FATDM-CycleGAN`, the only difference is we used CycleGAN as `DensityMatch` instead of StarGAN. The details of CycleGAN were presented in the original paper (Zhu et al., 2017).

For `FATDM-StarGAN`, each $\texttt{DensityMatch}^{Y}$ (or $\texttt{DensityMatch}^{Y,A}$) consists of a *generator* $\mathsf{G} : \mathcal{X} \times [N] \times [N] \rightarrow \mathcal{X}$ and a *discriminator* $\mathsf{D} : \mathcal{X} \rightarrow [N] \times \{0, 1\}$. The generator takes in real image x and a pair of domain labels i, j as input and generates a fake image; the discriminator aims to predict the domain label of the image generated by the generator and distinguish whether it is fake or real. G and D are learned simultaneously by solving the following optimizations:

$$\begin{aligned} \textbf{Discriminator's objective:} \quad & \min\ \mathcal{L}_{\mathsf{D}}^{\texttt{StarGAN}} := -\mathcal{L}_{adv}^{\texttt{StarGAN}} + \lambda_{cls}\mathcal{L}_{cls(\text{real})}^{\texttt{StarGAN}} \\ \textbf{Generator's objective:} \quad & \min\ \mathcal{L}_{\mathsf{G}}^{\texttt{StarGAN}} := \mathcal{L}_{adv}^{\texttt{StarGAN}} + \lambda_{cls}\mathcal{L}_{cls(\text{fake})}^{\texttt{StarGAN}} + \lambda_{rec}\mathcal{L}_{rec}^{\texttt{StarGAN}} \end{aligned} \tag{6}$$

where $\mathcal{L}_{adv}^{\texttt{StarGAN}}$ is the adversarial loss, $\mathcal{L}_{cls(\text{fake})}^{\texttt{StarGAN}}$, $\mathcal{L}_{cls(\text{real})}^{\texttt{StarGAN}}$ are domain classification loss with respect to fake and real images respectively, $\mathcal{L}_{rec}^{\texttt{StarGAN}}$ is the reconstruction loss. The specific formulations of these loss functions are in Choi et al. (2018). λ_{cls} and λ_{rec} are hyper-parameters that control the relative importance of domain classification and reconstruction losses, respectively, compared to the adversarial loss.

In our experiments, input images are resized to $(256, 256)$ and normalized into the range $[-1, 1]$. The dimension of representation space $\mathcal{Z}$ is set to 512. ω (hyper-parameter that controls accuracy-fairness trade-off) varies from 0 to 10 with step sizes 0.0002 for $\omega \in [0, 0.002]$, 0.002 for $\omega \in [0.002, 0.1]$ and 0.2 for $\omega \in [0.2, 10]$, and γ (hyper-parameter that controls accuracy-invariance trade-off) is set to 0.1 (after hyper-parameter tuning). Models (`FATDM` and baselines) are implemented by PyTorch library version 1.11 and is trained on multiple computer nodes (each model instance is trained on a single node which has 4 CPUs, 8GB of memory, and a single GPU (P100 or V100)). One domain's data is used for testing and the other domains' data is used for training (10% of training data is used for validation). Each model is trained with 10 epoches and the results are from the epoch with best performance on the validation set. Figure 6 visualizes the two-stage training process of `FATDM-StarGAN`. The detailed architectures of `FATDM-StarGAN` are shown in Tables 2-5. We have also provided all code for these models in supplemental material.

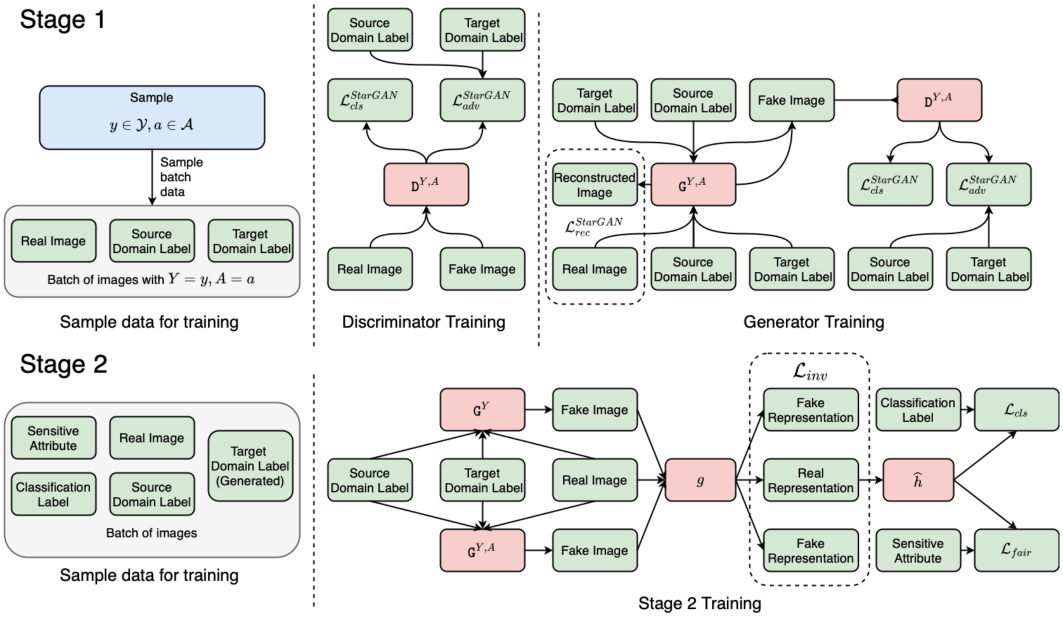


Figure 6: Two-stage training of `FATDM-StarGAN`. For stage 1, we only show the training process for `DensityMatch`^{Y,A} (training process for `DensityMatch`^{Y} is similar.)

Table 2: Architecture of StarGAN generators G^Y and $\mathsf{G}^{Y,A}$ - Density mapping functions $m_{i,j}^{y}$ and $m_{i,j}^{y,a}$ $\forall y \in \mathcal{Y}, a \in \mathcal{A}, i,j \in [N]$. This architecture is similar to the one in the original paper Choi et al. (2018) except for the first convolution layer where number of input channels is 1 (for grayscale images) and input shape is $(h, w, 1 + 2n_c)$. (h, w) is the size of input images, IN is instance batchnorm, and ReLU is Rectified Linear Unit. N: number of output channels, K: kernel size, S: stride szie, P: padding size are convolution and deconvolution layers' hyper-parameters.

Part	Input → Output Shape	Layer Information
Down-sampling	$(h, w, 1 + 2n_c) \to (h, w, 64)$	CONV-(N64, K7x7, S1, P3), IN, ReLU
	$(h, w, 64) \to \left(\frac{h}{2}, \frac{w}{2}, 128\right)$	CONV-(N128, K4x4, S2, P1), IN, ReLU
	$\left(\frac{h}{2}, \frac{w}{2}, 128\right) \to \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	CONV-(N256, K4x4, S2, P1), IN, ReLU
Bottleneck	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \to \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \to \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \to \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \to \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \to \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \to \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
Up-sampling	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \to \left(\frac{h}{2}, \frac{w}{2}, 128\right)$	DECONV-(N128, K4x4, S2, P1), IN, ReLU
	$\left(\frac{h}{2}, \frac{w}{2}, 128\right) \to (h, w, 64)$	DECONV-(N64, K4x4, S2, P1), IN, ReLU
	$(h, w, 64) \to (h, w, 3)$	CONV-(N3, K7x7, S1, P3), IN, ReLU

Table 3: Architecture of StarGAN discriminators. This architecture is similar to the one in the original paper Choi et al. (2018) except for the first convolution layer where number of input channels is 1 (for grayscale images). (h, w) is the size of input images, n_d is the number of domains, and Leaky ReLU is Leaky Rectified Linear Unit. N: number of output channels, K: kernel size, S: stride szie, P: padding size are convolution layers' hyper-parameters.

Layer	Input → Output Shape	Layer Information
Input Layer	$(h, w, 1) \rightarrow \left(\frac{h}{2}, \frac{w}{2}, 64\right)$	CONV-(N64, K4x4, S2, P1), Leaky ReLU
Hidden Layer	$\left(\frac{h}{2}, \frac{w}{2}, 64\right) \rightarrow \left(\frac{h}{4}, \frac{w}{4}, 128\right)$	CONV-(N128, K4x4, S2, P1), Leaky ReLU
Hidden Layer	$\left(\frac{h}{4}, \frac{w}{4}, 128\right) \rightarrow \left(\frac{h}{8}, \frac{w}{8}, 256\right)$	CONV-(N256, K4x4, S2, P1), Leaky ReLU
Hidden Layer	$\left(\frac{h}{8}, \frac{w}{8}, 256\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 512\right)$	CONV-(N512, K4x4, S2, P1), Leaky ReLU
Hidden Layer	$\left(\frac{h}{16}, \frac{w}{16}, 512\right) \rightarrow \left(\frac{h}{32}, \frac{w}{32}, 1024\right)$	CONV-(N1024, K4x4, S2, P1), Leaky ReLU
Hidden Layer	$\left(\frac{h}{32}, \frac{w}{32}, 1024\right) \rightarrow \left(\frac{h}{64}, \frac{w}{64}, 2048\right)$	CONV-(N2048, K4x4, S2, P1), Leaky ReLU
Output Layer (D_{src})	$\left(\frac{h}{64}, \frac{w}{64}, 2048\right) \rightarrow \left(\frac{h}{64}, \frac{w}{64}, 1\right)$	CONV-(N1, K3x3, S1, P1)
Output Layer (D_{cls})	$\left(\frac{h}{64}, \frac{w}{64}, 2048\right) \rightarrow (1, 1, n_d)$	CONV-(N(n_d), K$\frac{h}{64} \times \frac{w}{64}$, S1, P0)

Table 4: Architecture of feature mapping g. This architecture is similar to ResNet18 model He et al. (2016) except for the first convolution layer where number of input channels is 1 (for grayscale images) and the last layer where output dimension is n_z - dimension of representation space $\mathcal{Z}$. (h, w) is the size of input images, BN is batchnorm, MaxPool is max pooling, AvePool is average pooling, and ReLU is Rectified Linear Unit. N: number of output channels, K: kernel size, S: stride szie, P: padding size are convolution layers' hyper-parameters.

Part	Input → Output Shape	Layer Information
Input	$(h, w, 1) \rightarrow \left(\frac{h}{2}, \frac{w}{2}, 64\right)$	CONV-(N64, K7x7, S2, P3), BN, ReLU, MaxPool
Bottleneck	$\left(\frac{h}{2}, \frac{w}{2}, 64\right) \rightarrow \left(\frac{h}{4}, \frac{w}{4}, 64\right)$	Residual Block: CONV-(N64, K3x3, S1, P1), BN, ReLU, CONV-(N64, K3x3, S1, P1), BN
	$\left(\frac{h}{4}, \frac{w}{4}, 64\right) \rightarrow \left(\frac{h}{8}, \frac{w}{8}, 128\right)$	Residual Block: CONV-(N128, K3x3, S1, P1), BN, ReLU, CONV-(N128, K3x3, S1, P1), BN
	$\left(\frac{h}{8}, \frac{w}{8}, 128\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 256\right)$	Residual Block: CONV-(N256, K3x3, S1, P1), BN, ReLU, CONV-(N256, K3x3, S1, P1), BN
	$\left(\frac{h}{16}, \frac{w}{16}, 256\right) \rightarrow (1, 1, 512)$	Residual Block: CONV-(N512, K3x3, S1, P1), IN, ReLU, CONV-(N512, K3x3, S1, P1), BN, AvgPool
Output	$(1, 1, 512) \rightarrow n_z$	LINEAR-(512, n_z)

Table 5: Architecture of classifier $\widehat{h}$. n_z is the dimension of representation space $\mathcal{Z}$.

Layer	Input → Output Shape	Layer Information
Hidden Layer	$n_z \rightarrow \frac{n_z}{2}$	LINEAR-$\left(n_z, \frac{n_z}{2}\right)$, ReLU
Hidden Layer	$\frac{n_z}{2} \rightarrow \frac{n_z}{4}$	LINEAR-$\left(\frac{n_z}{2}, \frac{n_z}{4}\right)$, ReLU
Output Layer	$\frac{n_z}{4} \rightarrow 1$	LINEAR-$\left(\frac{n_z}{4}, 1\right)$, Sigmoid

C ADDITIONAL EXPERIMENTS

Experimental results with all unfairness and error metrics. In this section, we provide more experimental results about fairness and accuracy under domain generalization. In particular, we investigate fairness-accuracy trade-off on the two clinical image datasets including Cardiomegaly and Edema diseases with respect to different fairness criteria (i.e., Equalized Odds, Equal Opportunity), and unfairness (i.e., MD and EMD) and error (i.e., CE, MR, $\overline{\text{AUROC}}$, $\overline{\text{AUPR}}$, $\overline{F_1}$) measures. Figure 7 (Cardiomegaly disease - Equalized Odds), Figure 8 (Cardiomegaly disease - Equal Opportunity), Figure 9 (Edema disease - Equalized Odds), and Figure 10 (Edema disease - Equal Opportunity) show the unfairness-error curves of our models as well as baselines for these two datasets. As we can see, our model outperforms other baselines in terms of fairness-accuracy trade-off. The curve of our model is the bottom-leftmost compared to other baselines in all measures showing the clear benefit of (1) enforcing conditional invariant constraints for accuracy and fairness transfer and (2) using the two-stage training process to stabilize training compared to adversarial learning approach. We also quantify our observations by calculating the areas under these unfairness-error curves, in which the smaller area indicates the better accuracy-fairness trade-off. As shown in Tables 6 and 7, our model has the smallest areas under the curve and achieves significantly better fairness-accuracy trade-off for both equalized odd and equal opportunity compared to other methods.

Impact of the number of source domains. Our work focuses on transferring fairness and accuracy under domain generalization when the target domain data are inaccessible during training. Instead, it relies on a set of source domains to generalize to an unseen, novel target domain. We investigate the relationship between the fairness-accuracy trade-off on the target domain and the number of source domains during training. In particular, we evaluate the performances of `FATDM` and `ERM` on Edema dataset with different numbers of source domains. Similar to the previous experiment, we first construct the dataset for each domain by rotating images with θ degree, where $\theta \in \{0°, 15°, 30°\}$ when the number of domain is 3, $\theta \in \{0°, 15°, 30°, 45°\}$ when the number of domain is 4, and $\theta \in \{0°, 15°, 30°, 45°, 60°\}$ when the number of domain is 5. The number of images per domain is adapted to ensure the training set size is fixed for the three cases. We follow the leave-one-out domain setting in which one domain serves as the unseen target domain for evaluation while the rest domains are for training; the average results across target domains are reported.

Figure 11 shows error-unfairness curves of `FATDM` and `ERM` when training with 2, 3, and 4 source domains. We observe that training with more source domains does not always help the model achieve better fairness-accuracy trade-off on unseen target domains. In particular, the performances of both `FATDM` and `ERM` are the best when training with 2 source domains and the worst when training with 3 source domains. We conjecture the reason that adding more source domains may help reduce the discrepancy between source and target domains (term (ii) in Thm. 1 and Thm. 3), but it may make it more difficult to minimize the source error and unfairness (term (i) in Thm. 1 and Thm. 3) and to learn invariant representation across the source domains (term (iii) in Thm. 1 and Thm. 3). Thus, our suggestion in practice is to conduct an ablation study to find the optimal number of source domains.

Simultaneous and sequential training comparison. In all experiments we conducted so far, the fairness constraint $\mathcal{L}_{fair}$ is optimized simultaneously with the prediction error $\mathcal{L}_{acc}$ and the domain-invariant constraint $\mathcal{L}_{inv}$ for all methods. To investigate whether `FATDM` still attains a better accuracy-fairness trade-off when the processes of invariant representation learning and fair model training are decoupled, we conduct another set of experiments where models (`FATDM` (i.e., `FATDM-StarGAN`) and baselines `G2DM`, `DANN`, `CDANN`) are learned in a sequential matter: for each model, we first learn the representation mapping g by optimizing $\mathcal{L}_{inv}$ and $\mathcal{L}_{acc}$; using the representations generated by the fixed g, we then learn the fair classifier by optimizing $\mathcal{L}_{acc}$ and $\mathcal{L}_{fair}$. The models trained based on the above procedure are named `FATDM-seq`, `G2DM-seq`, `DANN-seq`, and `CDANN-seq`; and their corresponding error-unfairness curves are shown in Figure 12. The results show that `FATDM-seq` still attains the best accuracy-fairness trade-off at target domain compared to `G2DM-seq`, `DANN-seq`, `CDANN-seq`. Our method is effective no matter whether $\mathcal{L}_{fair}$ and $\mathcal{L}_{inv}$ are optimized simultaneously or sequentially.

The reason that our method consistently outperforms the baselines for both settings is that the invariant-representation learning in baseline methods only guarantees the transfer of accuracy but not fairness. Even though a fairness regularizer is imposed to ensure the model is fair at source domains (no matter whether invariant representations and fair classifier are trained simultaneously or sequentially), this fairness cannot be preserved at the target domain due to the potential distributional shifts. The key to ensuring the transfer of fairness is to learn representations such that $P(Z|Y, A)$ is domain-invariant; this must be done during the representation learning process. From Thm 3, we can see that unfairness at target domain $\epsilon^{\mathrm{EO}}_{D^T}$ can still blow up if $P^{Z|Y,A}$ is different across domains, regardless of how fair the model is at source domains (i.e., small $\epsilon^{\mathrm{EO}}_{D^S_i}$).

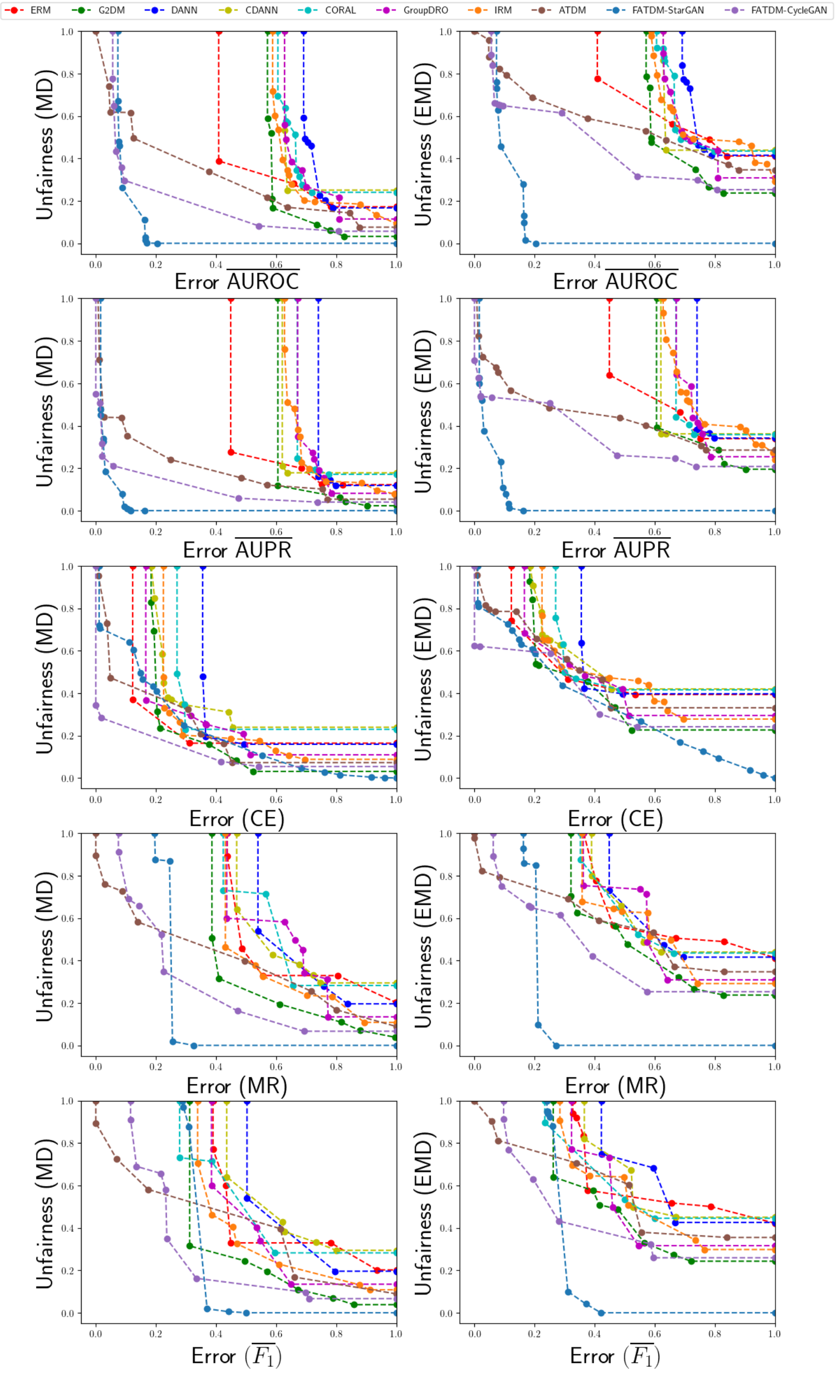


Figure 7: Error-unfairness curves with respect to equalized odds of `FATDM` and baselines on Cardiomegaly disease dataset.

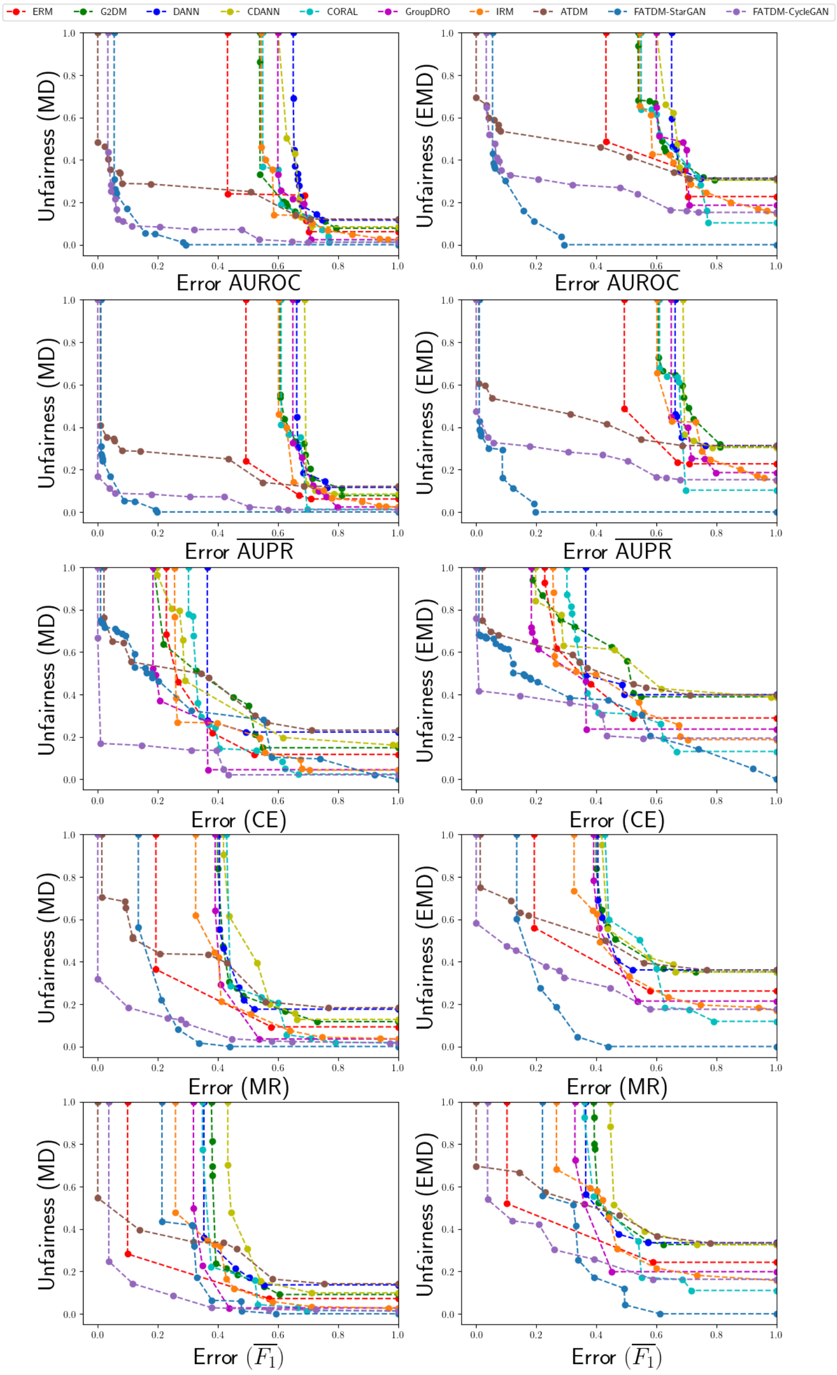


Figure 8: Error-unfairness curves with respect to equal opportunity of `FATDM` and baselines on Cardiomegaly disease dataset.

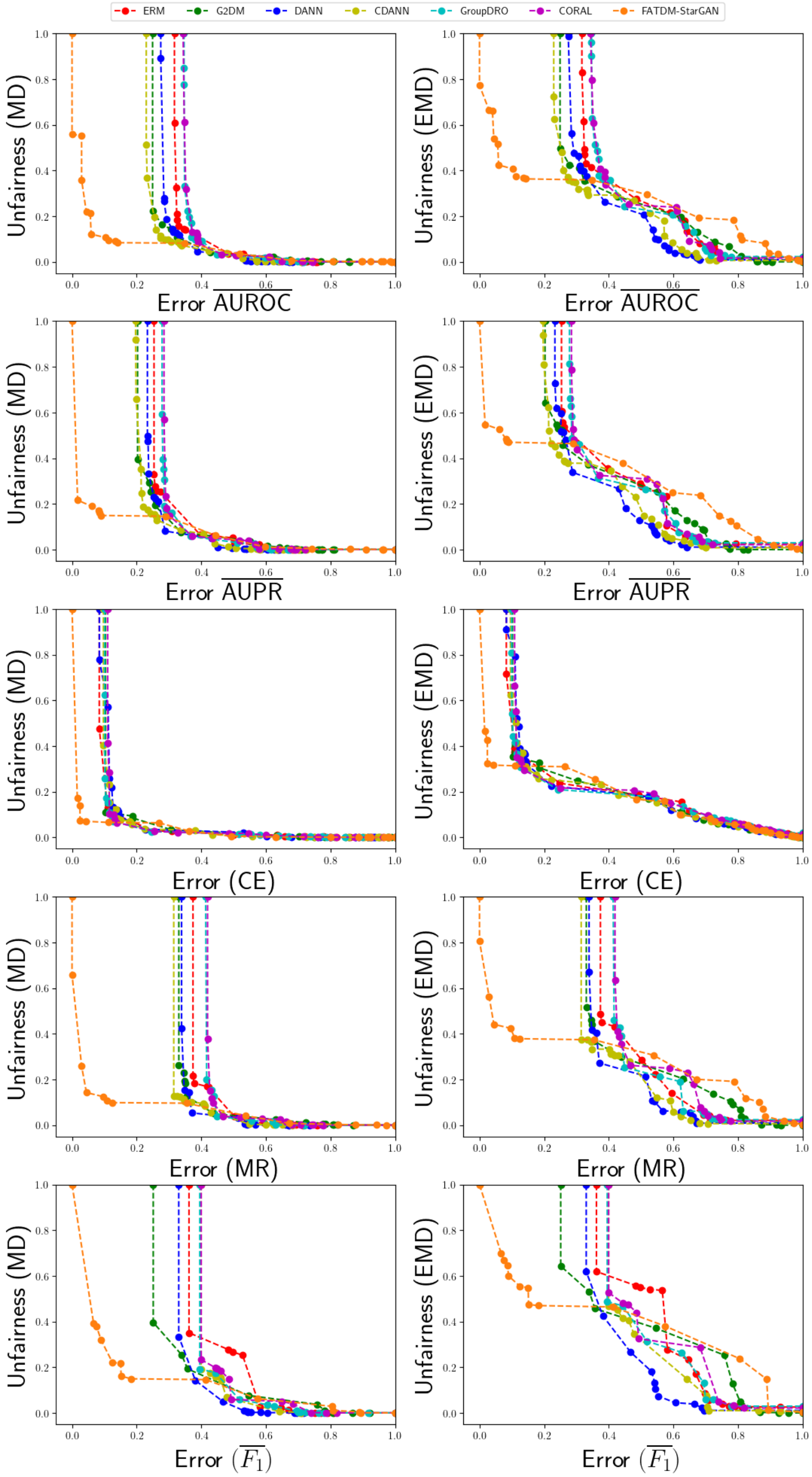


Figure 9: Error-unfairness curves with respect to equalized odds of FATDM and baselines on Edema disease dataset.

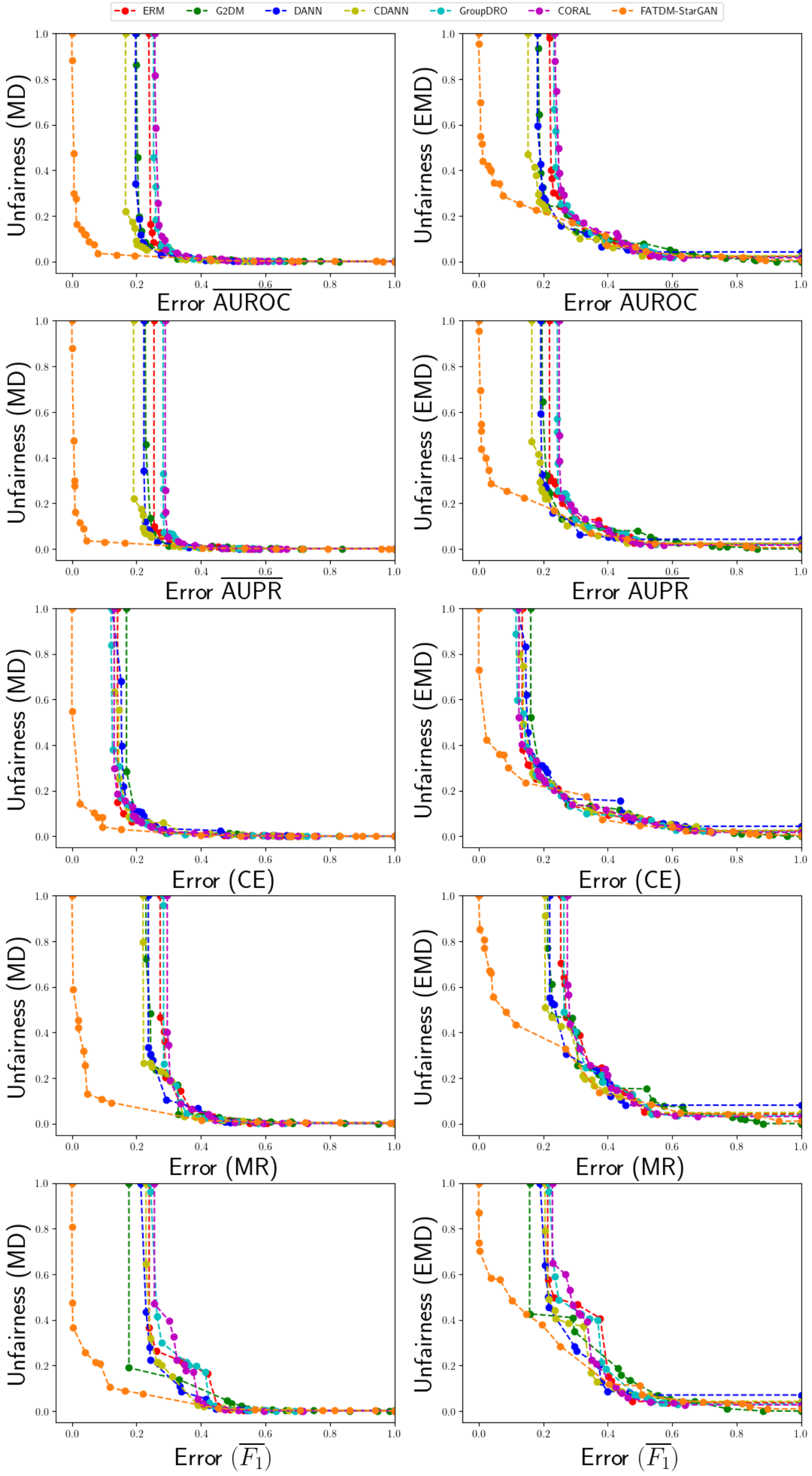


Figure 10: Error-unfairness curves with respect to equal opportunity of FATDM and baselines on Edema disease dataset.

Table 6: Area under the error-unfairness curves (Cardiomegaly disease dataset).

Error - Unfairness		Method							
		`ERM`	`G2DM`	`DANN`	`CDANN`	`CORAL`	`GroupDRO`	`IRM`	`FATDM`
Equalized Odds	$\overline{\text{AUROC}}$ - MD	0.5575	0.6093	0.7571	0.7224	0.7239	0.7039	0.6784	**0.0935**
	$\overline{\text{AUPRC}}$ - MD	0.5463	0.6301	0.7730	0.6883	0.7300	0.7152	0.6967	**0.0291**
	CE - MD	0.2861	0.2601	0.4622	0.4232	0.4424	0.3148	0.3370	**0.2152**
	MR - MD	0.6312	0.4906	0.6795	0.6667	0.6683	0.6382	0.5721	**0.2439**
	$\overline{F_1}$ - MD	0.5901	0.4150	0.6507	0.6547	0.5745	0.5360	0.5025	**0.3365**
	$\overline{\text{AUROC}}$ - EMD	0.7326	0.7106	0.8342	0.7931	0.8075	0.7845	0.7991	**0.1099**
	$\overline{\text{AUPRC}}$ - EMD	0.6901	0.7146	0.8308	0.7577	0.7918	0.7806	0.7945	**0.0437**
	CE - EMD	0.5158	0.4443	0.6143	0.5788	0.5873	0.4911	0.5274	**0.3384**
	MR - EMD	0.7056	0.5795	0.7137	0.6979	0.6902	0.6571	0.6483	**0.2045**
	$\overline{F_1}$ - EMD	0.6866	0.5328	0.7279	0.7019	0.6515	0.6027	0.6120	**0.2888**
Equal Opportunity	$\overline{\text{AUROC}}$ - MD	0.5128	0.6001	0.6999	0.6686	0.5935	0.6288	0.5910	**0.0750**
	$\overline{\text{AUPRC}}$ - MD	0.5419	0.6718	0.7086	0.7189	0.6423	0.6761	0.6435	**0.0262**
	CE - MD	0.3690	0.4272	0.5094	0.4492	0.3780	0.2737	0.3582	**0.2754**
	MR - MD	0.3203	0.5068	0.5252	0.5512	0.4897	0.4368	0.4173	**0.1778**
	$\overline{F_1}$ - MD	0.2134	0.4570	0.4608	0.5207	0.4017	0.3561	0.3510	**0.2737**
	$\overline{\text{AUROC}}$ - EMD	0.6119	0.7184	0.7649	0.7517	0.6720	0.7068	0.6780	**0.0947**
	$\overline{\text{AUPRC}}$ - EMD	0.6321	0.7684	0.7718	0.7877	0.6912	0.7335	0.7200	**0.0448**
	CE - EMD	0.5092	0.6093	0.6264	0.6141	0.4737	0.4340	0.4917	**0.3070**
	MR - EMD	0.4619	0.6420	0.6325	0.6532	0.5790	0.5515	0.5298	**0.1918**
	$\overline{F_1}$ - EMD	0.3876	0.6122	0.5942	0.6496	0.5101	0.4898	0.4889	**0.3108**

Table 7: Area under the error-unfairness curves (Edema disease dataset).

Error - Unfairness		Method						
		`ERM`	`G2DM`	`DANN`	`CDANN`	`CORAL`	`GroupDRO`	`FATDM`
Equalized Odds	$\overline{\text{AUROC}}$ - MD	0.3395	0.2765	0.2972	0.2548	0.3642	0.3627	**0.0633**
	$\overline{\text{AUPRC}}$ - MD	0.2865	0.2446	0.2561	0.2304	0.3052	0.2998	**0.0771**
	CE - MD	0.1096	0.1266	0.1243	0.1192	0.1269	0.1179	**0.0341**
	MR - MD	0.3929	0.3525	0.3509	0.3302	0.4303	0.4240	**0.0656**
	$\overline{F_1}$ - MD	0.4213	0.3219	0.3527	0.4178	0.4283	0.4189	**0.1369**
	$\overline{\text{AUROC}}$ - EMD	0.4277	0.3813	0.3637	0.3419	0.4419	0.4394	**0.2729**
	$\overline{\text{AUPRC}}$ - EMD	0.3868	0.3588	0.3285	0.3245	0.3958	0.3921	**0.3041**
	CE - EMD	0.2366	0.2401	0.2348	0.2334	0.2447	0.2339	**0.1827**
	MR - MD	0.4592	0.4435	0.4017	0.3904	0.4942	0.4792	**0.2802**
	$\overline{F_1}$ - MD	0.5186	0.4642	0.4132	0.4827	0.5180	0.5029	**0.3855**
Equal Opportunity	$\overline{\text{AUROC}}$ - MD	0.2488	0.2139	0.2085	0.1806	0.2696	0.2625	**0.0218**
	$\overline{\text{AUPRC}}$ - MD	0.2606	0.2381	0.2297	0.2035	0.2937	0.2874	**0.0168**
	CE - MD	0.1540	0.1839	0.1689	0.1572	0.1487	0.1446	**0.0234**
	MR - MD	0.2967	0.2652	0.2620	0.2516	0.3101	0.2999	**0.0468**
	$\overline{F_1}$ - MD	0.2848	0.2195	0.2534	0.2613	0.2973	0.2975	**0.0502**
	$\overline{\text{AUROC}}$ - EMD	0.2736	0.2472	0.2449	0.2155	0.2897	0.2841	**0.1121**
	$\overline{\text{AUPRC}}$ - EMD	0.2653	0.2451	0.2429	0.2176	0.2852	0.2812	**0.0912**
	CE - EMD	0.2083	0.2318	0.2355	0.2147	0.2055	0.2003	**0.1159**
	MR - MD	0.3409	0.3162	0.3258	0.3026	0.3442	0.3388	**0.1872**
	$\overline{F_1}$ - MD	0.3237	0.2756	0.3031	0.3008	0.3215	0.3271	**0.1779**

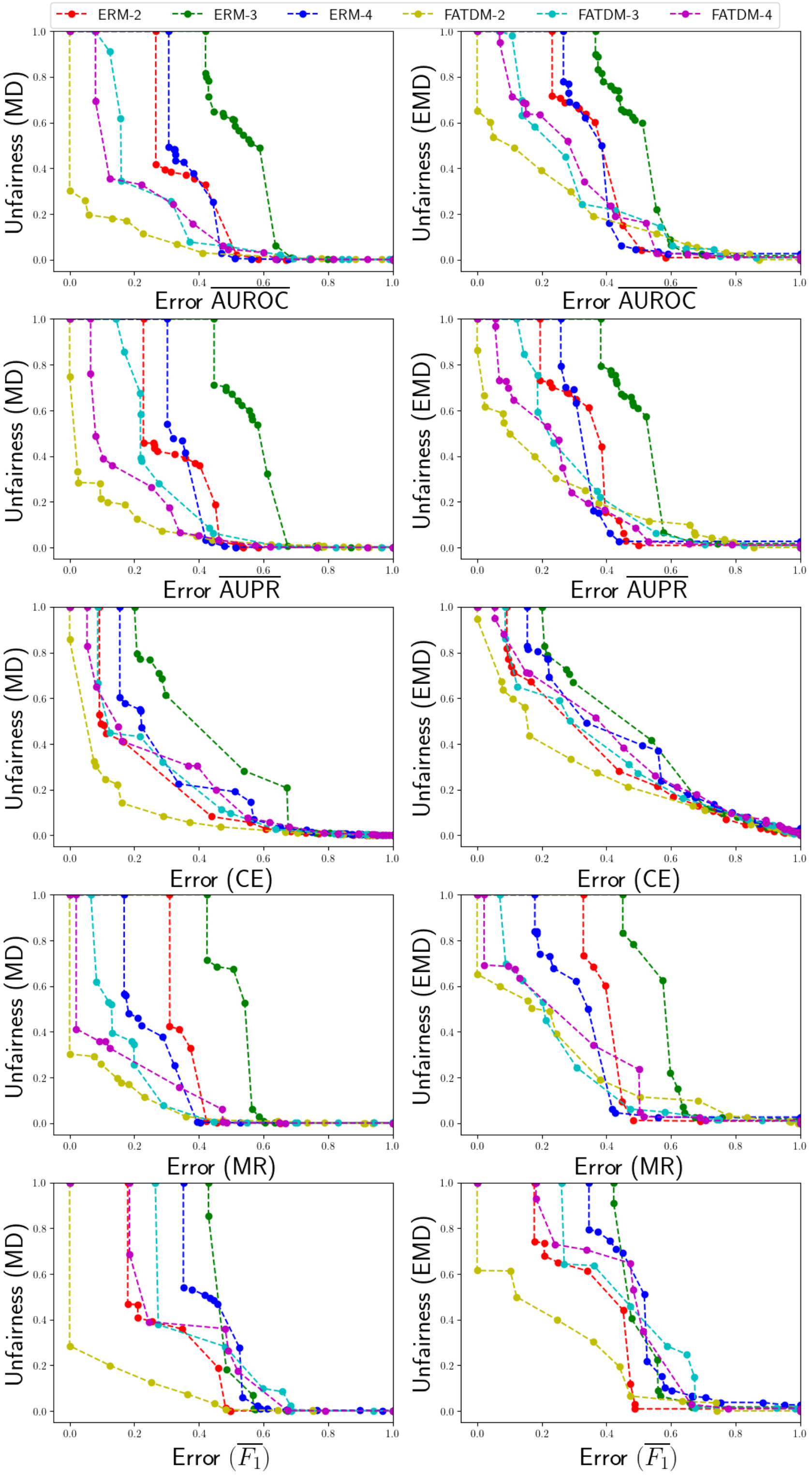
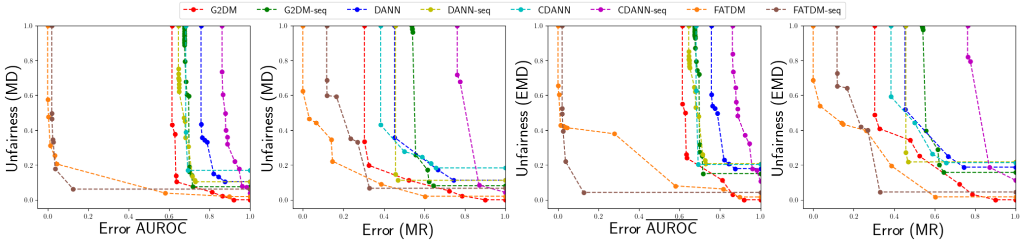
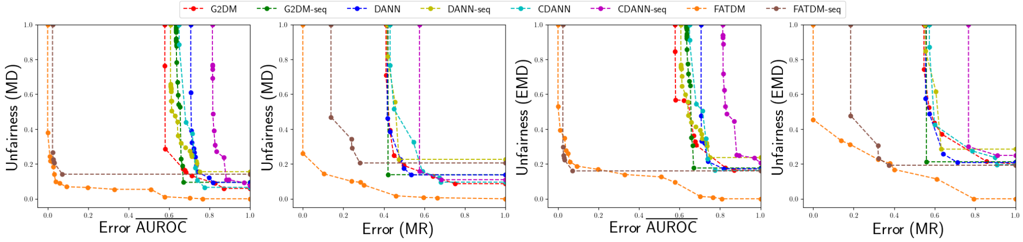


Figure 11: Error-unfairness curves with respect to equalized odds of `FATDM` and `ERM` on Edema disease dataset when training with different numbers of source domains. Names in the figure legend are in the form of `X-Y` where `X` is the model and `Y` is the number of source domains (e.g., `ERM-2` means training `ERM` on two source domains.)



(a) Equalized Odds



(b) Equal Opportunity

Figure 12: Fairness-accuracy trade-off (Pareto frontier) of models trained with simultaneous and sequential (i.e., models with '-seq' suffix) approaches, and `FATDM-CycleGAN` (i.e., use `CycleGAN` instead of `StarGAN` as density mapping functions) on Cardiomegaly disease dataset: error-unfairness curves are constructed by varying $\omega \in [0, 10]$ and the values of error and unfairness are normalized to $[0, 1]$. Lower-left points and the smaller area under the curve indicate the model has a better fairness-accuracy trade-off (Pareto optimality).

D ADDITIONAL RESULTS & LEMMAS

D.1 TIGHTER UPPER BOUND FOR ACCURACY

Corollary 5.1 *We can replace **term (ii)** in Thm. 1 with the following term to attain a tighter upper bound for accuracy:*

$$\sqrt{2}C\min_{i\in[N]}\left(d_{JS}\left(P^Y_{D^T},P^Y_{D^S_i}\right)+\sqrt{2\eta_{TV}\mathbb{E}_{z\sim P_{D^T_i}(z)}\left[d_{JS}\left(P^{X|Y}_{D^T},P^{X|Y}_{D^T_i}\right)^2\right]}\right).$$

where $\eta_{TV} = \sup_{P^X_{D_i}\neq P^X_{D_j}} \frac{\mathcal{D}_{TV}\left(P^Z_{D_i},P^Z_{D_j}\right)}{\mathcal{D}_{TV}\left(P^X_{D_i},P^X_{D_j}\right)} \leq 1$ *is called Dobrushin's coefficient (Polyanskiy & Wu, 2017).*

This result suggests that we can further optimize ***term (ii)*** in Thm. 1 by minimizing η_{TV}. It has been shown in Shui et al. (2022) that η_{TV} can be controlled by Lipschitz constant of the feature mapping $g : \mathcal{X} \to \mathcal{Z}$ when g follows Gaussian distribution. The Lipschitz constant of g, in turn, can be upper bounded by the Frobenius norm of Jacobian matrix with respect to g (Miyato et al., 2018). However, in practice, we found that computing Jacobian matrix of g is computationally expensive when dimension of representation Z is large, and optimizing it together with invariant constraints does not improve the performances of models in our experiments.

D.2 LEMMAS FOR PROVING THEOREM 1

Lemma 6 *Let X be the random variable in domains D_i and D_j, and $\mathcal{E}$ be an event that $P^X_{D_j} \geq P^X_{D_i}$, then we have:*

$$\int_{\mathcal{E}}\left|P^X_{D_j}-P^X_{D_i}\right|dX=\int_{\overline{\mathcal{E}}}\left|P^X_{D_j}-P^X_{D_i}\right|dX=\frac{1}{2}\int\left|P^X_{D_j}-P^X_{D_i}\right|dX$$

where $\overline{\mathcal{E}}$ is the complement of event $\mathcal{E}$.

Lemma 7 *Let X be the random variable in domains D_i and D_j, let $f : \mathcal{X} \to \mathbb{R}_+$ be a non-negative function bounded by C, then we have:*

$$\mathbb{E}_{D_j}[f(X)] - \mathbb{E}_{D_i}[f(X)] \leq \frac{C}{\sqrt{2}} \sqrt{\min \left(\mathcal{D}_{KL}\left(P^X_{D_i} \parallel P^X_{D_j}\right), \mathcal{D}_{KL}\left(P^X_{D_j} \parallel P^X_{D_i}\right) \right)}$$

where $\mathcal{D}_{KL}(\cdot \parallel \cdot)$ is the KL-divergence between two distributions.

Lemma 8 *Suppose loss function $\mathcal{L}$ is upper bounded by C and consider a classifier $\widehat{f} : \mathcal{X} \to \mathcal{Y}$. the expected classification error of $\widehat{f}$ in domain D_j can be upper bounded by its error in domain D_i:*

$$\epsilon^{Acc}_{D_j}\left(\widehat{f}\right) \leq \epsilon^{Acc}_{D_i}\left(\widehat{f}\right) + \sqrt{2} C d_{JS}\left(P^{X,Y}_{D_j}, P^{X,Y}_{D_i}\right)$$

where X, Y are random variables denoting feature and label in domains D_i and D_j.

Lemma 9 *Consider two distributions $P^X_{D_i}$ and $P^X_{D_j}$ over $\mathcal{X}$. Let $P^Z_{D_i}$ and $P^Z_{D_j}$ be the induced distributions over $\mathcal{Z}$ by mapping function $g : \mathcal{X} \to \mathcal{Z}$, then we have:*

$$d_{JS}(P^X_{D_i}, P^X_{D_j}) \geq d_{JS}(P^Z_{D_i}, P^Z_{D_j})$$

Lemma 10 (Phung et al., 2021) *Consider domain D with joint distribution $P^{X,Y}_D$ and labeling function $f_D : \mathcal{X} \to \mathcal{Y}^\Delta$ from feature space to label space. Given mapping function $g : \mathcal{X} \to \mathcal{Z}$ from feature to representation space, we define labeling function $h_D : \mathcal{Z} \to \mathcal{Y}^\Delta$ from representation space to label space as $h_D(Z)_Y = f_D(X)_Y \circ g^{-1}(Z) = \frac{\int_{g^{-1}(Z)} f_D(X)_Y P^X_D dX}{\int_{g^{-1}(Z)} P^X_D dX}$. Similarly, let $\widehat{f}$ be the hypothesis from feature space, then the corresponding hypothesis $\widehat{h}$ from representation space under the mapping function g is computed as $\widehat{h}(Z)_Y = \frac{\int_{g^{-1}(Z)} \widehat{f}(X)_Y P^X_D dX}{\int_{g^{-1}(Z)} P^X_D dX}$. Let $\epsilon^{Acc}_D(\widehat{f}) = \mathbb{E}_D\left[\mathcal{L}(\widehat{f}(X), Y)\right]$ and $\epsilon^{Acc}_D(\widehat{h}) = \mathbb{E}_D\left[\mathcal{L}(\widehat{h}(Z), Y)\right]$ be expected errors defined with respect to feature space and representation space, respectively. We have:*

$$\epsilon^{Acc}_D\left(\widehat{f}\right) = \epsilon^{Acc}_D\left(\widehat{h}\right)$$

D.3 LEMMAS FOR PROVING COROLLARY 1.1

Lemma 11 *Consider two random variables X, Y. Let $P^{X,Y}_{D_i}, P^{X,Y}_{D_j}$ be two joint distributions defined in domains D_i and D_j, respectively. Then, JS-divergence $\mathcal{D}_{JS}\left(P^{X,Y}_{D_i} \parallel P^{X,Y}_{D_j}\right)$ and KL-divergence $\mathcal{D}_{KL}\left(P^{X,Y}_{D_i} \parallel P^{X,Y}_{D_j}\right)$ can be decomposed as follows:*

$$\begin{aligned}
\mathcal{D}_{KL}\left(P^{X,Y}_{D_i} \parallel P^{X,Y}_{D_j}\right) &= \mathcal{D}_{KL}\left(P^Y_{D_i} \parallel P^Y_{D_j}\right) + \mathbb{E}_{D_i}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_j}\right)\right] \\
\mathcal{D}_{JS}\left(P^{X,Y}_{D_i} \parallel P^{X,Y}_{D_j}\right) &\leq \mathcal{D}_{JS}\left(P^Y_{D_i} \parallel P^Y_{D_j}\right) + \mathbb{E}_{D_i}\left[\mathcal{D}_{JS}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_j}\right)\right] \\
&\quad + \mathbb{E}_{D_j}\left[\mathcal{D}_{JS}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_j}\right)\right]
\end{aligned}$$

D.4 LEMMAS FOR PROVING THEOREM 2

Lemma 12 *Under Assumption in Theorem 2, the following holds for any domain D:*

$$\sqrt{\epsilon^{Acc}_D(\widehat{f})} = \sqrt{\mathbb{E}_D[\mathcal{L}(\widehat{f}(X), Y)]} \geq \sqrt{\frac{2c}{|\mathcal{Y}|}} d_{JS}(P^Y_D, P^{\widehat{Y}}_D)^2, \ \forall \widehat{f}$$

where $\widehat{Y}$ is the prediction made by randomized predictor $\widehat{f}$.

D.5 LEMMAS FOR PROVING THEOREM 3

Definition 13 *Given domain D_i with binary random variable A denoting the sensitive attribute, the unfairness measures that evaluate the violation of equalized odd ($\texttt{EO}$) and equal opportunity ($\texttt{EP}$) criteria between sensitive groups of this domain are defined as follows.*

$$\epsilon_{D_i}^{\texttt{EO}}\left(\widehat{f}\right) = \left|R_{D_i}^{0,0}\left(\widehat{f}\right) - R_{D_i}^{0,1}\left(\widehat{f}\right)\right| + \left|R_{D_i}^{1,0}\left(\widehat{f}\right) - R_{D_i}^{1,1}\left(\widehat{f}\right)\right|$$
$$\epsilon_{D_i}^{\texttt{EP}}\left(\widehat{f}\right) = \left|R_{D_i}^{1,0}\left(\widehat{f}\right) - R_{D_i}^{1,1}\left(\widehat{f}\right)\right|$$

where $R_{D_i}^{y,a}\left(\widehat{f}\right) = \mathbb{E}_{D_i}\left[\widehat{f}(X)_1|Y=y, A=a\right]$.

Lemma 14 *Given two domains D_i and D_j, under Definition 13, $R_{D_j}^{y,a}\left(\widehat{f}\right)$ can be bounded by $R_{D_i}^{y,a}\left(\widehat{f}\right)$ as follows.*

$$R_{D_j}^{y,a}\left(\widehat{f}\right) \le R_{D_i}^{y,a}\left(\widehat{f}\right) + \sqrt{2}d_{JS}\left(P_{D_j}^{X|Y=y,A=a}, P_{D_i}^{X|Y=y,A=a}\right) \quad \forall y,a \in \{0,1\}$$

Lemma 15 *Given two domains D_i and D_j, under Definition 13, the unfairness in domain D_j can be upper bounded by the unfairness measure in domain D_i as follows.*

$$\epsilon_{D_j}^{\texttt{EO}}\left(\widehat{f}\right) \le \epsilon_{D_i}^{\texttt{EO}}\left(\widehat{f}\right) + \sqrt{2}\sum_{y=0,1}\sum_{a=0,1} d_{JS}\left(P_{D_j}^{X|Y=y,A=a}, P_{D_i}^{X|Y=y,A=a}\right)$$
$$\epsilon_{D_j}^{\texttt{EP}}\left(\widehat{f}\right) \le \epsilon_{D_i}^{\texttt{EP}}\left(\widehat{f}\right) + \sqrt{2}\sum_{a=0,1} d_{JS}\left(P_{D_j}^{X|Y=1,A=a}, P_{D_i}^{X|Y=1,A=a}\right)$$

Lemma 16 *Consider domain D with distribution $P_D^{X,Y}$ and labeling function $f_D : \mathcal{X} \to \mathcal{Y}^\Delta$. Given mapping function $g : \mathcal{X} \to \mathcal{Z}$ from feature to representation space, we define labeling function $h_D : \mathcal{Z} \to \mathcal{Y}^\Delta$ from representation space to label space as $h_D(Z)_Y = f_D(X)_Y \circ g^{-1}(Z) = \frac{\int_{g^{-1}(Z)} f_D(X)_Y P_D^X dX}{\int_{g^{-1}(Z)} P_D^X dX}$. Similarly, let $\widehat{f}$ be the hypothesis from feature space, then the corresponding hypothesis $\widehat{h}$ from representation space under the mapping function g is computed as $\widehat{h}(Z)_Y = \frac{\int_{g^{-1}(Z)} \widehat{f}(X)_Y P_D^X dX}{\int_{g^{-1}(Z)} P_D^X dX}$. Under Definition 13, we have:*

$$\epsilon_D^{\texttt{EO}}\left(\widehat{f}\right) = \epsilon_D^{\texttt{EO}}\left(\widehat{h}\right)$$
$$\epsilon_D^{\texttt{EP}}\left(\widehat{f}\right) = \epsilon_D^{\texttt{EP}}\left(\widehat{h}\right)$$

D.6 LEMMAS FOR PROVING THEOREM 5

Lemma 17 *Consider two domains D_i and D_j, if there exist invertible mappings $m_{i,j}^y$ and $m_{i,j}^{y,a}$ such that $P_{D_i}^{X|y} = P_{D_j}^{m_{i,j}^y(X)|y}$ and $P_{D_i}^{X|y,a} = P_{D_j}^{m_{i,j}^{y,a}(X)|y,a}$, $\forall y \in \mathcal{Y}, a \in \mathcal{A}$, then $\mathcal{D}_{JS}\left(P_{D_i}^{Z|y} \parallel P_{D_j}^{Z|y}\right)$ and $\mathcal{D}_{JS}\left(P_{D_i}^{Z|y,a} \parallel P_{D_j}^{Z|y,a}\right)$ can be upper bounded by $\int_x P_{D_i}^{x|y}\mathcal{D}_{JS}\left(P^{Z|x} \parallel P^{Z|m_{i,j}^y(x)}\right) dx$ and $\int_x P_{D_i}^{x|y,a}\mathcal{D}_{JS}\left(P^{Z|x} \parallel P^{Z|m_{i,j}^{y,a}(x)}\right) dx$, respectively.*

E PROOFS

E.1 PROOFS OF THEOREMS

Proof of Theorem 1. First, we get the upper bound based on the representation space $\mathcal{Z}$. Then, we relate it with the feature space $\mathcal{X}$. Let $D_*^S \in \{D_i^S\}_{i=1}^N$ be the source domain that's nearest to the target domain D^T. According to Lemma 8, we have upper bound of the expected classification error for the target domain based on each of the source domain as follows.

$$\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{h}\right) \leq \epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right) + \sqrt{2}C d_{JS}\left(P^{Z,Y}_{D^T}, P^{Z,Y}_{D^S_i}\right) \quad \forall i \in [N]$$

Taking average of upper bounds based on all source domains, we have:

$$\begin{aligned}
\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{h}\right) &\leq \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right) + \frac{\sqrt{2}C}{N}\sum_{i=1}^{N} d_{JS}\left(P^{Z,Y}_{D^T}, P^{Z,Y}_{D^S_i}\right)\\
&\overset{(1)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right) + \frac{\sqrt{2}C}{N}\sum_{i=1}^{N} d_{JS}\left(P^{Z,Y}_{D^T}, P^{Z,Y}_{D^S_*}\right) + \frac{\sqrt{2}C}{N}\sum_{i=1}^{N} d_{JS}\left(P^{Z,Y}_{D^S_*}, P^{Z,Y}_{D^S_i}\right)\\
&\overset{(2)}{\leq} \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right) + \sqrt{2}C \min_{i\in[N]} d_{JS}\left(P^{Z,Y}_{D^T}, P^{Z,Y}_{D^S_i}\right) + \sqrt{2}C \max_{i,j\in[N]} d_{JS}\left(P^{Z,Y}_{D^S_i}, P^{Z,Y}_{D^S_j}\right)
\end{aligned} \tag{7}$$

Here we have $\overset{(1)}{\leq}$ by using triangle inequality for JS-distance: $d_{JS}(P,R) \leq d_{JS}(P,Q) + d_{JS}(Q,R)$ with P, Q, and $R = P_{D^T}, P_{D^S_*}$ and $P_{D^S_i}$, respectively. We have $\overset{(2)}{\leq}$ because $D^S_* \in \{D^S_i\}_{i=1}^N$ then $d_{JS}\left(P^{Z,Y}_{D^S_*}, P^{Z,Y}_{D^S_i}\right) \leq \max_{i,j\in[N]} d_{JS}\left(P^{Z,Y}_{D^S_i}, P^{Z,Y}_{D^S_j}\right)$. Similarly, we can obtain the upper bound based on the feature space $\mathcal{X}$ as follows.

$$\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right) \leq \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{f}\right) + \sqrt{2}C \min_{i\in[N]} d_{JS}\left(P^{X,Y}_{D^T}, P^{X,Y}_{D^S_i}\right) + \sqrt{2}C \max_{i,j\in[N]} d_{JS}\left(P^{X,Y}_{D^S_i}, P^{X,Y}_{D^S_j}\right) \tag{8}$$

However, the bounds in Eq. (7) and Eq. (8) are based on either feature space or representation space, which is not readily to use for practical algorithmic design because the actual objective is to minimize $\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right)$ in feature space by controlling Z in representation space. According to Lemmas 9 and 10, we can derive the bound that relates feature and representation spaces as follows.

$$\begin{aligned}
\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right) &= \epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{h}\right)\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right) + \sqrt{2}C \min_{i\in[N]} d_{JS}\left(P^{Z,Y}_{D^T}, P^{Z,Y}_{D^S_i}\right) + \sqrt{2}C \max_{i,j\in[N]} d_{JS}\left(P^{Z,Y}_{D^S_i}, P^{Z,Y}_{D^S_j}\right)\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{f}\right) + \sqrt{2}C \min_{i\in[N]} d_{JS}\left(P^{X,Y}_{D^T}, P^{X,Y}_{D^S_i}\right) + \sqrt{2}C \max_{i,j\in[N]} d_{JS}\left(P^{Z,Y}_{D^S_i}, P^{Z,Y}_{D^S_j}\right)
\end{aligned} \tag{9}$$

Proof of Corollary 1.1.

$$\begin{aligned}
d_{JS}\left(P^{Z,Y}_{D^S_i}, P^{Z,Y}_{D^S_j}\right) &= \sqrt{\mathcal{D}_{JS}\left(P^{Z,Y}_{D^S_i} \parallel P^{Z,Y}_{D^S_j}\right)}\\
&\overset{(1)}{\leq} \sqrt{\mathcal{D}_{JS}\left(P^{Y}_{D^S_i} \parallel P^{Y}_{D^S_j}\right) + 2\mathbb{E}_{z\sim P_{D^S_{i,j}}(z)}\left[\mathcal{D}_{JS}\left(P^{Z|Y}_{D^S_i} \parallel P^{Z|Y}_{D^S_j}\right)\right]}\\
&\overset{(2)}{\leq} d_{JS}\left(P^{Y}_{D^S_i}, P^{Y}_{D^S_j}\right) + \sqrt{2\mathbb{E}_{z\sim P_{D^S_{i,j}}(z)}\left[d_{JS}\left(P^{Z|Y}_{D^S_i}, P^{Z|Y}_{D^S_j}\right)^2\right]}
\end{aligned}$$

Here we have $\overset{(1)}{\leq}$ by using Lemma 11 to decompose the JS-divergence of the joint distributions and $\overset{(2)}{\leq}$ by using inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

This new upper bound, combined with Thm. 1 suggests learning representation Z such that $P^{Z|Y}_{D^S_i}$ is invariant across source domains, or in another word, $Z \perp D \mid Y$. This result is consistent with Thm. 4: when the target domain D^T is the mixture of source domains $\{D^S_i\}_{i=1}^N$, and when $P^Y_{D^S_i}$ and $P^{Z|Y}_{D^S_i}$ are invariant across source domains, we have $d_{JS}\left(P^{Z,Y}_{D^T}, P^{Z,Y}_{D^S_i}\right) = d_{JS}\left(P^{Z,Y}_{D^S_i}, P^{Z,Y}_{D^S_j}\right) = 0$, implying $\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right) \leq \frac{1}{N}\sum_{i=1}^N \epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{f}\right) = \epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{f}\right) \forall i \in [N]$.

Proof of Corollary 5.1 (tighter upper bound for accuracy). The bound in Eq. (9) is constructed using Lemma 9. Indeed, we can make this bound tighter using the strong data processing inequality for JS-divergence (Polyanskiy & Wu, 2017), as stated below.

$$\mathcal{D}_{JS}\left(P^Z_{D_i} \parallel P^Z_{D_j}\right) \leq \eta_{JS}\mathcal{D}_{JS}\left(P^X_{D_i} \parallel P^X_{D_j}\right) \leq \eta_{TV}\mathcal{D}_{JS}\left(P^X_{D_i} \parallel P^X_{D_j}\right)$$

where Z is random variable induced from random variable X, and $P^X_{D_i}$ and $P^X_{D_i}$ are two distribution over $\mathcal{X}$, and $\eta_{JS} = \sup_{P^X_{D_i} \neq P^X_{D_j}} \frac{\mathcal{D}_{JS}\left(P^Z_{D_i}, P^Z_{D_j}\right)}{\mathcal{D}_{JS}\left(P^X_{D_i}, P^X_{D_j}\right)} \leq \eta_{TV} = \sup_{P^X_{D_i} \neq P^X_{D_j}} \frac{\mathcal{D}_{TV}\left(P^Z_{D_i}, P^Z_{D_j}\right)}{\mathcal{D}_{TV}\left(P^X_{D_i}, P^X_{D_j}\right)} \leq 1$, $\mathcal{D}_{TV}$ is the total variation distance. η_{TV} is called the Dobrushin's coefficient (Polyanskiy & Wu, 2017).

Apply Lemma 11 and this inequality to the second term in the right hand side of Eq. (7) (similar to the proof of Corollary 1.1), we have:

$$\begin{aligned}
&\sqrt{2}C \min_{i\in[N]} d_{JS}\left(P^{Z,Y}_{D^T}, P^{Z,Y}_{D^S_i}\right) \\
&\leq \sqrt{2}C \min_{i\in[N]} \left(d_{JS}\left(P^Y_{D^T}, P^Y_{D^S_i}\right) + \sqrt{2\mathbb{E}_{z\sim P_{D^T_i}(z)}\left[d_{JS}\left(P^{Z|Y}_{D^T}, P^{Z|Y}_{D^T_i}\right)^2\right]}\right) \\
&\leq \sqrt{2}C \min_{i\in[N]} \left(d_{JS}\left(P^Y_{D^T}, P^Y_{D^S_i}\right) + \sqrt{2\eta_{TV}\mathbb{E}_{z\sim P_{D^T_i}(z)}\left[d_{JS}\left(P^{X|Y}_{D^T}, P^{X|Y}_{D^T_i}\right)^2\right]}\right) \qquad (10)
\end{aligned}$$

Proof of Theorem 2. Consider a source domain D^S_i and target domain D^T. Because JS-distance $d_{JS}(\cdot,\cdot)$ is a distance metric, we have triangle inequality:

$$d_{JS}(P^Y_{D^S_i}, P^Y_{D^T}) \leq d_{JS}(P^Y_{D^S_i}, P^{\widehat{Y}}_{D^S_i}) + d_{JS}(P^{\widehat{Y}}_{D^S_i}, P^{\widehat{Y}}_{D^T}) + d_{JS}(P^{\widehat{Y}}_{D^T}, P^Y_{D^T})$$

Since $X \xrightarrow{g} Z \xrightarrow{\widehat{h}} \widehat{Y}$, we have $d_{JS}(P^{\widehat{Y}}_{D^S_i}, P^{\widehat{Y}}_{D^T}) \leq d_{JS}(P^Z_{D^S_i}, P^Z_{D^T})$. Using Lemma 12, the following holds when $d_{JS}(P^Y_{D^S_i}, P^Y_{D^T}) \geq d_{JS}(P^Z_{D^S_i}, P^Z_{D^T})$

$$\begin{aligned}
\Big(d_{JS}(P^Y_{D^S_i}, P^Y_{D^T}) - d_{JS}(P^Z_{D^S_i}, P^Z_{D^T})\Big)^2 &\leq \Big(d_{JS}(P^Y_{D^S_i}, P^{\widehat{Y}}_{D^S_i}) + d_{JS}(P^{\widehat{Y}}_{D^T}, P^Y_{D^T})\Big)^2 \\
&\leq 2\Big(d_{JS}(P^Y_{D^S_i}, P^{\widehat{Y}}_{D^S_i})^2 + d_{JS}(P^{\widehat{Y}}_{D^T}, P^Y_{D^T})^2\Big) \\
&\leq \frac{2}{\sqrt{\frac{2c}{|\mathcal{Y}|}}}\left(\sqrt{\epsilon^{\texttt{Acc}}_{D^S_i}(\widehat{f})} + \sqrt{\epsilon^{\texttt{Acc}}_{D^T}(\widehat{f})}\right) \\
&\leq \sqrt{\frac{4|\mathcal{Y}|}{c}\left(\epsilon^{\texttt{Acc}}_{D^S_i}(\widehat{f}) + \epsilon^{\texttt{Acc}}_{D^T}(\widehat{f})\right)}
\end{aligned}$$

The last inequality is by AM-GM inequality.

Therefore, when $d_{JS}(P^Y_{D^S_i}, P^Y_{D^T}) \geq d_{JS}(P^Z_{D^S_i}, P^Z_{D^T})$, we have

$$\epsilon^{\texttt{Acc}}_{D^S_i}(\widehat{f}) + \epsilon^{\texttt{Acc}}_{D^T}(\widehat{f}) \geq \frac{c}{4|\mathcal{Y}|}\Big(d_{JS}(P^Y_{D^S_i}, P^Y_{D^T}) - d_{JS}(P^Z_{D^S_i}, P^Z_{D^T})\Big)^4$$

The above holds for any source domain D^S_i. Average over all N source domains, we have

$$\frac{1}{N}\sum_{i=1}^N \epsilon^{\texttt{Acc}}_{D^S_i}(\widehat{f}) + \epsilon^{\texttt{Acc}}_{D^T}(\widehat{f}) \geq \frac{c}{4|\mathcal{Y}|N}\sum_{i=1}^N \Big(d_{JS}(P^Y_{D^S_i}, P^Y_{D^T}) - d_{JS}(P^Z_{D^S_i}, P^Z_{D^T})\Big)^4$$

Proof of Theorem 3. The proof is based on Lemmas 15 and 16 and similar to the proof of Thm. 1. Let $D^S_* \in \{D^S_i\}_{i=1}^N$ be the source domain nearest to the target domain D^T. According to Lemma 15, we have upper bound of the unfairness measured with respect to the representation space for the target domain based on each of the source domain. For equal opportunity `(EP)`, we have:

$$\epsilon^{\mathrm{EP}}_{D^T}\left(\widehat{h}\right) \le \epsilon^{\mathrm{EP}}_{D^S_i}\left(\widehat{h}\right) + \sqrt{2}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^T}, P^{Z|Y=1,A=a}_{D^S_i}\right)$$

Taking average of upper bounds based on all source domains, we have:

$$\begin{aligned}
\epsilon^{\mathrm{EP}}_{D^T}\left(\widehat{h}\right) &\le \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\mathrm{EP}}_{D^S_i}\left(\widehat{h}\right) + \frac{\sqrt{2}}{N}\sum_{i=1}^{N}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^T}, P^{Z|Y=1,A=a}_{D^S_i}\right)\\
&\le \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\mathrm{EP}}_{D^S_i}\left(\widehat{h}\right) + \frac{\sqrt{2}}{N}\sum_{i=1}^{N}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^T}, P^{Z|Y=1,A=a}_{D^S_*}\right)\\
&+ \frac{\sqrt{2}}{N}\sum_{i=1}^{N}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^S_*}, P^{Z|Y=1,A=a}_{D^S_i}\right)\\
&\le \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\mathrm{EP}}_{D^S_i}\left(\widehat{h}\right) + \sqrt{2}\min_{i\in[N]}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^T}, P^{Z|Y=1,A=a}_{D^S_i}\right)\\
&+ \sqrt{2}\max_{i,j\in[N]}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^S_i}, P^{Z|Y=1,A=a}_{D^S_j}\right)
\end{aligned}$$

According to Lemmas 9 and 16. we can relate this bound to the feature space as follows.

$$\begin{aligned}
\epsilon^{\mathrm{EP}}_{D^T}\left(\widehat{f}\right) &= \epsilon^{\mathrm{EP}}_{D^T}\left(\widehat{h}\right)\\
&\le \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\mathrm{EP}}_{D^S_i}\left(\widehat{h}\right) + \sqrt{2}\min_{i\in[N]}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^T}, P^{Z|Y=1,A=a}_{D^S_i}\right)\\
&+ \sqrt{2}\max_{i,j\in[N]}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^S_i}, P^{Z|Y=1,A=a}_{D^S_j}\right)\\
&\le \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\mathrm{EP}}_{D^S_i}\left(\widehat{f}\right) + \sqrt{2}\min_{i\in[N]}\sum_{a\in\{0,1\}} d_{JS}\left(P^{X|Y=1,A=a}_{D^T}, P^{X|Y=1,A=a}_{D^S_i}\right)\\
&+ \sqrt{2}\max_{i,j\in[N]}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=1,A=a}_{D^S_i}, P^{Z|Y=1,A=a}_{D^S_j}\right)
\end{aligned}$$

Similarly, we got the upper bound for unfairness measure with respect to equalized odds as follows.

$$\begin{aligned}
\epsilon^{\mathrm{EO}}_{D^T}\left(\widehat{f}\right) \le \frac{1}{N}\sum_{i=1}^{N}\epsilon^{\mathrm{EO}}_{D^S_i}\left(\widehat{f}\right) &+ \sqrt{2}\min_{i\in[N]}\sum_{y\in\{0,1\}}\sum_{a\in\{0,1\}} d_{JS}\left(P^{X|Y=y,A=a}_{D^T}, P^{X|Y=y,A=a}_{D^S_i}\right)\\
&+ \sqrt{2}\max_{i,j\in[N]}\sum_{y\in\{0,1\}}\sum_{a\in\{0,1\}} d_{JS}\left(P^{Z|Y=y,A=a}_{D^S_i}, P^{Z|Y=y,A=a}_{D^S_j}\right) \qquad (11)
\end{aligned}$$

Proof of Theorem 4. Consider two source domains, D^S_i and D^S_j, if $P^Y_{D^S_i} = P^Y_{D^S_j}$, we can learn the mapping function $g = P_\theta\left(Z|X\right)$ such that $P^{Z|Y}_{D^S_i} = P^{Z|Y}_{D^S_j}$. Note that this mapping function always exists. In particular, the trivial solution for Z that satisfies $P^{Z|Y}_{D^S_i} = P^{Z|Y}_{D^S_j}$ is making $Z \perp Y, D$ (e.g.,

$P_\theta\left(Z|X\right)=\mathcal{N}\left(\mathbf{0},\mathbf{I}\right)$). Then we have:

$$\begin{aligned}
\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right) &= \mathbb{E}_{z\sim P^Z_{D^S_i},y\sim h_{D^S_i}(z)}\left[\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\right]\\
&= \mathbb{E}_{y\sim P^Y_{D^S_i},z\sim P^{Z|Y}_{D^S_i}}\left[\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\right]\\
&= \mathbb{E}_{y\sim P^Y_{D^S_j},z\sim P^{Z|Y}_{D^S_j}}\left[\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\right]\\
&= \mathbb{E}_{z\sim P^Z_{D^S_j},y\sim h_{D^S_j}(z)}\left[\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\right]\\
&= \epsilon^{\texttt{Acc}}_{D^S_j}\left(\widehat{h}\right)
\end{aligned}$$

For unseen target domain D^T in Λ, we have:

$$\begin{aligned}
\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{h}\right) &= \mathbb{E}_{D^T}\left[\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\right]\\
&= \int_{\mathcal{Z}\times\mathcal{Y}}\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)P^{Y,Z}_{D^T}\,dY\,dZ\\
&= \int_{\mathcal{Z}\times\mathcal{Y}}\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\sum_{i=1}^{N}\pi_i P^{Y,Z}_{D^S_i}\,dY\,dZ\\
&= \sum_{i=1}^{N}\pi_i\int_{\mathcal{Z}\times\mathcal{Y}}\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)P^{Y,Z}_{D^S_i}\,dY\,dZ\\
&= \sum_{i=1}^{N}\pi_i\mathbb{E}_{D^S_i}\left[\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\right]\\
&= \mathbb{E}_{D^S_i}\left[\mathcal{L}\left(\widehat{h}\left(Z\right),Y\right)\right]\quad\forall i\in[N]\\
&= \epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right)\quad\forall i\in[N]
\end{aligned}$$

By Lemma 10, we have $\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{h}\right)=\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{h}\right)=\epsilon^{\texttt{Acc}}_{D^T}\left(\widehat{f}\right)=\epsilon^{\texttt{Acc}}_{D^S_i}\left(\widehat{f}\right)$.

For fairness, we only give the proof for equalized odds (`EO`), we can easily get the similar derivation for equal opportunity. For any Z that satisfies $P^{Z|Y=y,A=a}_{D^S_i}=P^{Z|Y=y,A=a}_{D^S_j}\ \forall y,a\in\{0,1\}$, we have:

$$\begin{aligned}
\epsilon^{\texttt{EO}}_{D^S_i}\left(\widehat{h}\right) &= \sum_{y\in\{0,1\}}\mathcal{D}\left(P^{\widehat{h}(Z)_1|Y=y,A=0}_{D^S_i}\parallel P^{\widehat{h}(Z)_1|Y=y,A=1}_{D^S_i}\right)\\
&= \sum_{y\in\{0,1\}}\mathcal{D}\left(P^{\widehat{h}(Z)_1|Y=y,A=0}_{D^S_j}\parallel P^{\widehat{h}(Z)_1|Y=y,A=1}_{D^S_j}\right)\\
&= \epsilon^{\texttt{EO}}_{D^S_j}\left(\widehat{h}\right)
\end{aligned}$$

For unseen target domain D^T in Λ, we have:

$$\begin{aligned}
\epsilon^{\texttt{EO}}_{D^T}\left(\widehat{h}\right) &= \sum_{y\in\{0,1\}}\mathcal{D}\left(P^{\widehat{h}(Z)_1|Y=y,A=0}_{D^T}\parallel P^{\widehat{h}(Z)_1|Y=y,A=1}_{D^T}\right)\\
&= \sum_{y\in\{0,1\}}\mathcal{D}\left(\sum_{i=1}^{N}\pi_i P^{\widehat{h}(Z)_1|Y=y,A=0}_{D^S_i}\parallel \sum_{i=1}^{N}\pi_i P^{\widehat{h}(Z)_1|Y=y,A=1}_{D^S_i}\right)\\
&= \sum_{y\in\{0,1\}}\mathcal{D}\left(P^{\widehat{h}(Z)_1|Y=y,A=0}_{D^S_i}\parallel P^{\widehat{h}(Z)_1|Y=y,A=1}_{D^S_i}\right)\quad\forall i\in[N]\\
&= \epsilon^{\texttt{EO}}_{D^S_i}\left(\widehat{h}\right)\quad\forall i\in[N]
\end{aligned}$$

Similar to the proof of accuracy, Z that satisfies $P_{D_i^S}^{Z|Y=y,A=a} = P_{D_j^S}^{Z|Y=y,A=a} \ \ \forall y, a \in \{0,1\}, i, j \in [N]$ always exists. The trivial solution for is Z that satisfies $Z \perp Y, A, D$.

By Lemma 16, we have $\epsilon_{D^T}^{\mathtt{EO}}\left(\widehat{h}\right) = \epsilon_{D_i^S}^{\mathtt{EO}}\left(\widehat{h}\right) = \epsilon_{D^T}^{\mathtt{EO}}\left(\widehat{f}\right) = \epsilon_{D_i^S}^{\mathtt{EO}}\left(\widehat{f}\right)$.

For equal opportunity (`EP`), Z only need to satisfy the condition for positive label, i.e., $P_{D_i^S}^{Z|Y=1,A=a} = P_{D_j^S}^{Z|Y=1,A=a} \ \forall a \in \{0,1\}, i, j \in [N]$.

Proof of Theorem 5. According to Lemma 17, we have:

$$\mathcal{D}_{JS}\left(P_{D_i}^{Z|y} \parallel P_{D_j}^{Z|y}\right) \leq \int_x P_{D_j}^{x|y} \mathcal{D}_{JS}\left(P_{D_i}^{Z|x} \parallel P_{D_i}^{Z|m_{i,j}^y(x)}\right) dx \tag{12}$$

Then, minimizing $D_{JS}\left(P_{D_i}^{Z|y} \parallel P_{D_j}^{Z|y}\right)$ can be achieved by minimizing $\mathcal{D}_{JS}\left(P^{Z|x} \parallel P^{Z|m_{i,j}^y(x)}\right) \ \ \forall x \in \mathcal{X}$. We can upper bound $\mathcal{D}_{JS}\left(P^{Z|x} \parallel P^{Z|m_{i,j}^y(x)}\right)$ as follows

$$\begin{aligned}
\mathcal{D}_{JS}\left(P^{Z|x} \parallel P^{Z|m_{i,j}^y(x)}\right) &\leq \mathcal{D}_{TV}\left(P^{Z|x} \parallel P^{Z|m_{i,j}^y(x)}\right) \\
&\leq \sqrt{2}\, d_{1/2}\left(P^{Z|x}, P^{Z|m_{i,j}^y(x)}\right) \\
&\overset{(1)}{=} \sqrt{2}\, d_{1/2}\left(\mathcal{N}\left(\mu(x); \sigma^2 \mathbf{I}_d\right), \mathcal{N}\left(\mu\left(m_{i,j}^y(x)\right); \sigma^2 \mathbf{I}_d\right)\right)
\end{aligned} \tag{13}$$

where $\mathcal{D}_{TV}$ and $d_{1/2}$ are total variation distance and Hellinger distance between two distributions, respectively. We have $\overset{(1)}{=}$ because of our choice for representation mapping $g(x) := P^{Z|x} = \mathcal{N}\left(\mu(x); \sigma^2 \mathbf{I}_d\right)$. According to Devroye et al. (2018), the Hellinger distance between two multivariate normal distributions over $\mathbb{R}^d$ has a closed form as follows

$$\begin{aligned}
&d_{1/2}\left(\mathcal{N}\left(\mu_1; \Sigma_1\right), \mathcal{N}\left(\mu_2; \Sigma_2\right)\right) \\
&= \sqrt{1 - \frac{det\left(\Sigma_1\right)^{1/4} det\left(\Sigma_2\right)^{1/4}}{det\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{1/2}} \exp\left(-\frac{1}{8}\left(\mu_1 - \mu_2\right)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \left(\mu_1 + \mu_2\right)\right)}
\end{aligned} \tag{14}$$

where $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ are mean vectors and covariance matrices of the two normal distributions. In Eq. (14), let $\mu_1 = \mu(x)$, $\mu_2 = \mu\left(m_{i,j}^y(x)\right)$, $\Sigma_1 = \Sigma_2 = \sigma^2 \mathbf{I}_d$, then we have:

$$\begin{aligned}
&d_{1/2}\left(\mathcal{N}\left(\mu(x); \sigma^2 \mathbf{I}_d\right), \mathcal{N}\left(\mu\left(m_{i,j}^y(x)\right); \sigma^2 \mathbf{I}_d\right)\right) \\
&= \sqrt{1 - \exp\left(-\frac{1}{8d\sigma^2}\left(\mu\left(x\right) - \mu\left(m_{i,j}^y(x)\right)\right)^T \left(\mu\left(x\right) - \mu\left(m_{i,j}^y(x)\right)\right)\right)} \\
&= \sqrt{1 - \exp\left(-\frac{1}{8d\sigma^2}\left\|\mu\left(x\right) - \mu\left(m_{i,j}^y(x)\right)\right\|_2^2\right)}
\end{aligned} \tag{15}$$

From Eq. (15), we can see that Helinger distance between two representation distributions $P^{Z|x}$ and $P^{Z|m_{i,j}^y(x)}$ is the function of their means $\mu\left(x\right)$ and $\mu\left(m_{i,j}^y(x)\right)$. Combining this with Eq. (12) and Eq. (13), we conclude that minimizing $d_{JS}\left(P_{D_i^S}^{Z|y}, P_{D_j^S}^{Z|y}\right)$ can be reduced to minimizing $\left\|\mu(x) - \mu\left(m_{i,j}^y(x)\right)\right\|_2$ which can be implemented as the mean square error between $\mu(x)$ and $\mu\left(m_{i,j}^y(x)\right)$ in practice. Proof for $d_{JS}\left(P_{D_i^S}^{Z|y,a}, P_{D_j^S}^{Z|y,a}\right)$ is derived in the similar way.

E.2 PROOFS OF LEMMAS

Proof of Lemma 6. We have:

$$
\begin{aligned}
\int_{\mathcal{E}} \left|P^X_{D_j} - P^X_{D_i}\right| dX &= \int_{\mathcal{E}} \left(P^X_{D_j} - P^X_{D_i}\right) dX \\
&= \int_{\mathcal{E}\cup\overline{\mathcal{E}}} \left(P^X_{D_j} - P^X_{D_i}\right) dX - \int_{\overline{\mathcal{E}}} \left(P^X_{D_j} - P^X_{D_i}\right) dX \\
&= \int_{\overline{\mathcal{E}}} \left(P^X_{D_i} - P^X_{D_j}\right) dX \\
&= \int_{\overline{\mathcal{E}}} \left|P^X_{D_j} - P^X_{D_i}\right| dX \\
&= \frac{1}{2} \int \left|P^X_{D_j} - P^X_{D_i}\right| dX
\end{aligned}
$$

Proof of Lemma 7. We have:

$$
\begin{aligned}
\mathbb{E}_{D_j}\left[f(X)\right] &= \int_{\mathcal{X}} f(X) P^X_{D_j} dX = \int_{\mathcal{X}} f(X) P^X_{D_i} dX + \int_{\mathcal{X}} f(X) \left(P^X_{D_j} - P^X_{D_i}\right) dX \\
&= \mathbb{E}_{D_i}\left[f(X)\right] + \int_{\mathcal{X}} f(X) \left(P^X_{D_j} - P^X_{D_i}\right) dX \\
&= \mathbb{E}_{D_i}\left[f(X)\right] + \int_{\mathcal{E}} f(X) \left(P^X_{D_j} - P^X_{D_i}\right) dX + \int_{\overline{\mathcal{E}}} f(X) \left(P^X_{D_j} - P^X_{D_i}\right) dX \\
&\overset{(1)}{\leq} \mathbb{E}_{D_i}\left[f(X)\right] + \int_{\mathcal{E}} f(X) \left(P^X_{D_j} - P^X_{D_i}\right) dX \\
&\overset{(2)}{\leq} \mathbb{E}_{D_i}\left[f(X)\right] + C \int_{\mathcal{E}} \left(P^X_{D_j} - P^X_{D_i}\right) dX \\
&= \mathbb{E}_{D_i}\left[f(X)\right] + C \int_{\mathcal{E}} \left|P^X_{D_j} - P^X_{D_i}\right| dX \\
&\overset{(3)}{\leq} \mathbb{E}_{D_i}\left[f(X)\right] + \frac{C}{2} \int \left|P^X_{D_j} - P^X_{D_i}\right| dX \\
&\overset{(4)}{\leq} \mathbb{E}_{D_i}\left[f(X)\right] + \frac{C}{2} \sqrt{2 \min\left(\mathcal{D}_{KL}\left(P^X_{D_i} \parallel P^X_{D_j}\right), \mathcal{D}_{KL}\left(P^X_{D_j} \parallel P^X_{D_i}\right)\right)} \\
&= \mathbb{E}_{D_i}\left[f(X)\right] + \frac{C}{\sqrt{2}} \sqrt{\min\left(\mathcal{D}_{KL}\left(P^X_{D_i} \parallel P^X_{D_j}\right), \mathcal{D}_{KL}\left(P^X_{D_j} \parallel P^X_{D_i}\right)\right)}
\end{aligned}
$$

where $\mathcal{E}$ is the event that $P^X_{D_j} \geq P^X_{D_i}$ and $\overline{\mathcal{E}}$ is the complement of $\mathcal{E}$. We have $\overset{(1)}{\leq}$ because $\int_{\overline{\mathcal{E}}} f(X) \left(P^X_{D_j} - P^X_{D_i}\right) dX \leq 0$; $\overset{(2)}{\leq}$ because $f(X)$ is non-negative function and is bounded by C; $\overset{(3)}{\leq}$ by using Lemma 6; $\overset{(4)}{\leq}$ by using Pinsker's inequality between total variation norm and KL-divergence.

Proof of Lemma 8. Applying Lemma 7 and replacing X by (X, Y), f by loss function $\mathcal{L}$, D_i by $D_{i,j}$, we have:

$$
\begin{aligned}
\epsilon^{\mathtt{Acc}}_{D_j}\left(\widehat{f}\right) - \mathbb{E}_{D_{i,j}}\left[\mathcal{L}(\widehat{f}(X), Y)\right] &= \mathbb{E}_{D_j}\left[\mathcal{L}(\widehat{f}(X), Y)\right] - \mathbb{E}_{D_{i,j}}\left[\mathcal{L}(\widehat{f}(X), Y)\right] \\
&\leq \frac{C}{\sqrt{2}} \sqrt{\min\left(\mathcal{D}_{KL}\left(P^{X,Y}_{D_j} \parallel P^{X,Y}_{D_{i,j}}\right), \mathcal{D}_{KL}\left(P^{X,Y}_{D_{i,j}} \parallel P^{X,Y}_{D_j}\right)\right)} \\
&\leq \frac{C}{\sqrt{2}} \sqrt{\mathcal{D}_{KL}\left(P^{X,Y}_{D_j} \parallel P^{X,Y}_{D_{i,j}}\right)} \qquad (16)
\end{aligned}
$$

Applying Lemma 7 again and replacing X by (X,Y), f by loss function $\mathcal{L}$, D_j by $D_{i,j}$, we have:

$$\begin{aligned}
\mathbb{E}_{D_{i,j}}\left[\mathcal{L}(\widehat{f}(X),Y)\right]-\epsilon^{\text{Acc}}_{D_i}\left(\widehat{f}\right) &= \mathbb{E}_{D_{i,j}}\left[\mathcal{L}(\widehat{f}(X),Y)\right]-\mathbb{E}_{D_i}\left[\mathcal{L}(\widehat{f}(X),Y)\right] \\
&\le \frac{C}{\sqrt{2}}\sqrt{\min\left(\mathcal{D}_{KL}\left(P^{X,Y}_{D_i}\parallel P^{X,Y}_{D_{i,j}}\right),\mathcal{D}_{KL}\left(P^{X,Y}_{D_{i,j}}\parallel P^{X,Y}_{D_i}\right)\right)} \\
&\le \frac{C}{\sqrt{2}}\sqrt{\mathcal{D}_{KL}\left(P^{X,Y}_{D_i}\parallel P^{X,Y}_{D_{i,j}}\right)} \qquad (17)
\end{aligned}$$

Adding Eq. (16) to Eq. (17), we have:

$$\begin{aligned}
\epsilon^{\text{Acc}}_{D_j}\left(\widehat{f}\right)-\epsilon^{\text{Acc}}_{D_i}\left(\widehat{f}\right) &\le \frac{C}{\sqrt{2}}\left(\sqrt{\mathcal{D}_{KL}\left(P^{X,Y}_{D_i}\parallel P^{X,Y}_{D_{i,j}}\right)}+\sqrt{\mathcal{D}_{KL}\left(P^{X,Y}_{D_j}\parallel P^{X,Y}_{D_{i,j}}\right)}\right) \\
&\overset{(1)}{\le} \frac{C}{\sqrt{2}}\sqrt{2\left(\mathcal{D}_{KL}\left(P^{X,Y}_{D_i}\parallel P^{X,Y}_{D_{i,j}}\right)+\mathcal{D}_{KL}\left(P^{X,Y}_{D_j}\parallel P^{X,Y}_{D_{i,j}}\right)\right)} \\
&= \frac{C}{\sqrt{2}}\sqrt{4\mathcal{D}_{JS}\left(P^{X,Y}_{D_i}\parallel P^{X,Y}_{D_j}\right)} \\
&= \sqrt{2}C d_{JS}\left(P^{X,Y}_{D_i},P^{X,Y}_{D_j}\right)
\end{aligned}$$

Here we have $\overset{(1)}{\le}$ by using Cauchy–Schwarz inequality.

Proof of Lemma 9. Note that the JS-divergence $\mathcal{D}_{JS}\left(P^X_{D_i}\parallel P^X_{D_j}\right)$ can be understood as the mutual information between a random variable X associated with the mixture distribution $P^X_{D_{i,j}}=\frac{1}{2}\left(P^X_{D_i}+P^X_{D_j}\right)$ and the equiprobable binary random variable T used to switch between $P^X_{D_i}$ and $P^X_{D_j}$ to create the mixture distribution $P^X_{D_{i,j}}$. In particular, we have:

$$\begin{aligned}
\mathcal{D}_{JS}\left(P^X_{D_i}\parallel P^X_{D_j}\right) &= \frac{1}{2}\left(\mathcal{D}_{KL}\left(P^X_{D_i}\parallel P^X_{D_{i,j}}\right)+\mathcal{D}_{JS}\left(P^X_{D_j}\parallel P^X_{D_{i,j}}\right)\right) \\
&= \frac{1}{2}\int\left(\log P^X_{D_i}-\log P^X_{D_{i,j}}\right)P^X_{D_i}dX \\
&+\frac{1}{2}\int\left(\log P^X_{D_j}-\log P^X_{D_{i,j}}\right)P^X_{D_j}dX \\
&= \left(\frac{1}{2}\int\log\left(P^X_{D_i}\right)P^X_{D_i}dx+\frac{1}{2}\int\log\left(P^X_{D_j}\right)P^X_{D_j}dX\right) \\
&-\int\log\left(P^X_{D_{i,j}}\right)P^X_{D_{i,j}}dX \\
&= -H(X|T)+H(X) \\
&= I(X;T)
\end{aligned}$$

where $H(X)$ is the entropy of X, $H(X|T)$ is the entropy of X conditioned on T, and $I(X;T)$ is the mutual information between X and T. Similarly, we also have $\mathcal{D}_{JS}((P^Z_{D_i}\parallel P^Z_{D_j}))=I(Z;T)$. Because Z is induced from X by the mapping function h then we have $Z\perp T\mid X$ and the Markov chain $T\to X\to Z$. According to data processing inequality for mutual information (Polyanskiy & Wu, 2014), we have $I(X;T)\ge I(Z;T)$ which implies $\mathcal{D}_{JS}((P^X_{D_i}\parallel P^X_{D_j}))\ge\mathcal{D}_{JS}((P^Z_{D_i}\parallel P^Z_{D_j}))$. Taking square root on both sides, we have $d_{JS}(P^X_{D_i},P^X_{D_j})\ge d_{JS}(P^Z_{D_i},P^Z_{D_j})$.

Proof of Lemma 10. We have:

$$
\begin{aligned}
\epsilon_D^{\texttt{Acc}}\left(\widehat{h}\right) &= \mathbb{E}_{z\sim P_D^Z, y\sim h_D(z)}\left[\mathcal{L}\left(\widehat{h}\left(Z\right), Y\right)\right] \\
&= \sum_{y=1}^{|\mathcal{Y}|} \mathbb{E}_{z\sim P_D^Z}\left[\mathcal{L}\left(\widehat{h}\left(Z\right), y\right) h_D(Z)_y\right] \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{Z}} \mathcal{L}\left(\widehat{h}\left(Z\right), y\right) h_D(Z)_y P_D^Z dZ \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{Z}} \mathcal{L}\left(\widehat{h}\left(Z\right), y\right) \frac{\int_{g^{-1}(Z)} f_D(X)_y P_D^X dX}{\int_{g^{-1}(Z)} P_D^X dX} \int_{g^{-1}(Z)} P_D^X dX dZ \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{Z}} \mathcal{L}\left(\widehat{h}\left(Z\right), y\right) \int_{g^{-1}(Z)} f_D(X)_y P_D^X dX dZ \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{Z}} \int_{g^{-1}(Z)} \mathcal{L}\left(\widehat{h}\left(g(X)\right), y\right) f_D(X)_y P_D^X dX dZ \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{Z}} \int_{\mathcal{X}} \mathbb{1}\left(X \in g^{-1}(Z)\right) \mathcal{L}\left(\widehat{h}\left(Z\right), y\right) f_D(X)_y P_D^X dX dZ \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{X}} \int_{\mathcal{Z}} \mathbb{1}\left(Z = g(X)\right) \mathcal{L}\left(\widehat{h}\left(Z\right), y\right) f_D(X)_y P_D^X dX dZ \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{X}} \mathcal{L}\left(\widehat{h}\left(g(X)\right), y\right) f_D(X)_y P_D^X dX dZ \\
&= \sum_{y=1}^{|\mathcal{Y}|} \int_{\mathcal{X}} \mathcal{L}\left(\widehat{f}\left(X\right), y\right) f_D(X)_y P_D^X dX \\
&= \epsilon_D^{\texttt{Acc}}\left(\widehat{f}\right)
\end{aligned}
$$

Proof of Lemma 11. We show the decomposition for KL-divergence first and then use the result to derive the decomposition for JS-divergence. We have:

$$
\begin{aligned}
&\mathcal{D}_{KL}\left(P_{D_i}^{X,Y} \parallel P_{D_j}^{X,Y}\right) \\
&= \mathbb{E}_{D_i}\left[\log P_{D_i}^{X,Y} - \log P_{D_j}^{X,Y}\right] \\
&= \mathbb{E}_{D_i}\left[\log P_{D_i}^{Y} + \log P_{D_i}^{X|Y}\right] - \mathbb{E}_{D_i}\left[\log P_{D_j}^{Y} + \log P_{D_j}^{X|Y}\right] \\
&= \mathbb{E}_{D_i}\left[\log P_{D_i}^{Y} - \log P_{D_j}^{Y}\right] + \mathbb{E}_{D_i}\left[\log P_{D_i}^{X|Y} - \log P_{D_j}^{X|Y}\right] \\
&= \mathbb{E}_{D_i}\left[\log P_{D_i}^{Y} - \log P_{D_j}^{Y}\right] + \mathbb{E}_{y\sim P_{D_i}^Y}\left[\mathbb{E}_{x\sim P_{D_i}^{X|y}}\left[\log P_{D_i}^{X|Y} - \log P_{D_j}^{X|Y}\right]\right] \\
&= \mathcal{D}_{KL}\left(P_{D_i}^{Y} \parallel P_{D_j}^{Y}\right) + \mathbb{E}_{D_i}\left[\mathcal{D}_{KL}\left(P_{D_i}^{X|Y} \parallel P_{D_j}^{X|Y}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{D}_{JS}\left(P^{X,Y}_{D_i} \parallel P^{X,Y}_{D_j}\right)\\
&= \frac{1}{2}\left(\mathcal{D}_{KL}\left(P^{X,Y}_{D_i} \parallel P^{X,Y}_{D_{i,j}}\right)\right) + \frac{1}{2}\left(\mathcal{D}_{KL}\left(P^{X,Y}_{D_j} \parallel P^{X,Y}_{D_{i,j}}\right)\right)\\
&= \frac{1}{2}\left(\mathcal{D}_{KL}\left(P^{Y}_{D_i} \parallel P^{Y}_{D_{i,j}}\right)\right) + \frac{1}{2}\left(\mathbb{E}_{D_i}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right)\\
&+ \frac{1}{2}\left(\mathcal{D}_{KL}\left(P^{Y}_{D_j} \parallel P^{Y}_{D_{i,j}}\right)\right) + \frac{1}{2}\left(\mathbb{E}_{D_j}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_j} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right)\\
&= \mathcal{D}_{JS}\left(P^{Y}_{D_i} \parallel P^{Y}_{D_j}\right) + \frac{1}{2}\left(\mathbb{E}_{D_i}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right) + \frac{1}{2}\left(\mathbb{E}_{D_j}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_j} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right)\\
&\leq \mathcal{D}_{JS}\left(P^{Y}_{D_i} \parallel P^{Y}_{D_j}\right) + \frac{1}{2}\left(\mathbb{E}_{D_i}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right) + \frac{1}{2}\left(\mathbb{E}_{D_i}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_j} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right)\\
&+ \frac{1}{2}\left(\mathbb{E}_{D_j}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_j} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right) + \frac{1}{2}\left(\mathbb{E}_{D_j}\left[\mathcal{D}_{KL}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_{i,j}}\right)\right]\right)\\
&= \mathcal{D}_{JS}\left(P^{Y}_{D_i} \parallel P^{Y}_{D_j}\right) + \mathbb{E}_{D_i}\left[\mathcal{D}_{JS}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_j}\right)\right] + \mathbb{E}_{D_j}\left[\mathcal{D}_{JS}\left(P^{X|Y}_{D_i} \parallel P^{X|Y}_{D_j}\right)\right]
\end{aligned}
$$

Proof of Lemma 12.

$$
\begin{aligned}
\mathbb{E}_D\left[\mathcal{L}(\widehat{f}(X), Y)\right] &= \mathbb{E}_D\left[\sum_{\widehat{y}\in\mathcal{Y}} \widehat{f}(X)_{\widehat{y}} L(\widehat{y}, Y)\right]\\
&\overset{(1)}{\geq} c\,\mathbb{E}_X\left[\sum_{\widehat{y}\in\mathcal{Y}} \widehat{f}(X)_{\widehat{y}} \Pr(Y \neq \widehat{y}|X)\right]\\
&\overset{(2)}{=} c\,\mathbb{E}_X\left[1 - \widehat{f}(X)^T f(X)\right]\\
&\overset{(3)}{\geq} \frac{c}{2}\,\mathbb{E}_X\left[\left\|\widehat{f}(X) - f(X)\right\|_2^2\right]\\
&\overset{(4)}{\geq} \frac{c}{2}\frac{1}{|\mathcal{Y}|}\mathbb{E}_X\left[\left(\left\|\widehat{f}(X) - f(X)\right\|_1\right)^2\right]\\
&\overset{(5)}{\geq} \frac{c}{2}\frac{1}{|\mathcal{Y}|}\left(\left\|\mathbb{E}_X\left[\widehat{f}(X) - f(X)\right]\right\|_1\right)^2\\
&= \frac{c}{2}\frac{1}{|\mathcal{Y}|}\left\|P^{\widehat{Y}}_D - P^Y_D\right\|_1^2\\
&\overset{(6)}{\geq} \frac{2c}{|\mathcal{Y}|}\mathcal{D}_{JS}\left(P^Y_D \parallel P^{\widehat{Y}}_D\right)^2\\
&= \frac{2c}{|\mathcal{Y}|}\cdot d_{JS}\left(P^Y_D, P^{\widehat{Y}}_D\right)^4
\end{aligned}
$$

Here we have $\overset{(1)}{\geq}$ is because of the assumption that $L(\widehat{y}, y)$ is lower bounded by c when $\widehat{y} \neq y$; $\overset{(2)}{=}$ is because $\widehat{f}(X)^T\mathbf{1} = ||\widehat{f}(X)||_1 = 1$; $\overset{(3)}{\geq}$ is because $||\widehat{f}(X)||_2 \leq ||\widehat{f}(X)||_1 = 1$; $\overset{(4)}{\geq}$ is because $||\widehat{f}(X)||_2 \geq \frac{1}{\sqrt{|\mathcal{Y}|}}||\widehat{f}(X)||_1$; $\overset{(5)}{\geq}$ is by using Jensen's inequality; $\overset{(6)}{\geq}$ is by using JS-divergence lower bound of total variation distance.

Proof of Lemma 14. Similar to the proof in Lemma 8, we apply Lemma 7 for $R^{y,a}_{D_i}$ and $R^{y,a}_{D_j}$ and note that $\widehat{f}(X)_y$ is bounded by 1. Then $\forall y, a \in \{0,1\}$, we have:

$$\begin{aligned}
&R^{y,a}_{D_j} - \mathbb{E}_{D_{i,j}}\left[\widehat{f}(X)_y|Y=y,A=a\right] \\
&= \mathbb{E}_{D_j}\left[\widehat{f}(X)_y|Y=y,A=a\right] - \mathbb{E}_{D_{i,j}}\left[\widehat{f}(X)_y|Y=y,A=a\right] \\
&\le \frac{1}{\sqrt{2}}\sqrt{\min\left(\mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_i} \parallel P^{X|Y=y,A=a}_{D_{i,j}}\right), \mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_{i,j}} \parallel P^{X|Y=y,A=a}_{D_i}\right)\right)} \\
&\le \frac{1}{\sqrt{2}}\sqrt{\mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_j} \parallel P^{X|Y=y,A=a}_{D_{i,j}}\right)} \qquad (18)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{D_{i,j}}\left[\widehat{f}(X)_y|Y=y,A=a\right] - R^{y,a}_{D_i} \\
&= \mathbb{E}_{D_{i,j}}\left[\widehat{f}(X)_y|Y=y,A=a\right] - \mathbb{E}_{D_i}\left[\widehat{f}(X)_y|Y=y,A=a\right] \\
&\le \frac{1}{\sqrt{2}}\sqrt{\min\left(\mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_j} \parallel P^{X|Y=y,A=a}_{D_{i,j}}\right), \mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_{i,j}} \parallel P^{X|Y=y,A=a}_{D_j}\right)\right)} \\
&\le \frac{1}{\sqrt{2}}\sqrt{\mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_i} \parallel P^{X|Y=y,A=a}_{D_{i,j}}\right)} \qquad (19)
\end{aligned}$$

Adding Eq. (18) to Eq. (19), we have:

$$\begin{aligned}
&R^{y,a}_{D_j} - R^{y,a}_{D_i} \\
&\le \frac{1}{\sqrt{2}}\left(\sqrt{\mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_i} \parallel P^{X|Y=y,A=a}_{D_{i,j}}\right)} + \sqrt{\mathcal{D}_{KL}\left(P^{X|Y=y,A=a}_{D_j} \parallel P^{X|Y=y,A=a}_{D_{i,j}}\right)}\right) \\
&\le \sqrt{2}d_{JS}\left(P^{X|Y=y,A=a}_{D_j}, P^{X|Y=y,A=a}_{D_{i,j}}\right)
\end{aligned}$$

Proof of Lemma 15. We give the proof for unfairness measure w.r.t. to equal opportunity first and then use this result to derive the proof for unfairness measure w.r.t. to equalized odd. Without loss of generality, assign group indices 1, 0 be such that $R^{1,0}_{D_j}\left(\widehat{f}\right) \ge R^{1,1}_{D_j}\left(\widehat{f}\right)$. Then we have:

$$\begin{aligned}
\epsilon^{\text{EP}}_{D_j}\left(\widehat{f}\right) &= \left|R^{1,0}_{D_j}\left(\widehat{f}\right) - R^{1,1}_{D_j}\left(\widehat{f}\right)\right| \\
&= R^{1,0}_{D_j}\left(\widehat{f}\right) - R^{1,1}_{D_j}\left(\widehat{f}\right) \\
&= R^{1,0}_{D_j}\left(\widehat{f}\right) - \mathbb{E}_{D_j}\left[\widehat{f}(X)_1|Y=1,A=1\right] \\
&= R^{1,0}_{D_j}\left(\widehat{f}\right) + \mathbb{E}_{D_j}\left[1-\widehat{f}(X)_1|Y=1,A=1\right] - 1 \\
&= R^{1,0}_{D_j}\left(\widehat{f}\right) + R^{1,1}_{D_j}\left(\mathbf{1}-\widehat{f}\right) - 1
\end{aligned}$$

where $\mathbf{1}$ is vector with all 1's. By Lemma 14, we have:

$$\begin{aligned}
R^{1,0}_{D_j}\left(\widehat{f}\right) &\le R^{1,0}_{D_i}\left(\widehat{f}\right) + \sqrt{2}d_{JS}\left(P^{X|Y=1,A=0}_{D_j}, P^{X|Y=1,A=0}_{D_i}\right) \\
R^{1,1}_{D_j}\left(\mathbf{1}-\widehat{f}\right) &\le R^{1,1}_{D_i}\left(\mathbf{1}-\widehat{f}\right) + \sqrt{2}d_{JS}\left(P^{X|Y=1,A=1}_{D_j}, P^{X|Y=1,A=1}_{D_i}\right)
\end{aligned}$$

Sum above two inequalities and add -1 at both sides, we have,

$$\begin{aligned}
\epsilon^{\text{EP}}_{D_j}\left(\widehat{f}\right) &= R^{1,0}_{D_j}\left(\widehat{f}\right) + R^{1,1}_{D_j}\left(\mathbf{1}-\widehat{f}\right) - 1 \\
&\le R^{1,0}_{D_i}\left(\widehat{f}\right) + R^{1,1}_{D_i}\left(\mathbf{1}-\widehat{f}\right) - 1 + \sqrt{2}\sum_{a=0,1} d_{JS}\left(P^{X|Y=1,A=a}_{D_j}, P^{X|Y=1,A=a}_{D_i}\right) \\
&\le \epsilon^{\text{EP}}_{D_i}\left(\widehat{f}\right) + \sqrt{2}\sum_{a=0,1} d_{JS}\left(P^{X|Y=1,A=a}_{D_j}, P^{X|Y=1,A=a}_{D_i}\right) \qquad (20)
\end{aligned}$$

Similarly, we have:

$$\left|R^{0,0}_{D_j}\left(\widehat{f}\right)-R^{0,1}_{D_j}\left(\widehat{f}\right)\right|\le\left|R^{0,0}_{D_i}\left(\widehat{f}\right)-R^{0,1}_{D_i}\left(\widehat{f}\right)\right|+\sqrt{2}\sum_{a=0,1}d_{JS}\left(P^{X|Y=0,A=a}_{D_j},P^{X|Y=0,A=a}_{D_i}\right) \tag{21}$$

Sum both Eq. (20) and Eq. (21), we have:

$$\epsilon^{\text{EO}}_{D_j}\left(\widehat{f}\right)\le\epsilon^{\text{EO}}_{D_i}\left(\widehat{f}\right)+\sqrt{2}\sum_{y=0,1}\sum_{a=0,1}d_{JS}\left(P^{X|Y=y,A=a}_{D_j},P^{X|Y=y,A=a}_{D_i}\right)$$

Proof of Lemma 16. Similar to the proof of Lemma 10, $R^{y,a}_{D_i}\left(\widehat{f}\right)=R^{y,a}_{D_i}\left(\widehat{h}\right)\ \forall y,a\in\{0,1\}$. Then, we have:

$$\begin{aligned}
\epsilon^{\text{EO}}_{D_i}\left(\widehat{f}\right)&=\left|R^{0,0}_{D_i}\left(\widehat{f}\right)-R^{0,1}_{D_i}\left(\widehat{f}\right)\right|+\left|R^{1,0}_{D_i}\left(\widehat{f}\right)-R^{1,1}_{D_i}\left(\widehat{f}\right)\right|\\
&=\left|R^{0,0}_{D_i}\left(\widehat{h}\right)-R^{0,1}_{D_i}\left(\widehat{h}\right)\right|+\left|R^{1,0}_{D_i}\left(\widehat{h}\right)-R^{1,1}_{D_i}\left(\widehat{h}\right)\right|\\
&=\epsilon^{\text{EO}}_{D_i}\left(\widehat{h}\right)\\
\epsilon^{\text{EP}}_{D_i}\left(\widehat{f}\right)&=\left|R^{1,0}_{D_i}\left(\widehat{f}\right)-R^{1,1}_{D_i}\left(\widehat{f}\right)\right|\\
&=\left|R^{1,0}_{D_i}\left(\widehat{h}\right)-R^{1,1}_{D_i}\left(\widehat{h}\right)\right|\\
&=\epsilon^{\text{EP}}_{D_i}\left(\widehat{h}\right)
\end{aligned}$$

Proof of Lemma 17. $\forall y\in\mathcal{Y}$, we have:

$$\begin{aligned}
\mathcal{D}_{JS}\left(P^{Z|y}_i\parallel P^{Z|y}_j\right)&\overset{(1)}{=}\mathcal{D}_{JS}\left(\int_{\mathcal{X}}P^{Z|x}P^{x|y}_i dx\parallel\int_{\mathcal{X}}P^{Z|m^y_{i,j}(x)}P^{m^y_{i,j}(x)|y}_j dm^y_{i,j}(x)\right)\\
&\overset{(2)}{=}\mathcal{D}_{JS}\left(\int_{\mathcal{X}}P^{Z|x}P^{x|y}_i dx\parallel\int_{\mathcal{X}}P^{Z|m^y_{i,j}(x)}P^{m^y_{i,j}(x)|y}_j dx\right)\\
&\overset{(3)}{=}\mathcal{D}_{JS}\left(\int_{\mathcal{X}}P^{Z|x}P^{x|y}_i dx\parallel\int_{\mathcal{X}}P^{Z|m^y_{i,j}(x)}P^{x|y}_i dx\right)\\
&\overset{(4)}{\le}\int_{\mathcal{X}}P^{x|y}_i\mathcal{D}_{JS}\left(P^{Z|x}\parallel P^{Z|m^y_{i,j}(x)}\right)dx
\end{aligned}$$

Here we have $\overset{(1)}{=}$ is because of law of total probability and $Z\perp Y|X$; $\overset{(2)}{=}$ is because $m^y_{i,j}$ is invertible function; $\overset{(3)}{=}$ is because $P^{x|y}_i=P^{m^y_{i,j}(x)|y}_j\quad\forall x\in\mathcal{X}$; $\overset{(4)}{\le}$ is because of joint complexity of JS divergence. By similar derivation, $\forall y\in\mathcal{Y},a\in\mathcal{A}$, we have:

$$\mathcal{D}_{JS}\left(P^{Z|y,a}_i\parallel P^{Z|y,a}_j\right)\le\int_{\mathcal{X}}P^{x|y,a}_i\mathcal{D}_{JS}\left(P^{Z|x}\parallel P^{Z|m^{y,a}_{i,j}(x)}\right)dx$$