Chi-square Information for Invariant Learning

Prasanna Sattigeri [1] **Soumya Ghosh** [1] **Samuel C. Hoffman** [1]

Abstract

Invariant learning aims to train models robust to nuisance confounding that may be present in the data. This is typically achieved by minimizing some measure of dependence between learned representations or predictions and confounding factors. However, accurate estimation as well as reliable minimization of typically used dependence measures can be challenging. Chi-square divergence based dependence measure has recently been found effective in enforcing fairness through learning invariant representations. We show that with an appropriate parameterization, this choice both improves dependence estimation quality and simplifies its minimization. Empirically, we find that our proposal is effective at fair predictor learning and domain generalization.

1. Introduction

Consider a supervised setting where we are interested in learning a mapping between an outcome y and covariates $\mathbf{x}$. Additionally, we assume that a sensitive attribute or a nuisance factor s is observed and may be correlated with $\mathbf{x}$ and/or y. A model, parameterized by θ, that learns to predict $\hat{y} = f_\theta(\mathbf{x})$ ignoring s, is said to be invariant to s. Such invariance is desirable in many applications. For instance, we may wish to learn fair machine learning models, where a model's predictions are invariant to attributes such as race or gender, or we may wish our models to be robust to domain shift by being invariant to the domain (Fernando et al., 2013; Ganin & Lempitsky, 2015) or even generalize to unseen domains (Muandet et al., 2013; Ghifary et al., 2015). The notion of invariance can be formalized via conditional independence statements involving y, $\mathbf{x}$ and s, where the exact form is dependent on the application. For example, in fairness, the property $\hat{y} \perp\!\!\!\perp s$, where model predictions are not dependent on sensitive attributes such as race and gender, is called demographic parity (Barocas et al., 2019). In practice, the independence criterion is replaced by dependence measures. Thus, estimation and minimization of dependence between two variables is crucial for learning models that are robust or invariant to nuisance factors. Mutual Information (MI), defined as the Kullback-Leibler (KL) divergence between a joint distribution and the product of marginal distributions, is a popular choice for measuring dependence. In (Micheas & Zografos, 2006), the authors consider generalizations of mutual information that involve replacing the KL-divergence with other f-divergences. Here, we focus on one such f-divergence — the χ^2-divergence and the corresponding dependence measure. It is an upper bound to Mutual Information (MI) and Gebelein-Rényi Maximum Correlation Coefficient (HGR) dependence measures that satisfies Rényi properties. When one of the variables is binary, Chi-square information is equivalent to HGR (Mary et al., 2019). Hence, it can also be employed as a surrogate for minimizing MI or HGR. Estimating the dependence measure between two variables (χ^2 or otherwise) is challenging. It typically requires either the estimation of complex joint and/or marginal densities of the variables or the ability to sample from these distributions; both of which are daunting, especially in high dimensions (Gao et al., 2015; 2017). However, the conditional distributions are often easier to approximate (Poole et al., 2019) and may even be known in one direction in certain applications such as representation learning. In this work, we exploit this insight to develop tractable approximations to the χ^2 dependence measure. Our contributions include,

- We show that the Chi-square information between two variables X and Y can be expressed in a convenient form that involves the product of conditional densities $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{y})$. Additionally, we propose unbiased Monte-Carlo estimates for this dependence measure based on this form for discrete and continuous cases.
- Through synthetic experiments and real world fairness applications, we demonstrate that estimating and minimizing the dependence of two variables by estimating their conditional distributions is simpler and just as effective compared to learning the joint distribution $p(\mathbf{x}, \mathbf{y})$ and/or the marginal distributions $p(\mathbf{x})$ and $p(\mathbf{y})$.
- Finally, through domain generalization application, we

[1]IBM Research. Correspondence to: Prasanna Sattigeri <psattig@us.ibm.com>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

demonstrate the proposed approach can be effective in minimizing the dependence between high-dimensional continuous variables.

2. Estimating Chi-square Information using Conditional Distributions

In this section, we introduce Monte-Carlo estimators for computing the Chi-square mutual information (shortened as CHI2MI) between two random variables X and Y. We assume the existence of the marginal distributions P_X, P_Y and the joint distribution $P_{X,Y}$ over $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X}\times\mathcal{Y}$, respectively, with known density $p(\mathbf{y}|\mathbf{x})$ and with either known density $p(\mathbf{x}|\mathbf{y})$ or the ability to sample from distribution $P_{X|Y}$. Let us first define the χ^2-divergence between two distributions as:

Definition 1 (χ^2-divergence)**.** *Let P_X and Q_X be two distributions that admit density $p(\mathbf{x})$ and $q(\mathbf{x})$ respectively. The χ^2-divergence between these two distributions is*

$$D_{\chi^2}(P_X||Q_X) = \int \Big(\Big(\frac{p(\mathbf{x})}{q(\mathbf{x})}\Big)^2 - 1\Big) q(\mathbf{x}) d\mathbf{x} \quad (1)$$

This divergence belongs to a general family of distributions called f-divergences that include commonly used divergences such as Kullback-Leibler (KL) and Jenson-Shannon (JS) divergences. The Chi-square mutual information $MI_{\chi^2}(X;Y)$ is defined in terms of χ^2-divergence as:

Definition 2 (Chi-square mutual information)**.** *Let P_X, P_Y and $P_{X,Y}$ denote the marginal distributions and the joint distribution of X and Y admitting densities $p(\mathbf{x})$, $p(\mathbf{y})$ and $p(\mathbf{x},\mathbf{y})$, respectively. The Chi-square dependence measure between two random variables X and Y is*

$$\begin{aligned} MI_{\chi^2}(X;Y) &= D_{\chi^2}(P_{X,Y}||P_X P_Y) \\ &= \int\int \frac{p(\mathbf{x},\mathbf{y})^2}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x}d\mathbf{y} - 1 \end{aligned} \quad (2)$$

We can obtain the well known Mutual Information by replacing the χ^2-divergence with KL-divergence between the joint distribution $p(\mathbf{x},\mathbf{y})$ and product of marginals $p(\mathbf{x})p(\mathbf{y})$.

Remark 1. *Let $P_{X|Y}$ and $P_{Y|X}$ be the conditional distributions that admit densities $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{y}|\mathbf{x})$, respectively. $MI_{\chi^2}(X;Y)$ between X and Y can be expressed as:*

$$MI_{\chi^2}(X;Y) = \int\int p(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y} \mid \mathbf{x}) d\mathbf{x}d\mathbf{y} - 1, \quad (3)$$

Proof can be found in the Appendix A. When $\mathbf{x},\mathbf{y}$ are conditionally independent, we have $MI_{\chi^2}(X;Y) = \int p(\mathbf{x}|\mathbf{y})p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} = \int p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y} - 1 = 0$. We can similarly define the conditional Chi-square dependence measure $MI_{\chi^2}(X;Y|Z)$ between X and Y conditioned on Z as,

$$\begin{aligned} MI_{\chi^2}(X;Y|Z) &= \mathbb{E}_{p(\mathbf{z})}[D_{\chi^2}(P_{X,Y|Z}||P_{X|Z}P_{Y|Z})] \\ &= \mathbb{E}_{p(\mathbf{z})}\Big[\int\int p(\mathbf{x}|\mathbf{y},\mathbf{z})p(\mathbf{y}|\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{y}\Big] - 1 \end{aligned} \quad (4)$$

where $P_{X|Z}$, $P_{Y|Z}$ and $P_{X,Y|Z}$ denote the marginal conditional distributions and joint conditional distribution of X and Y conditioned on Z.

When one or both of the variables are discrete the above form of $MI_{\chi^2}(X;Y)$ leads to simple Monte-Carlo estimates based on the conditional distributions (Derivation can be found in Appendix B).

When both X and Y are continuous, we can express $MI_{\chi^2}(X;Y)$ as

$$\begin{aligned} MI_{\chi^2}(X;Y) &= \int\int p(\mathbf{x}|\mathbf{y})p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} - 1 \\ &= \int \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[p(\mathbf{y}|\mathbf{x})]d\mathbf{y} - 1 \end{aligned} \quad (5)$$

Let us denote $\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}[p(\mathbf{y}|\mathbf{x})]$ as $f(\mathbf{y})$ and $\hat{f}(\mathbf{y}_j) = \frac{1}{N}\sum_{i=0}^{N-1} p(\mathbf{y}_j \mid \mathbf{x}_{i,j})$ as the empirical estimate where $\mathbf{x}_{i,j} \sim p(\mathbf{x}|\mathbf{y}_j)$. By replacing the expectation inside the integral with a Monte-Carlo estimate, we arrive at the following estimate:

$$\hat{MI}_{\chi^2}(X;Y) = \int \hat{f}(\mathbf{y}) d\mathbf{y} - 1 \quad (6)$$

Let us first consider a special case, when one of the variables is scalar and, without loss of generality, let us assume that the scalar variable is Y. In this case, the integral in Equation 6 can be efficiently and accurately estimated using numerical quadrature (Davis & Rabinowitz, 1967). The parameters $\phi_{\mathbf{x}\to y}$ of the conditional density $p(y \mid \mathbf{x})$ can to be estimated from the observed data after choosing a suitable form for the distribution. Similar to the discrete case, we only need samples from $P_{X|Y}$. This is especially beneficial when X is high-dimensional since we can employ implicit generative models to sample without simplistic assumption on the form of the distribution.

When numerical quadrature is infeasible, we can estimate the integral in 6 using *Importance Sampling* (IS) with a proposal distribution $q(\mathbf{x},\mathbf{y}) = q(\mathbf{y})p(\mathbf{x} \mid \mathbf{y})$:

$$\begin{aligned} MI_{\chi^2}(X;Y) &= \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\Big[\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\Big] - 1 \\ &= \mathbb{E}_{q(\mathbf{x},\mathbf{y})}\Big[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\Big] - 1 \\ &= \mathbb{E}_{q(\mathbf{y})p(\mathbf{x}|\mathbf{y})}\Big[\frac{p(y \mid x)}{q(\mathbf{y})}\Big] - 1 \end{aligned} \quad (7)$$

Using M samples from the proposal distribution, we arrive at the CHI2MI-MC estimator:

$$\hat{MI}_{\chi^2}(X;Y) = \frac{1}{M}\sum_{i=0}^{M-1}\frac{p(\mathbf{y}_i \mid \mathbf{x}_i)}{q(\mathbf{y}_i)} - 1 \quad \text{where } (\mathbf{x}_i, \mathbf{y}_i) \sim q(\mathbf{y})p(\mathbf{x}|\mathbf{y}). \tag{8}$$

In Appendix C we show the effect of M and choice of proposal distribution for a synthetic dataset where we can analytically compute ground truth $MI_{\chi^2}(X;Y)$.

3. Applications

We investigate the use of CHI2MC estimate in two applications that involve minimizing dependence or conditional dependence. First, we show that the estimate can be used to enforce fairness criteria when learning predictors that performs similar or better than methods that involve KDE-based density estimation. Next, we demonstrate, through the domain generalization application, the scalablity of the estimate to high-dimensional variables where it is difficult to learn good KDE based density estimates.

3.1. Learning Fair Predictors

Let us denote random variables X, Y and S as the covariates, the ground truth outcome and the sensitive attribute. We denote $\hat{Y} = h_\theta(X)$ as the prediction using the model $h_\theta(\cdot)$. Two common metrics used to measure fairness are statistical parity (Barocas et al., 2019) and equality of odds (Hardt et al., 2016). These metrics can be expressed as independence statements. Statistical parity, also called demographic parity or *Independence*, implies $\hat{Y} \perp\!\!\!\perp S$. We can achieve equality of odds, also called *Separation*, by rendering the predictions conditionally independent of the sensitive attribute given the true outcomes ($\hat{Y} \perp\!\!\!\perp S|Y$). Finally, equality of opportunity is a special case of equality of odds where the predictions are conditionally independent of the sensitive attribute given the true outcome is positive: $\hat{Y} \perp\!\!\!\perp S|Y = 1$. The difference in equal opportunity (DEO) is, therefore, defined as the absolute value of the difference in true positive rates between groups.

In this section, we use the scalable CHI2MC estimates to directly minimize $MI_{\chi^2}(S;\hat{Y})$ and $MI_{\chi^2}(S;\hat{Y}|Y)$ to achieve demographic parity and equalized odds, respectively. Concretely, we solve the following optimization problem to improve demographic parity:

$$\min_\theta \mathbb{E}_{p(\mathbf{x},y)}[l(y, h_\theta(\mathbf{x}))] + \lambda MI_{\chi^2}(S; h_\theta(X)) \tag{9}$$

To encourage equality of odds, $MI_{\chi^2}(S; h_\theta(X)|Y)$ is instead minimized. Appendix F shows the exact expressions for the estimates. The estimation involves learning the parameters ϕ of the conditional distribution $p_\phi(s|\hat{y})$ in the case of $MI_{\chi^2}(S; h_\theta(X))$ and the conditional distribution $p_\phi(s|\hat{y}, y)$ in the case of $MI_{\chi^2}(S; h_\theta(X)|Y)$.

Experiments: We benchmark our method on four popular algorithmic fairness datasets: COMPAS (violent recidivism), Adult, German Credit Data, and Drug Consumption. We follow the preprocessing steps and classification tasks from (Mary et al., 2019). The $\hat{Y}$-predictor (θ) network we use for these experiments has two hidden layers with sizes 20 and 10 with BatchNorm (Ioffe & Szegedy, 2015), ReLU non-linearities, and Dropout ($p = 0.2$) after each hidden layer. We reproduce the results from (Mary et al., 2019) using this network (NN + χ^2-KDE in Table 1). We use a simple linear model to estimate S.

For both models we use an Adam optimizer and alternate training the S-predictor (ϕ) and $\hat{Y}$-predictor for 5 steps and 20 steps, respectively, 100 times (with a 2 epoch "warm-start" training for the $\hat{Y}$-predictor first). We do 10-fold cross validation repeated 20 times for each dataset (except Adult for which we use the provided test split). Using a validation set consisting of 30% or 50% of the training set (thus the actual training set size is reduced to 70% or 50% as well), we compute the validation DEO and use early stopping to pick the best iteration for each fold subject to a balanced accuracy threshold of 0.65 or 0.55 (drug only). Finally, the best hyperparameters are chosen using the average validation DEO. We use $\lambda = 1.0$ for all results and report the best results for batch size from $\{100, 250, 1000\}$.

For COMPAS, we were unable to reproduce similar accuracy results to (Mary et al., 2019) for either NN + χ^2-KDE or NN + CHI2MC but DEO was still competitive. Similarly, we note that the accuracy on Adult for NN + χ^2-KDE is abnormally low. Overall, our method shows competitive performance on this task. As noted in (Mary et al., 2019), many of these datasets make poor benchmarks for deep learning models due to their small size. German, and Drug contain only 1000, and 1885 total samples, respectively. COMPAS and Adult are significantly larger and consequently our method performs well on these datasets.

3.2. Domain Generalization

DIVA (Ilse et al., 2019) is an Variational Autoencoder (VAE)-based generative approach for learning domain-invariant feature representations. The generative process involves a decoder $p_\theta(\mathbf{x}|\mathbf{z}_y, \mathbf{z}_d, \mathbf{z}_x)$ that transforms latent representations $\mathbf{z}_y$, $\mathbf{z}_d$ and $\mathbf{z}_x$, drawn from independent Gaussian prior $p(\mathbf{z}_x)$ and learnable conditional priors $p(\mathbf{z}_d|d)$, $p(\mathbf{z}_y|y)$, that capture the class label, domain label and other variables, respectively, into the observed sample $\mathbf{x}$. Variational posteriors are obtained using three separate encoders $q_{\phi_y}(\mathbf{z}_y|\mathbf{x})$, $q_{\phi_d}(\mathbf{z}_d|\mathbf{x})$, $q_{\phi_x}(\mathbf{z}_x|\mathbf{x})$ which enforce the following factorization of the marginal distribution over latent: $q_\phi(\mathbf{z}) = q_{\phi_y}(\mathbf{z}_y)q_{\phi_d}(\mathbf{z}_d)q_{\phi_x}(\mathbf{z}_x)$.

Table 1. Results showing the average and standard deviation of the accuracy and Difference in Equality of Opportunity (DEO) metrics for benchmark datasets.

Method	COMPAS		Adult		German		Drug	
	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Naive SVM (Donini et al., 2018)	0.72±0.01	0.14±0.02	0.80	0.09	0.74±0.05	0.12±0.05	0.81±0.02	0.22±0.04
Mary et. al.(Mary et al., 2019)	0.96* ± 0.00	0.00±0.00	0.83	0.03	0.73±0.03	0.25±0.14	0.78±0.05	0.00±0.00
NN + χ^2-KDE	0.83±0.01	0.09±0.06	0.76	0.00	0.73±0.04	0.19±0.16	0.85±0.01	0.06±0.12
NN + CHI2MC	0.83±0.01	0.10±0.05	0.84	0.02	0.68±0.05	0.24±0.17	0.84±0.02	0.10±0.01

Table 2. The effect of CHI2MC regularization on the ROC AUC performance of the model on an unseen test domain when unlabeled samples from additional domain are included in the training.

Dataset	Unsupervised Domain	Test Domain	$\lambda = 0.0$ (DIVA*)	$\lambda = 0.1$	$\lambda = 1.0$
Rotated MNIST	$\mathcal{M}_{60^\circ}$	$\mathcal{M}_{75^\circ}$	76.9 ± 1.2	77.1 ± 0.8	$\mathbf{78.4 \pm 0.7}$
Malaria Cell Images	C59P20	C116P77	70.3 ± 2.5	71.7 ± 2.9	$\mathbf{75.5 \pm 2.5}$

We study the semi-supervised variant of DIVA that was found to be more challenging than the fully supervised setting. We modify the DIVA semi-supervised lower bound $\mathcal{F}_{SS-DIVA}$ (See Appendix G for derivation) to explicitly minimize the dependence between the domain and label specific representations, Z_d and Z_y, respectively. Maximizing the following lower bound minimizes any leakage of information between the two latent representations and render them even more separated.

$$\mathcal{F}_{SS-CHI2MC} := \mathcal{F}_{SS-DIVA} - \lambda \hat{MI}_{\chi^2}(Z_d; Z_y) \quad (10)$$

Assuming N labeled samples and M unlabeled samples, the MI_{χ^2} estimate is computed using $M + N$ samples $(\mathbf{z}_{y,i}, \mathbf{z}_{d,i})$ from the empirical distribution consisting of T datapoints as below:

$$\hat{MI}_{\chi^2}(Z_d; Z_y) = \frac{T}{N+M} \sum_{i=0}^{N+M-1} p_{\phi_{\mathbf{z},d}}(\mathbf{z}_{d,i} \mid \mathbf{z}_{y,i}) - 1 \quad (11)$$

Experiments: We show improvement over the DIVA baseline in cases where additional unlabeled data is included in the training data. In Table 3.1, we show results for this on two settings: the rotated MNIST (Ghifary et al., 2015) and malaria cell images (Rajaraman et al., 2018) datasets. We follow the procedure in (Ilse et al., 2019) to construct the domains in each dataset. Rotated MNIST consists of 6 domains. Domain $\mathcal{M}_{0^\circ}$ is formed by sampling 100 images from each of the 10 classes in the original MNIST dataset. Additional domains are created by rotating the digits by 15, 30, 45, 60 and 75 degrees and are denoted as $\mathcal{M}_{15^\circ}$, $\mathcal{M}_{30^\circ}$, $\mathcal{M}_{45^\circ}$, $\mathcal{M}_{60^\circ}$ and $\mathcal{M}_{75^\circ}$, respectively. Malaria cell images is a more challenging dataset consisting of 27558 single red blood cell images taken from 50 and 150 healthy and infected patients, respectively. 10 patients among these with the highest amount of cells are chosen as 10 domains with total 5922 images. The domains are denoted by the patient IDs and the task is to predict the healthy or infected state.

We study the performance of the model on an unseen test domain when maximizing the $\mathcal{F}_{SS-CHI2MC}$ objective for the two datasets. All the model architectures and hyperparameters pertaining to DIVA are set to the default values in the code accompanying the paper (Ilse et al., 2019). We do not perform any grid search for these hyperparameters. We provide these details in the Appendix H. We use a Gaussian distribution for $p_{\phi_{z,d}}(\mathbf{z}_d|\mathbf{z}_y) = \mathcal{N}(a_{\phi_{\mathbf{z},d}}(\mathbf{z}_y), I)$. The architecture of $a_{\phi_{\mathbf{z},d}}(\cdot)$ consists of a fully-connected network with a single hidden layer of dimension 20. Table 3.1 shows the average and standard error of ROC AUC on the test domain with increasing CHI2MC regularization strength (λ), where $\lambda = 0$ corresponds to vanilla DIVA model. We repeat the experiments 10 times for each dataset. We can observe that minimizing $\hat{MI}_{\chi^2}(Z_d; Z_y)$ consistently leads to improved performance on both datasets.

4. Conclusions

In this paper, we introduced an alternate form of chi-square mutual information that simplifies and improves its estimation quality. We show that the CHI2MC estimates based on this form are effective for invariant learning and can handle discrete and continuous variables. We employed CHI2MC for learning fair predictors and achieve similar or better performance on benchmarks datasets. Furthermore, we demonstrate that CHI2MC is effective in learning high-dimensional invariant representations that lead to improved domain generalization performance.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Barber, D. and Agakov, F. V. Information maximization in noisy channels: A variational approach. In *Advances in Neural Information Processing Systems*, pp. 201–208, 2004.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

Barrett, M., Kementchedjhieva, Y., Elazar, Y., Elliott, D., and Søgaard, A. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6331–6336, 2019.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Davis, P. J. and Rabinowitz, P. Numerical integration. 1967.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. Unsupervised visual domain adaptation using subspace alignment. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 2960–2967. IEEE, 2013.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

Gao, S., Ver Steeg, G., and Galstyan, A. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286, 2015.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. Estimating mutual information for discrete-continuous mixtures. In *Advances in neural information processing systems*, pp. 5986–5997, 2017.

Ghifary, M., Bastiaan Kleijn, W., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559, 2015.

Greenfeld, D. and Shalit, U. Robust learning with the hilbert-schmidt independence criterion. *arXiv preprint arXiv:1910.00270*, 2019.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. Diva: Domain invariant variational autoencoders. *arXiv preprint arXiv:1905.10427*, 2019.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 448–456. JMLR.org, 2015.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019.

Lopez-Paz, D., Hennig, P., and Schölkopf, B. The randomized dependence coefficient. In *Advances in neural information processing systems*, pp. 1–9, 2013.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Maria Carlucci, F., Russo, P., Tommasi, T., and Caputo, B. Hallucinating agnostic images to generalize across domains. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Mary, J., Calauzenes, C., and El Karoui, N. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pp. 4382–4391, 2019.

Micheas, A. C. and Zografos, K. Measuring stochastic dependence using ϕ-divergence. *Journal of Multivariate Analysis*, 97(3):765–784, 2006.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5715–5725, 2017.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Ver Steeg, G. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pp. 9084–9093, 2018.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.

Ozair, S., Lynch, C., Bengio, Y., Van den Oord, A., Levine, S., and Sermanet, P. Wasserstein dependency measure for representation learning. In *Advances in Neural Information Processing Systems*, pp. 15578–15588, 2019.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Poole, B., Ozair, S., Oord, A. v. d., Alemi, A. A., and Tucker, G. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568, 2018.

Roy, P. C. and Boddeti, V. N. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2586–2594, 2019.

Shu, R., Bui, H. H., Narui, H., and Ermon, S. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., and Zuo, W. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2272–2281, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

A. Proof of Remark 1

Remark 1. *Let $P_{X|Y}$ and $P_{Y|X}$ be the conditional distributions that admit densities $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{y}|\mathbf{x})$, respectively. $MI_{\chi^2}(X;Y)$ between X and Y can be expressed as:*

$$MI_{\chi^2}(X;Y) = \int\int p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y} \mid \mathbf{x})d\mathbf{x}d\mathbf{y} - 1, \quad (3)$$

Proof. Starting from the definition of Chi-square dependence measure (See definition 1),

$$\begin{aligned}
MI_{\chi^2}(X;Y) &= D_{\chi^2}(P_{X,Y}||P_X P_Y) \\
&= \int\int \frac{p(\mathbf{x},\mathbf{y})^2}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x}d\mathbf{y} - 1 \\
MI_{\chi^2}(X;Y) + 1 &= \int\int \Big(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\Big)^2 p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int \left(\frac{p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right)^2 p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int \left(\frac{p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right)^2 p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int \left(\frac{p(\mathbf{x} \mid \mathbf{y})}{p(\mathbf{x})}\right)^2 p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int \left(\frac{p(\mathbf{x} \mid \mathbf{y})}{p(\mathbf{x})}\right)^2 p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int \frac{p(\mathbf{x} \mid \mathbf{y})^2}{p(\mathbf{x})} p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int \frac{p(\mathbf{x} \mid \mathbf{y})^2}{p(\mathbf{x})} p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\
&= \int\int \frac{p(\mathbf{x} \mid \mathbf{y})p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} d\mathbf{x}d\mathbf{y} \\
&= \int\int \frac{p(\mathbf{x} \mid \mathbf{y})p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})} d\mathbf{x}d\mathbf{y} \\
&= \int\int p(\mathbf{x} \mid \mathbf{y})p(\mathbf{y} \mid \mathbf{x})d\mathbf{x}d\mathbf{y}
\end{aligned} \quad (12)$$

□

B. $MI_{\chi^2}(X;Y)$ when one of the variables is discrete

Case 1: Both variables are discrete.

$$\hat{MI}_{\chi^2}(X;Y) = \sum_j \sum_i p(Y = j \mid X = i)p(X = i \mid Y = j) - 1 \quad (13)$$

$$\hat{MI}_{\chi^2}(X;Y|Z) = \frac{1}{K}\sum_{k=0}^{K-1}\sum_{j}\sum_{i} p(Y=j \mid X=i,\mathbf{z}_k)p(X=i \mid Y=j,\mathbf{z}_k) - 1 \quad \text{where } \mathbf{z}_k \sim p(\mathbf{z}) \tag{14}$$

Case 2: One of the variables is discrete. Without loss of generality, let us assume Y is discrete.

$$\hat{MI}_{\chi^2}(X;Y) = \sum_j \int p(Y=j \mid \mathbf{x})p(\mathbf{x}|Y=j)d\mathbf{x} - 1 = \sum_j \mathbb{E}_{p(\mathbf{x}|Y=j)}[p(Y=j \mid \mathbf{x})] - 1 \tag{15}$$

If we draw N_j samples from each empirical conditional distribution $p(\mathbf{x}|Y=j)$, we can approximate the $MI_{\chi^2}(X;Y)$ as

$$\hat{MI}_{\chi^2}(X;Y) = \sum_{j=0}^{M-1}\frac{1}{N_j}\sum_{i=0}^{N_j-1} p(Y=j \mid \mathbf{x}_{i,j}) - 1 \quad \text{where } \mathbf{x}_{i,j} \sim p(\mathbf{x}|Y=j) \tag{16}$$

Samples from the conditional distribution $p(\mathbf{x}|Y=j)$ can be directly obtained from the observed empirical data without learning the distribution. Thus, we can obtain this estimate by first assuming a form for the conditional distribution $p(Y=j \mid \mathbf{x})$ with appropriate parametrization $\phi_{\mathbf{x}\to y}$ and subsequently learning the parameters by maximizing the conditional likelihood of the observed data.

We can extend this to conditional χ^2 mutual information $MI_{\chi^2}(X;Y|Z)$ as

$$\hat{MI}_{\chi^2}(X;Y|Z) = \frac{1}{K}\sum_{k=0}^{K-1}\sum_{j=0}^{M-1}\frac{1}{N_j}\sum_{i=0}^{N_j-1} p(Y=j \mid \mathbf{x}_{i,j,k},\mathbf{z}_k) - 1 \quad \text{where } \mathbf{z}_k \sim p(\mathbf{z}) \text{ and } \mathbf{x}_{i,j,k} \sim p(\mathbf{x}|Y=j,\mathbf{z}_k) \tag{17}$$

Samples $\mathbf{z}_k$ and $\mathbf{x}_{i,j,k}$ can be easily obtained from the training samples for discrete Z. For continuous Z we can learn a sampler for the distribution $P_{X|Y,Z}$.

C. Experiments on Synthetic Data

Consider two variables X and Y that are drawn from D-dimensional zero mean, unit variance Gaussian distributions with varying degrees of correlation i.e., $p(\mathbf{x}) = p(\mathbf{y}) = \mathcal{N}(0,\mathbf{I})$ and $p(\mathbf{x},\mathbf{y}) = \mathcal{N}(0,\Sigma)$. Let us denote, $\mathbf{z} = [\mathbf{x},\mathbf{y}]^T$. We are then interested in computing,

$$MI_{\chi^2}(X;Y) = \int\int \frac{p(\mathbf{x},\mathbf{y})^2}{p(\mathbf{x})p(\mathbf{y})}d\mathbf{x}d\mathbf{y} - 1 = \int \frac{\mathcal{N}(\mathbf{z} \mid 0,\Sigma)^2}{\mathcal{N}(\mathbf{z} \mid 0,\mathbf{I})}d\mathbf{z} - 1 \tag{18}$$

Rewrite as,

$$MI_{\chi^2}(X;Y) = \int \frac{\mathcal{N}(\mathbf{z} \mid 0,\Sigma)\mathcal{N}(\mathbf{z} \mid 0,\Sigma)}{\mathcal{N}(\mathbf{z} \mid 0,\mathbf{I})}d\mathbf{z} - 1 \tag{19}$$

Product of two Gaussian densities, $\mathcal{N}(\mathbf{z} \mid 0,A)\mathcal{N}(\mathbf{z} \mid 0,A) = K\mathcal{N}(\mathbf{z} \mid 0,A/2)$, with $K = (2\pi)^{-D/2}|2A|^{-1/2}$. Plugging this result in the above equation we get,

$$\begin{aligned} MI_{\chi^2}(X;Y) &= \int \frac{(2\pi)^{-D/2}|2\Sigma|^{-1/2}\mathcal{N}(\mathbf{z} \mid 0,\Sigma/2)}{\mathcal{N}(\mathbf{z} \mid 0,\mathbf{I})}d\mathbf{z} - 1, \\ &= (2\pi)^{-D/2}|2\Sigma|^{-1/2}\int \frac{|\Sigma/2|^{-1/2}e^{-0.5\mathbf{z}^T(\Sigma/2)^{-1}\mathbf{z}}}{e^{-0.5\mathbf{z}^T\mathbf{I}\mathbf{z}}}d\mathbf{z} - 1 \\ &= (2\pi)^{-D/2}|2\Sigma|^{-1/2}|\Sigma/2|^{-1/2}\int e^{-\frac{1}{2}\mathbf{z}^T(2\Sigma^{-1}-\mathbf{I})^{-1}\mathbf{z}}d\mathbf{z} - 1, \end{aligned} \tag{20}$$

Realizing that the quantity inside the integral is an unnormalized Gaussian, we get,

$$MI_{\chi^2}(X;Y) = |2\Sigma|^{-1/2}|\Sigma/2|^{-1/2}|(2\Sigma^{-1}-\mathbf{I})^{-1}|^{1/2} - 1 = \frac{|(2\Sigma^{-1}-\mathbf{I})^{-1}|^{1/2}}{|\Sigma|} - 1 \tag{21}$$

In Figure 1, we show the effect of the number of samples and data dimension on the CHI2MI-MC estimate for a synthetic example. The two variable X and Y are drawn from 1 and 10-D zero mean, unit variance Gaussian distributions with varying degrees of correlation. We can analytically compute the ground truth $MI_{\chi^2}(X;Y)$ in this case (the derivation and additional details can be found in the Appendix). CHI2MI-MC estimate converges to ground truth with increasing number of samples from the proposal distribution.

D. Discussion: $MI_{KL}(X;Y)$ and $MI_{\chi^2}(X;Y)$

Let us look at the IS estimate of the KL-Mutual Information $MI_{KL}(X;Y)$, again using a proposal distribution for the

joint distribution of the form $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y})p(\mathbf{x}|\mathbf{y})$:

$$\begin{aligned} MI_{KL}(X;Y) &= D_{KL}(P_{X,Y}||P_X P_Y) \\ &= \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\Big[\log\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\Big] \\ &= \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\Big[\log\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\Big] \\ &= \mathbb{E}_{q(\mathbf{x},\mathbf{y})}\Big[\log\Big[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\Big]\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\Big] \\ &= \mathbb{E}_{q(\mathbf{y})p(\mathbf{x}|\mathbf{y})}\Big[\log\Big[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\Big]\frac{p(\mathbf{y})}{q(\mathbf{y})}\Big] \end{aligned} \tag{22}$$

The above estimate still relies on potentially complex $p(\mathbf{y})$. On the other-hand, using the inequality $\log(\mathbf{x}) \leq \mathbf{x} - 1$ for $\mathbf{x} > 0$, it is easy to show that for any distribution $q(\mathbf{y})$, $\hat{MI}_{\chi^2}(X;Y)$ is an unbiased estimate of the upper bound to $MI_{KL}(X;Y)$,

$$\begin{aligned} MI_{KL}(X;Y) &\leq \mathbb{E}_{p(\mathbf{y})p(\mathbf{x}|\mathbf{y})}\Big[\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}\Big] - 1 \\ &= \mathbb{E}_{q(\mathbf{y})p(\mathbf{x}|\mathbf{y})}\Big[\frac{p(\mathbf{y}|\mathbf{x})}{q(\mathbf{y})}\Big] - 1 \\ &:= MI_{\chi^2}(X;Y) = \mathbb{E}_{q(\mathbf{x},\mathbf{y})}[\hat{MI}_{\chi^2}(X;Y)] \end{aligned} \tag{23}$$

E. Related Work and Background

Invariant representations and predictors that are sufficiently responsive to the task under consideration while being invariant to any confounding factors has long been considered as crucial for successful learning and generalization (Kolchinsky et al., 2019).

Fair Machine Learning: The goal of fair machine learning is to learn models that are invariant to sensitive attributes. Reliance of a model on these sensitive attributes can lead to unwanted bias towards a subpopulation. One approach to algorithmic fairness is to learn representations from covariates X such that they do not contain information about the sensitive attributes (Zemel et al., 2013; Louizos et al., 2015). Such approaches are referred to as pre-processing methods. Post-processing methods (Hardt et al., 2016; Pleiss et al., 2017) learn to transform the predictions of a trained model to satisfy a measure of fairness. In-processing algorithmic fairness methodologies (Calmon et al., 2017; Kamishima et al., 2011) are designed to directly learn predictors $\hat{Y} = h_\theta(X)$ that possess the desired invariance.

Adversarial in-processing methods (Zhang et al., 2018; Barrett et al., 2019) involve learning an adversary $a_\phi(\cdot)$ that is trained to predict the sensitive variable from the predictions $\hat{Y}$. Fair predictors $h_\theta(\cdot)$ can be learned by minimizing risk $R(\theta)$ or the expected prediction error between the predicted outcome and ground truth outcome variable captured by loss $l(y, h_\theta(\mathbf{x}))$, typically negative log likelihood, while maximizing risk $R(\phi)$ or the expected prediction error between the predicted sensitive attribute and ground truth sensitive variable captured by loss $l(s, a_\phi(\hat{y}))$. This leads to solving the following alternating optimization over the parameters θ and ϕ.

$$\begin{aligned} &\arg\min_{\phi} \mathbb{E}_{p(\mathbf{x},s)}[l(s, a_\phi(h_\theta(\mathbf{x})))] \\ &\arg\min_{\theta} \mathbb{E}_{p(\mathbf{x},y)}[l(y, h_\theta(\mathbf{x}))] - \lambda\mathbb{E}_{p(\mathbf{x},s)}[l(s, a_\phi(h_\theta(\mathbf{x})))] \end{aligned} \tag{24}$$

Recently, (Roy & Boddeti, 2019) showed that maximizing the entropy of the adversary that is trying to predict the sensitive attribute was more effective than maximising the adversary's loss, as done in adversarial debiasing.

Domain Generalization: In domain adaptation and generalization, the domain or environment of the data samples can be understood as the sensitive attributes. Domain adaptation methods typically assume access to labeled or unlabeled samples from the test domain. The invariance is achieved by matching or aligning the feature distribution across domains (Fernando et al., 2013; Ganin & Lempitsky, 2015; Yan et al., 2017; Shu et al., 2018). Domain generalization is a more challenging scenario where no samples from the test/target domain are observed during training. Adversarial approaches (Motiian et al., 2017; Maria Carlucci et al., 2019), with access to multiple training domains, learn domain agnostic representations that successfully fool an adversary that is trying to predict the source domain of the samples from these representations. The recently proposed Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) paradigm aims to achieve domain generalization by learning representations that render a classifier built on these representations invariant across domains or environments.

Measuring and Minimizing Dependence: Several recent approaches (Ozair et al., 2019; Lopez-Paz et al., 2013; Moyer et al., 2018; Greenfeld & Shalit, 2019) propose to measure or enforce these independence relations by minimizing directly or indirectly a dependence measure between the variables such as MI_{KL} or Hilbert Schmidt Independence Criterion (HSIC). It can be easily shown that maximising the entropy of the adversary's loss, as done in (Roy & Boddeti, 2019), is equivalent to minimising the upper-bound of MI_{KL} in (Barber & Agakov, 2004) by using a distribution that is uniform over the support $u(y)$ as the variational approximation $q(y)$.

$$\begin{aligned} MI_{KL}(X;Y) &\leq \mathbb{E}_{p(\mathbf{x},y)}\Big[\log\frac{p(y \mid \mathbf{x})}{u(y)}\Big] \\ &= \mathbb{E}_{p(\mathbf{x})}[KL(p(y|\mathbf{x})||u(y)] \end{aligned} \tag{25}$$

Most similar to our work, in (Mary et al., 2019) the authors formulate a constrained optimization problem that aims to

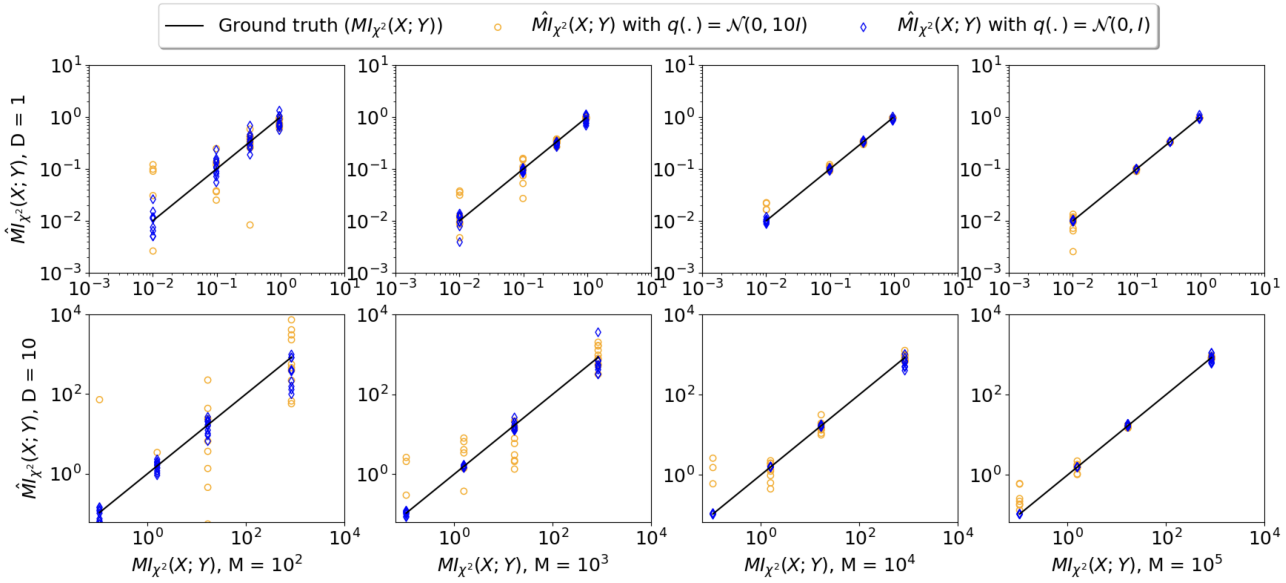

Figure 1. Effect of the number of samples, data dimension and the proposal distribution on the CHI2MI-MC estimate. The plots show the estimates over 10 runs with varying levels of ground truth chi-square mutual information. M corresponds to the number of samples from the proposal distribution used for the estimate and D corresponds to the dimension of the variables.

enforce independence by ensuring that the HGR dependency measure is below a small ϵ. The HGR constraint is relaxed and instead the L_1 norm of the χ^2-divergence between the joint and the marginal distributions is minimized. The densities are estimated using a Kernel Density Estimator (KDE) and the divergence is computed over a mesh grid which can scale poorly with data dimensions and grid granularity. We address these issues with the CHI2MC estimate of the chi-square mutual information.

F. Demographic Parity and Equality of Odds Estimation

Demographic Parity is enforced by minimizing the estimate $\hat{MI}^{\phi}_{\chi^2}(S; h_\theta(X))$. When Y and S are discrete and binary, following Equation 16 we can estimate $MI^{\phi}_{\chi^2}(D; h_\theta(X))$ as

$$\hat{MI}^{\phi}_{\chi^2}(S; h_\theta(X)) = \frac{1}{N_{j=0}} \sum_{i=0}^{N_{j=0}-1} p_\phi(S=0|h_\theta(\mathbf{x}_{i,j=0})) + \frac{1}{N_{j=1}} \sum_{i=0}^{N_{j=1}-1} p_\phi(S=1|h_\theta(\mathbf{x}_{i,j=1})) - 1 \quad \text{where } \mathbf{x}_{i,j} \sim p(\mathbf{x}|S=j) \tag{26}$$

Here, the estimate involves learning the parameters ϕ of conditional distribution $p_\phi(s|\hat{y})$.

Equality of Odds is enforced by minimizing the estimate $\hat{MI}^{\phi}_{\chi^2}(S; h_\theta(X)|Y)$. Using the learned conditional distribution $p_\phi(s|\hat{y}, y)$, the conditional mutual information is estimated as:

$$\hat{MI}^{\phi}_{\chi^2}(S; h_\theta(X)|Y) = \frac{1}{2} \sum_{k=0}^{1} \sum_{j=0}^{1} \frac{1}{N_j} \sum_{i=0}^{N_j-1} p_\phi(S=j \mid h_\theta(\mathbf{x}_{i,j,k}), Y=k) - 1 \quad \text{where } \mathbf{x}_{i,j,k} \sim p(\mathbf{x}|S=j, Y=k) \tag{27}$$

G. DIVA for Domain Generalization

DIVA starts with the following variational lower bound per input $\mathbf{x}$ for the model:

$$\begin{aligned}\mathcal{L}(d, \mathbf{x}, y) = \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{x}) q_{\phi_d}(\mathbf{z}_d|\mathbf{x}) q_{\phi_x}(\mathbf{z}_x|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_l, \mathbf{z}_d, \mathbf{z}_x)] \\ - \beta KL(q_{\phi_y}(\mathbf{z}_y|\mathbf{x})||p(\mathbf{z}_y|y)) \\ - \beta KL(q_{\phi_d}(\mathbf{z}_d|\mathbf{x})||p(\mathbf{z}_d|d)) \\ - \beta KL(q_{\phi_x}(\mathbf{z}_x|\mathbf{x})||p(\mathbf{z}_x))\end{aligned} \tag{28}$$

To encourage the separation of the domain and label-specific latent spaces, the authors additionally learn classifiers to predict label y and domain d from $\mathbf{z}_y$ and $\mathbf{z}_d$, respectively. This leads to the following modification to the lower bound:

$$\begin{aligned}\mathcal{F}_{DIVA}(d, \mathbf{x}, y) = \mathcal{L}(d, \mathbf{x}, y) + \alpha_y \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{x})}[\log q_{w_y}(y|\mathbf{z}_y)] \\ + \alpha_d \mathbb{E}_{q_{\phi_d}(\mathbf{z}_d|\mathbf{x})}[\log q_{w_d}(d|\mathbf{z}_d)]\end{aligned} \tag{29}$$

The semi-supervised variant of DIVA that was found to be more challenging than the fully supervised setting. Here, the variational lower-bound is given as:

$$\begin{aligned}\mathcal{L}_u(d, \mathbf{x}, y) = \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{x}), q_{\phi_d}(\mathbf{z}_d|\mathbf{x}) q_{\phi_x}(\mathbf{z}_x|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_l, \mathbf{z}_d, \mathbf{z}_x)] \\ - \beta KL(q_{\phi_d}(\mathbf{z}_d|\mathbf{x})||p(\mathbf{z}_d|d)) \\ - \beta KL(q_{\phi_x}(\mathbf{z}_x|\mathbf{x})||p(\mathbf{z}_x)) \\ + \beta \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{x}) q_{w_y}(y|\mathbf{z}_y)}[p(\mathbf{z}_y|y) - q_{\phi_y}(\mathbf{z}_y|\mathbf{x})] \\ + \mathbb{E}_{q_{\phi_y}(\mathbf{z}_y|\mathbf{x}) q_{w_y}(y|\mathbf{z}_y)}[p(\mathbf{y}) - q_{w_y}(y|\mathbf{z}_y)]\end{aligned} \tag{30}$$

Assuming N labeled samples and M unlabeled samples, the semi-supervised objective is given as,

$$\begin{aligned}\mathcal{F}_{SS-DIVA} := \sum_{n=1}^{N} \mathcal{F}_{DIVA}(d_n, \mathbf{x}_n, y_n) + \sum_{m=1}^{M} \mathcal{L}(d_m, \mathbf{x}_m, y_m) \\ + \alpha_d \mathbb{E}_{q_{\phi_d}(\mathbf{z}_d|\mathbf{x}_m)}[\log q_{w_d}(d_m|\mathbf{z}_d)]\end{aligned} \tag{31}$$

H. Training Details for Domain Generalization Experiment

We provide the training details that are specific to DIVA+CHI2MC here and refer to relevant sections in the appendix section of DIVA (Ilse et al., 2019) for the common aspects. Instead of searching the best hyperparameters, we fix them to default values found in the code to see the effect of CHI2MC regularization.

Rotated MNIST: The training procedure and DIVA architecture details can be found in section 5.1.1 and 5.1.2 of (Ilse et al., 2019). We set $\alpha_d = 2000.0$, $\alpha_y = 4200.0$ and $\beta_{max} = 1.0$.

Malaria Cell Images: The training procedure and DIVA architecture details can be found in section 5.2.2 and 5.1.4 of (Ilse et al., 2019). We set $\alpha_d = 100000.0$, $\alpha_y = 75000.0$ and $\beta_{max} = 1.0$.

For both datasets, the dimensions of $\mathbf{z}_d$, $\mathbf{z}_y$ and $\mathbf{z}_x$ are set to 64. We alternate the training of the DIVA model parameters and the CHI2MC estimators parameters ($\phi_{z,d}$) for 1 steps each. We use a Gaussian distribution for $p_{\phi_{z,d}}(\mathbf{z}_d|\mathbf{z}_y) = \mathcal{N}(a_{\phi_{z,d}}(\mathbf{z}_y), I)$. The architecture of $a_{\phi_{z,d}}(\cdot)$ consists of a fully-connected network with a single hidden layer of dimension 20. The parameters $\phi_{z,d}$ are learned by maximising the likelihood $p_{\phi_{z,d}}(\mathbf{z}_d|\mathbf{z}_y)$. We use ADAM optimizer with default setting for all the parameters.