Stabilizing Adversarial Invariance Induction by Discriminator Matching

Anonymous authors
Paper under double-blind review

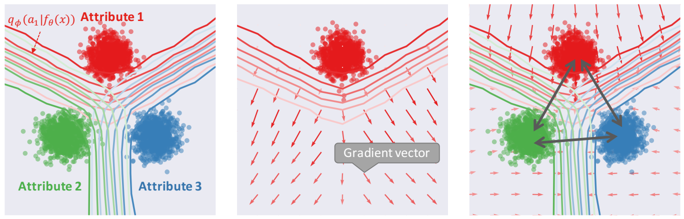
Abstract

Incorporating the desired invariance into representation learning is a key challenge in many situations, e.g., for domain generalization and privacy/fairness constraints. An adversarial invariance induction (AII) shows its power on this purpose, which maximizes the proxy of the conditional entropy between representations and attributes by adversarial training between an attribute discriminator and feature extractor (Xie et al., 2017). However, the practical behavior of AII is still unclear as the previous analysis assumes the optimality of the attribute classifier, which is rarely held in practice. This paper first analyzes the practical behavior of AII both theoretically and empirically, indicating that AII has theoretical difficulty as it maximizes variational *upper* bound of the actual conditional entropy, and AII catastrophically fails to induce invariance even in simple cases as suggested by the above theoretical findings. We then argue that a simple modification to AII can significantly stabilize the adversarial induction framework and achieve better invariant representations. Our modification is based on the property of conditional entropy; it is maximized if and only if the divergence between all pairs of marginal distributions over z between different attributes is minimized. The proposed method, *invariance induction by discriminator matching*, modify AII objective to explicitly consider the divergence minimization requirements by defining a proxy of the divergence by using the attribute discriminator. Empirical validations on both the toy dataset and four real-world datasets (related to applications of user anonymization and domain generalization) reveal that the proposed method provides superior performance when inducing invariance for nuisance factors.

1 Introduction

Extensive studies have demonstrated that deep neural networks (DNNs) can uncover complicated variations in data to provide powerful representations that are useful for classification tasks (Hinton et al., 2006; Krizhevsky et al., 2012). However, in some scenarios, the learned representation should be invariant to some attribute of the input data. A motivating example is a task called domain generalization (Blanchard et al., 2011), which requires learning a domain-invariant representation that applies to unseen domains (e.g., the data of an unseen user or different image sources). When practitioners apply DNNs to data that include a large amount of user information (such as images with usernames (Edwards & Storkey, 2016) or data from wearables (Iwasawa et al., 2017)), the desired representations should not include user identifying information. For legal and ethical reasons, machine learning algorithms must make fair decisions that are independent of sensitive variables such as gender, age, or race (Louizos et al., 2016). Therefore, this study aims to answer the following question: how can we systematically incorporate the desired invariance into representation learning?

Invariance induction is a systematic solution to this problem, which often introduces an additional regularization term that measures the level of invariance. One theoretically sound metric is the conditional entropy of attributes given representations of data, denoted by $H(a|z)$ where a and z denote the random variables of attributes and representations, as it is maximized if and only if the representations are invariant to attributes. However, exact calculation is intractable as it requires integration over high dimensional z, and $p(a|z)$ is unknown in general.



(a) $q_\phi(a|z)$ (b) AII (c) Proposal

Figure 1: Comparison between the previous method and the proposed method. The detailed description of this dataset is shown in Section 3. The arrow in (b) and (c) represents the gradient vector of each method. (a) Both methods utilize the learned critic function $q_\phi(a|z)$, which represented as counterplot in the figure. (b) AII update f_θ by keeping a distribution away from the decision boundary. Note that, only the information from the decision boundary is used in AII. (c) Our proposal minimize the proxy of divergence between distributions, which uses $q_\phi(a|z)$. The proxy minimization keeps the distribution near the decision boundary, as indicated by the gradient vector fields.

Recently, Xie et al. (2017) introduced a sensible way to approximate the conditional entropy, which uses an external attribute classifier to estimate the conditional distribution and uses the estimates to regularize an encoder f_θ, where θ represents the parameter of the neural network encoder. Mathematically, the method optimizes following min–max game:

$$\min_\theta \max_\phi \mathbb{E}[\log q_\phi(a|z{=}f_\theta(x))], \tag{1}$$

where $q_\phi(a|.)$ is a conditional probability distribution approximated by an attribute classifier parameterized by ϕ. Since f_θ and q_ϕ is an adversarial relationship, this framework is called *adversarial invariance induction (AII)*. Intuitively speaking, the optimization push representations to the region where learned classier can not correctly distinguish the attributes (as shown in Figure 1-a, b). As Xie et al. (2017) demonstrated the game possesses an equilibrium where the encoder maximizes the true conditional entropy, under the assumption that $q_\phi(a|z)$ correctly estimates the true $p(a|z)$. A similar approach was extensively used in domain generalization, fair-prediction, and privacy-protection contexts (Edwards & Storkey, 2016; Motiian et al., 2017; Xie et al., 2017; Iwasawa et al., 2017).

Build upon the above achievements, we first analyze the practical behavior of AII both theoretically and empirically. Note that, although the connection to the conditional entropy is intuitive, the connection does not guarantee the practical behavior of AII, as the assumption of $q_\phi(a|z) = p(a|z)$ is rarely holds in practice, and the interpretation does not give any indication of whether the gradient for the encoder is informative or not. This result suggests that optimization of AII, without the assumption of the optimality of $q_\phi(a|z)$, does not need to maximize the conditional entropy, as the maximizing (unbounded) upper bound does not gives any guarantees. We also empirically verify that AII catastrophically fails to induce invariance even in simple situations as suggested by the above theoretical findings.

We then argue that the simple modification to AII attains better property from the optimization perspective while achieving the same goal asymptotically. Our modification is based on the property of the conditional entropy, i.e., it is maximized if and only if the divergence between all pairs of marginal distributions over z between different attributes are minimized. This property suggests that the invariance induction algorithm should also minimize the divergence, which is not considered on AII and possibly induce unstable behavior. In contrast, our proposed method, *invariance induction by discriminator matching (IIDM)*, explicitly considers the divergence minimization requirements by minimizing the proxy of the divergence between the marginals, which push the representations with different attributes are recognized similarly by the discriminator (as shown in Figure 1-c). Thanks to the consideration of the divergence minimization requirements, IIDM try to keep the distributions near the decision boundary, which allvaite the possibility that $q_\phi(a|z)$ gives wrong information. We discuss the relationship among our proxy, the divergence between the marginals, and non-saturating heuristics used in different but related community (Goodfellow et al., 2014).

The main contributions of this paper can be summarized as follows. (1) We highlight the practical issues of AII, which is a state-of-the-art framework for invariance induction, both theoretically and empirically (Section 3). (2) We propose a modification to AII by explicitly consider the property of the maximum conditional entropy (Section 4). Empirical validations on both the toy dataset and four real-world datasets demonstrate that the proposed method provides superior performance and faster convergence at the level of invariance induction. For example, in the experiments on toy dataset, the proposed method converges to mostly the maximum conditional entropy, while the optimization of AII is catastrophically unstable.

2 Related Works

The goal of the invariance induction is to learn attribute-invariant spaces given a training dataset made of tuples of $\{(x_n, y_n, a_n)\}_{n=1}^{N}$ (supervised setting) or pairs of $\{(x_n, y_n)\}_{n=1}^{N}$ (referred to as unsupervised setting (Jaiswal et al., 2018)), where x_n is an observation and y_n is a target of x_n, $a_n \in \mathcal{A}$ is a realization of a categorical random variable a. In particular, we focus on the supervised settings as it has a broader range of applications, such as fairness aware machine learning and privacy preservation. Assume f_θ is a function (encoder) that parameterized by a neural network, which maps observations to representations $z \in \mathcal{Z}$. A natural way to induce invariance is to define differentiable dependency measurements between attribute and representations of the data and use it as a regularization term to learn f_θ: $\mathbb{E}\left[L(f_\theta(x_n), y_n) + \lambda V(f_\theta(x_n), a_n)\right]$, where a λ is the weighting parameter, V somehow measures invariance of the representations regarding the attribute, and L is a loss function that represents how much information about y is present in the representations.

Adversarial invariance induction (AII) is a recently proposed approach for measuring the V by an external neural network. That is, if the external network can accurately predict a from $z = f_\theta(x)$, AII considers z to have considerable information about a. The external neural network is often called a discriminator or adversary in this context. Information from the discriminator is used to update the weights of the encoder f_θ so that the updated representations have less information about a (Eq. 1). The merit of this framework is that it does not depend on the pre-defined metrics, such as L2 distance and maximum mean discrepancy (Zemel et al., 2013; Li et al., 2014; Louizos et al., 2016). However, this advantage often comes at the cost of the optimization difficulty, which motivates us to identify the cause of the difficulty and to develop better methods.

Another related area of our work is generative adversarial networks (GAN) (Goodfellow et al., 2014) and domain adversarial networks (DAN) (Ajakan et al., 2014; Gan et al., 2016). Initially, this approach has a similar min–max formulation with our adversarial invariance induction framework. However, we rarely use its naive min–max optimization in practice. Instead, they often incorporate several heuristics, e.g., non-saturating heuristics in GAN, and asymmetric mapping in domain adaptation (Tzeng et al., 2017), which gives the same fixed points but have better convergence property. These works motivate us to replace the min–max game of the adversarial invariance induction problem by deriving alternative interpretations of adversarial invariance induction frameworks.

To this ends, this paper proposes discriminator matching, which derived from the divergence minimization interpretation of conditional entropy maximization. The derived algorithm relates the proposal in (Jolicoeur-Martineau, 2018a), which point out the misconception that regards generative adversarial networks as divergence minimization framework, and modifications to recover the divergence minimization interpretations. (Jolicoeur-Martineau, 2018a), which point out the misconception that regards generative adversarial networks as divergence minimization framework, and modifications to recover the divergence minimization interpretations. In a similar spirit, (Jolicoeur-Martineau, 2018b) presents relativistic discriminator training. Note that, our scenario differ with GAN and DAN in many ways, e.g., (1) we typically need to match the several distributions, and (2) there are no explicit target distribution (data distribution in GAN and source data distribution in DAN), which makes the applications of heuristics in GAN and DAN into our scenario nontrivial. We focus on the problem of invariance induction and show that the invariance induction problem can be interpreted as divergence minimization involving multiple distributions, and this interpretation is significant in practice.

3 Analysis on Adversarial Invariance Induction

In the remainder of this paper, we assume the use of alternating optimization to solve the adversarial game. At each iteration, AII first updates the discriminator κ times by minimizing the negative log-likelihood $\mathbb{E}\left[-\log q_{\phi}(a = a_n | f_{\theta}(x_n))\right]$. AII then updates the encoder by maximizing the negative log-likelihood while fixing q_{ϕ}. In supervised classification case, the encoder is trained with a classifier h_{ψ} by mixing the adversarial and supervised objectives (typically cross-entropy) to ensure that the representations are useful for predicting target variables.

The advantage of this approach is that it avoids using pre-defined metrics, which often ease computation but reduce efficiency. It learns the critic function itself by approximating measurements that are hard to compute but possibly increase efficiency. In the case of invariance induction, the learned critic function approximates the conditional entropy between z and a. Formally, AII objective is the variational upper bound of the conditional entropy:

$$\begin{aligned} H(a|z) = \mathbb{E}_{p_{\theta}(z,a)}\left[-\log p(a_n|z_n)\right] &= \mathbb{E}_{p_{\theta}(z_n,a_n)}\left[-\log q_{\phi}(a_n|z_n)\right] - D_{KL}(p(a|z)||q_{\phi}(a|z)) \\ &\leq \mathbb{E}_{p_{\theta}(z_n,a_n)}\left[-\log q_{\phi}(a_n|z_n)\right] = H_q(a|z), \end{aligned}$$

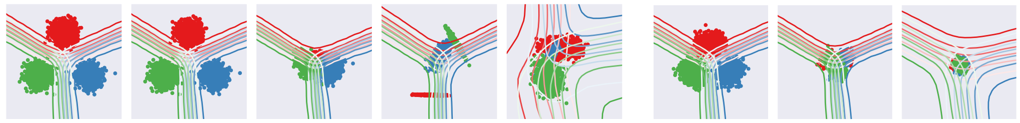
where D_{KL} represents the Kullback–Leibler (KL) divergence, which is greater than zero by definition, and $p_{\theta}(z)$ is an empirical distribution of representations. The bound is tight when $D_{KL}(p(a|z)||q_{\phi}(a|z)) = 0$. Maximizing $H_q(a|z)$ yields the encoder update in AII framework exactly. The updation of the discriminator is equivalent to minimizing this KL divergence; therefore, the update of the discriminator tighten the bound. As (Xie et al., 2017) explained, if the discriminator correctly approximates the true conditional distribution $p(a|z)$, the objective of the updation of f_{θ} is equal to maximizing $H(a|z)$. Xie et al. (2017) also demonstrated that the min–max game has an equilibrium, at which f_{θ} maximizes the conditional entropy $H(a|z)$.

Unfortunately, it is not practical to assume the optimality of $q_{\phi}(a|z)$ due to the limitation of the capacity of the ϕ and computational costs to increase κ. It means, in a more practical case where $q_{\phi}(a|z)$ is not optimal, AII maximize the *upper* bounds of the conditional entropy. In general, maximizing the upper bound of the function of interest f does not guarantee the minimizing the f (in our case, f is the conditional entropy). Moreover, in this case, $H_q(a|z)$ itself is not upper-bounded, as assigning zero probability for correct pair of z, a makes $H_q(a|z)$ infinity. Note that, the above observations do not immediately mean that maximizing $H_q(a|z)$ is useless, rather it implies that there are regions where maximizing $H_q(a|z)$ does not increase $H(a|z)$. For example, in the most exaggerated case, moving the entire distribution in a specific direction increases the upper bound while the actual conditional entropy remains constant.

To clarify the problem of AII, we design the toy dataset, which comprises samples from several Gaussian distributions with different means ($[\sin(\frac{i}{K}\pi), \cos(\frac{i}{K}\pi)]$, and $i \in 1, 2, \cdots K$) and the same variance, assuming that each distribution corresponds to different attributes. We test AII on this synthesized dataset while varying the number of distributions K and κ (number of updates of ϕ per iteration). Specifically, we first train the discriminator 100 times with a batch size of 128 and update q_{ϕ} and f_{θ} iteratively using stochastic gradient descent with a learning rate=0.1. Figure 2-(b) visualizes how distributions move during the optimization of AII on synthesized data ($K = 3$ and κ=1). Each figure corresponds to a different timestep of the alternating optimization. The red line in Figure 3-a represents the quantitative results on this configuration, where solid lines represent the approximated conditional entropy by using a post-hoc classifier q_{eval}, which is parameterized by the neural network having the same architecture as that of q_{ϕ}. For reference, the theoretical maximum value of the conditional entropy (gray line) and the negative log-likelihood of q_{ϕ} (dashed lines) are also depicted. Figure 3-(b) compares the performances of AII on different configurations of $\{K, \kappa\}$, where color denotes different K, marker denotes different κ, and the dashed line denotes optimal values.

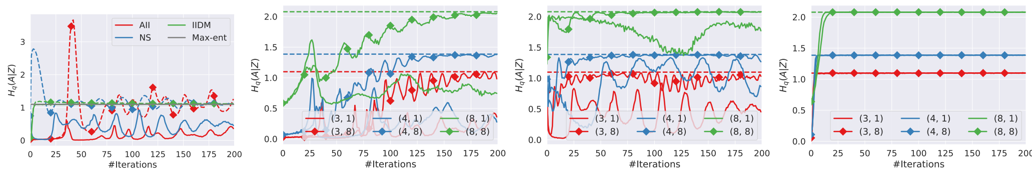
As shown in the visualization of AII and the vibration of the red lines in Figure 3-a, AII is catastrophically unstable, even such a simple case. Figure 3-a also indicates that, at some points (e.g., around 50 iterations), the approximated conditional entropy becomes significantly large, whereas the true conditional entropy remains constant or even decreases. The issue may be alleviated if the discriminator has a sufficiently large capacity and is trained many times at each iteration, as suggested by a faster convergence of AII when $\kappa = 8$, but this is an impractical assumption. Note that the toy dataset is much simpler than real datasets, and it is fair to say that identifying the supremum is more challenging.

It is noteworthy that the above formulation resembles the original formulation of GAN (Goodfellow et al., 2014), which is never used practically. Although no works in the invariance induction community have been explicitly considered yet, one can derive a slightly better objective by transferring



(a) AII (1, 3, 35, 40, 200 steps each) (b) IIDM (1, 3, 200 steps each)

Figure 2: Visualizing behaviors of AII and IIDM on toy datasets. (a) The behavior of AII after different iterations. (b) The behavior of IIDM after different iterations. The following url shows gif version of this visualization: `https://drive.google.com/open?id=1N2kuTCfwjQBGgv3dxTSD9oBklAenLj-J`.



(a) $K = 3$, $\kappa = 1$ (b) AII (c) NS (d) IIDM

Figure 3: Quantitative comparison of AII, NS and IIDM (proposed method) on the toy datasets.

the non-saturating heuristic used in the GAN via label flipping:

$$V_{NS}(\theta; \phi) := \mathbb{E}_{p_\theta(z,a)}\left[-\sum_{a_j \neq a_n} \log q_\phi(a_j|z_n)\right], \tag{2}$$

which also enhances the discriminator to misclassify the attribute of the data. In the remainder of this paper, we refer to this version as the *non-saturating version* and denote it by NS. As discussed in the GAN community, this objective has the same fixed points as the original objective but provides meaningful gradients because it not saturated even if the discriminator is a supremum. However, it is unclear how it relates to the conditional entropy maximization. Later, we discuss the connection between the proposal and this objective, and how our proposal attains better properties while inheriting the non-saturating nature of the NS objective.

4 Invariance Induction by Discriminator Matching

4.1 Property of Maximum Conditional Entropy

The above theoretical and empirical analyses suggest that maximizing the approximated conditional entropy by using AII framework does not ensure the maximization of the true conditional entropy, as the former is the upper bound of the latter. We derive an alternative approach to maximize the conditional entropy by rethinking it from a divergence minimization perspective. Specifically, we assume that A is a uniform categorical random variable. The maximum conditional entropy holds the following property:

Proposition 1. *The maximum conditional entropy* $H(a|z)$ *is* $-\log\frac{1}{K}$, *and* $H(a|z)$ *is maximized if and only if* $p(z|a_i) = p(z|a_j)$ *for all* $a_i \neq a_j \in \mathcal{A}$ *and* $z \in \mathcal{Z}$.

The proof is shown in the Appendix A. This proposition means that maximizing conditional entropy is asymptotically equal to minimizing pairwise divergence. Specifically, assume D is a divergence measurement over a space of a possible probability distribution, i.e., positive-definitive function ($D(p||q) \geq 0$ for all p, q and $D(p||q) = 0$ if and only if $p = q$). For simplicity, we denote the marginal distribution over a random variable z associated with an ttribute $a = a_i$ as $p^i_\theta(z)$. Then, the following corollary holds.

Corollary 1. *If* f_θ *gives an attribute invariant representation (i.e., conditional entropy* $H(a|z)$ *is maximized), then* $D(p^i_\theta(z)||p^j_\theta(z)) = 0$ *for all* a_i, a_j *and vice versa.*

The above analysis suggests that the invariance induction algorithm should also minimize the divergence between marginal distributions of different attributes a. However, the original AII does not consider such a constraint as AII keeps the distribution $p_\theta^i(z)$ away from the non-desired point where a discriminator correctly predicts the attribute but does not ensure that it approaches some target distribution $p_\theta^{j\neq i}(z)$.

4.2 Discriminator Matching

Based on the analysis, we propose simple modifications to AII considering the divergence minimization requirements. In the remainder of the paper, we assume to use KL-divergence as divergence measurements, though it could be a design choice in practice. Similar to AII, IIDM is based on the alternating training of attribute classifier q_ϕ and feature extractor f_θ, but IIDM deceives the discriminator differently. Using the conditional distribution $q_\phi(a|z)$, we denote $q_\phi^i(a) = \int p_\theta^i(z)q_\phi(a|z)dz$. Formally, IIDM minimizes the following discriminator matching loss for all pairs of a_i and $a_j \neq a_i$:

$$V_{dm}(p_\theta^i(z)||p_\theta^j(z);\phi) := \mathbb{E}_{z_j\sim p_\theta^j(z)}\left[D_{KL}(q_\phi^i(a)||q_\phi(a|z_j))\right]. \tag{3}$$

Minimizing the Eq. 3 with respect to the encoder parameter θ push $q_\phi(a|z_j)$ to $q_\theta^i(a)$, where z_j is drawn from p_θ^j and $q_\theta^i(a)$ represents the average discriminator's perception over the sample from attribute $a_i \neq a_j$. In other words, IIDM deceives a discriminator by ensuring that the representations from different attributes are recognized equally by the discriminator.

IIDM's objective V_{dm} explicitly relates to the divergence $D_{KL}(p_\theta^i(z)||p_\theta^j(z))$. Specifically, the following inequality holds:

$$D_{KL}(p_\theta^i(z)||p_\theta^j(z)) \geq \mathbb{E}_{z_j\sim p_\theta^j(z)}\left[D_{KL}(q_\phi^i(a)||q_\phi(a|z_j))\right]. \tag{4}$$

We first use data processing inequality $D_{KL}(p_\theta^i(z)||p_\theta^j(z)) \geq D_{KL}(q_\phi^i(a)||q_\phi^j(a))$ (Gerchinovitz et al., 2017; Barber et al., 2018), and then use Jensen's inequality. The detailed proof is shown in the Appendix B.

Intuitively speaking, the divergence minimization perspective restricts the update of the encoder to consider the location of the marginal distributions of different attributes, in addition to the decision boundary (which is all the information source of the AII). Note that, the inequality indicates that the $V_{dm}(p_\theta^i(z)||p_\theta^j(z);\phi)$ is the lower bound of the $D_{KL}(p_\theta^i(z)||p_\theta^j(z))$, so minimizing the former one does not ensure the minimization of the later one in general. The problem seems the same as that in the case of AII; however, the proposed method still has the advantage that $V_{dm}(\theta,\phi)$ is also lower bounded (by zero), so minimizing it does not induce catastrophic failures. In contrast, AII's objective is not upper-bounded and maximizing it causes catastrophic behavior, as shown in the simulation results. In addition, we empirically found that IDM try to keep the distributions near the decision boundary, which alleviate the possibility that $q_\phi(a|z)$ gives wrong information as shown in Figure 1-c. In a special case, if the discriminator is invertible, $V_{dm}(\theta,\phi) = 0$ ensures that $p_\theta^i(z) = p_\theta^j(z)$, as shown in (Barber et al., 2018) Restricting the invertibility of the discriminator is an interesting direction, but we did not add such a regularization as restricting neural networks is difficult in general and open research areas (Jacobsen et al., 2018; Behrmann et al., 2018; Ardizzone et al., 2018). Instead, we empirically validate that the proposed method reliably learns invariant representations without such a regularization.

It is worth mentioning that the proposed method is closely related to the non-saturating version of AII. Specifically, the proposed method minimizes

$$\begin{aligned}\mathbb{E}_{z_j\sim p_\theta^j(z)}\left[D_{KL}(q_\phi^i(a)||q_\phi(a|z_j))\right] &= \mathbb{E}_{z_j\sim p_\theta^j(z)}\left[\sum_{a\in\mathcal{A}} q_\phi^i(a)\log\frac{q_\phi^i(a)}{q_\phi(a|z_j))}\right] && (5)\\ &= \mathbb{E}_{z_j\sim p_\theta^j(z)}\left[\sum_{a\in\mathcal{A}} -q_\phi^i(a)\log q_\phi(a|z_j))\right] + C, && (6)\end{aligned}$$

and the NS objective (Eq. 2) can be rewritten as

$$V_{NS}(\theta;\phi) = \sum_{a_j\in\mathcal{A}}\sum_{a_j\neq a_i}\mathbb{E}_{z_j\sim p_\theta^j(z)}\left[\sum_{a\in\mathcal{A}} -p^i(a)\log q_\phi(a|z_j))\right], \tag{7}$$

where $p^i(a)$ is equal to 1.0 for $a = a_i$ and zero otherwise. Ignoring the constant term, the only difference is whether to push $q_\phi(a|z_j)$ to average perception by a discriminator of a different attribute $q^i_\phi(a)$, or true marginal $p^i(a)$. Because the discriminator matching loss considers $q_\phi(a \neq a_j|z_j)$, similar to the NS objective, its gradient does not vanish even if q_ϕ is the supremum, as with the NS loss. This property encourages faster convergence, especially when the representations have a significant amount of information regarding attributes. It differs from NS as it explicitly considers the divergence minimization requirements, while NS does not as it only considers the current decision boundary of the discriminator but does not consider information from $p^i_\theta(z)$ directly (similar to AII). Therefore, NS push representations as far as possible as long as it successfully changes the discriminator's prediction regardless of whether it aligns marginal distributions.

To clarify the benefit of the proposal, we test NS and IIDM on the synthesized datasets. All the experimental settings are identical to those of the simulation in Section 3. Figure 2-(b) visualizes how distributions move during the optimization of IIDM on synthesized data ($K = 3$ and κ=1). The quantitative results on this configuration are represented in Figure 3-a, where color represents the different methods (red: AII, blue: NS, and green: IIDM). Figure 3-(c, d) compares the performances of NS and IIDM on different configurations of $\{K, \kappa\}$, where color denotes different K, marker denotes different κ, and the dashed line denotes the optimal values. We can make the following observations. (1) NS is superior to AII, but it still unstable near the optimal point, as indicated by the vibration of the plots. (2) IIDM consistently reaches the theoretical maximum values in all configurations, as depicted in Figure 3-d.

One implementation choice is how to calculate $q^i_\phi(a)$ in Eq. 6. The straightforward approach is through Monte Carlo approximation: $q^i_\phi(a) = \int p^i_\theta(z)q_\phi(a|z)dz = \mathbb{E}_{p^i_\theta(z)}[q_\phi(a|z)]$. Although it is an unbiased estimation, the variance is large if the number of samples is small. The average can be calculated from all samples (or a sufficiently large number of samples from each K attributes) at every iteration, but it requires additional computation other than standard mini-batch estimation. We address these issues by using the moving centroid mechanism. Specifically, instead of estimating $q^i_{\theta,\phi}(a)$ every time with sufficiently large samples, the proposed method stores the moving average of discriminator perceptions:

$$Q^i_t(a) = \gamma Q^i_{t-1}(a) + (1-\gamma)q^i_t(a), \tag{8}$$

where Q^j_{t-1} is a previous centroid, $q^i_t(a)$ is the new estimation of the centroid based on a single batch, and γ is the decay parameter for controlling the speed at which the centroids change. We initialized Q^j_0 by computing the centroids of all training data points.

Then we can use the standard mini-batch method to calculate

$$\mathcal{V}_{IIDM}(\theta;\phi) = \mathbb{E}_{z_n,a_n \sim p_\theta(z,a)}\left[\sum_{a_i \neq a_n}[\mathcal{D}_{KL}(Q^i(a)||q_\phi(a|z_n))\right]. \tag{9}$$

Finally, the encoder and the classifier are trained by minimizing:

$$\min_{\theta,\psi}\big[L(\theta,\psi) + \lambda V_{IIDM}(\theta;\phi)\big], \tag{10}$$

where ψ is the parameter of the classifier, L is any classification loss, and λ is a weighting parameter.

4.3 Analysis

Remark 1. Continuous nuisance attribute: One may question the applicability of our method to the continuous attribute case, e.g., the goal is to learn age-invariant representations. Firstly, the treatment of continuous R.V. is an open question. The most straightforward yet practically used answer is to discretize the continuous R.V. For example, (Xie et al., 2017) divides an age variable into two groups. We can apply the proposed method with similar discretization.

Remark 2. Semantic alignments: One well-known problem of invariant feature learning is determining how to incorporate semantic alignment, i.e., how to align only the pair of samples that have the same semantic (typically, the target label). For this purpose, (Li et al., 2018b) proposes adversarial training based semantic alignment method, which prepares multiple domain classification networks where each classifier specialize for each class label y. Another merit of the proposed method is that we can enforce semantic alignment with a simple modification, without any additional computational costs. Individually, semantic alignment can be carried out by merely computing the centroids for each (attribute, label) tuple and aligning the perceptions of $\{x, y, a\}$ between only centroids of the same label $y' = y$ but different attributes $a' \neq a$. Since most applications of invariant feature learning require that L_y is also minimized, we also test this modification for all the later-described experiments.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

In addition to the simulation results, we provide experimental results on two tasks (four datasets) relevant to invariant feature learning: (1) user anonymization (Opportunity and USC datasets), and (2) domain generalization (MNISTR and PACS datasets). All experiments were implemented in PyTorch and were run on GPUs (either GTX 1080 or Tesla V100).

For user anonymization tasks, Opportunity and USC datasets were used. This task is to learn anonymized representations (z that does not contain user-identifiable information) while maintaining classification performance. The **Opportunity** dataset (Sagha et al., 2011) consists of sensory data regarding human activity in a breakfast scenario. Each record consists of 113 real-value sensory readings, excluding time information. We considered the task of recognizing 18 classes[1]. Following previous studies (Yang et al., 2015; Iwasawa et al., 2017), we use a sliding window procedure with 30 frames and a 50% overlap. The number of samples was 57,790 in total. We parameterize the encoder using convolutional neural networks (CNN) with three convolution-ReLU-pooling repetitions followed by one fully connected layer and classification by logistic regression, following previous studies (Yang et al., 2015; Iwasawa et al., 2017). The discriminator is a simple feedforward network with 800–400 hidden units. The **USC-HAD** dataset is another activity recognition dataset that consists of 14 subjects (Zhang & Sawchuk, 2012). The data include 12 activity classes[2] that correspond to people's most essential and daily activities. MotionNode, which is a 6 DOF inertial measurement unit, is used to record the output from accelerometers that record six real sensory values. The sliding window procedure, with 30 frames and a 50% overlap, produced 172,169 samples.

The MNISTR and PACS are two typical datasets of domain generalization tasks. The **MNISTR** dataset, derived from MNIST, was introduced by (Ghifary et al., 2015). Its labels comprise the ten digits; domains are created by rotating the images in multiples of 15 degrees: 0, 15, 30, 45, 60, and 75. The domains are labeled with the angle by which they are rotated, e.g., M15 and M30. Each image is cropped to 16×16 pixel in accordance with a previous study (Ghifary et al., 2015)[3]. Similar to (Ghifary et al., 2015), we used two convolution layers with 32 and 48 filters of 5×5 kernels, followed by a max-pooling layer and two fully connected layers with 100 hidden units. A discriminator with 100 hidden units is connected to the output of the first fully connected layer. The **PACS** dataset is a relatively new benchmark dataset designed for cross-domain recognition (Li et al., 2017). The dataset has 9991 images in total across seven categories (dog, elephant, giraffe, guitar, house, horse, and person) and four domains of different stylistic depictions (photo, painting, cartoon, and sketch). The diverse depiction styles provide a significant domain gap. We use the ImageNet pre-trained AlexNet CNN (Krizhevsky et al., 2012) as a base network, following previous studies(Li et al., 2017; 2018a). A discriminator with 1024 hidden units is connected to the output of the last fully connected layer.

Baselines: To demonstrate the efficacy of the proposed method, we compared it with the following methods. (1) A **CNN** trained on the aggregation of data from all source domains. Although there are special treatments for domain generalization, (Li et al., 2017) reports that CNN outperforms many domain generalization methods on the PACS dataset. (2) **AII** (Xie et al., 2017), is a main baseline. (3) **AII+GP** uses a variant of AII with an additional gradient penalty regularization used in GAN

[1] open door 1, open door 2, close door 1, close door 2, open fridge, close fridge, open dishwasher, close dishwasher, open drawer 1, close drawer 1, open drawer 2, close drawer 2, open drawer 3, close drawer 3, clean table, drink from cup, toggle switch, and null

[2] walking forward, walking left, walking right, walking upstairs, walking downstairs, running forward, jumping, sitting, standing, sleeping, elevator up, and elevator down

[3] Specifically, we used the dataset distributed at `https://github.com/ghif/mtae`.

Table 1: Performance comparison of user anonymization tasks. The value is the lowest user-classification accuracy with specific performance degradation $(0.01, 0.03$ points) from CNN.

dataset	Opp-S1		Opp-S2		Opp-S3		Opp-S4		USC	
threshold	0.01	0.03	0.01	0.03	0.01	0.03	0.01	0.03	0.01	0.03
CNN	0.939	0.939	0.973	0.973	0.984	0.967	0.983	0.983	0.683	0.683
AII	0.631	0.517	0.590	0.590	0.694	0.659	0.589	0.586	0.512	**<u>0.179</u>**
AII+GP	0.619	0.619	0.521	0.521	**0.471**	**0.471**	0.673	0.510	0.580	0.569
NS	0.635	0.452	0.614	0.523	0.484	0.484	**0.499**	**0.482**	None	None
IIDM	**<u>0.462</u>**	**<u>0.417</u>**	**<u>0.415</u>**	**<u>0.415</u>**	**<u>0.409</u>**	**<u>0.409</u>**	**<u>0.486</u>**	0.486	**0.499**	0.499
IIDM+	**0.502**	**0.433**	**0.474**	**0.474**	0.495	0.495	0.631	**<u>0.461</u>**	**<u>0.478</u>**	**0.478**

(Mescheder et al., 2018). (4) **RevGrad** is a slightly modified version of AII, which uses the gradient reversal layer (Ganin et al., 2016) to train all the networks (encoder, classifier, and discriminator) at the same time. (5) **NS** is a non-saturating version of AII introduced in section 3 of this paper. (6) **CrossGrad** (Shankar et al., 2018) is regarded as a state-of-the-art method in domain generalization tasks. Note that it does not intend to learn invariant representation, so we use CrossGrad only for comparing domain generalization performance. (7) **IIDM** is our proposal. We used the gradient penalty as well.

Optimization: For all datasets and methods, we used RMSprop for each optimization. For all datasets except PACS, we set the learning rate to 0.001 and the batch size to 128. For PACS, we set the learning rate to $5e-5$ and the batch size to 64. The number of iterations was $10k$, $5k$, $20k$, $30k$, and $50k$ for MNISTR, PACS, Opp, and USC, respectively. For a fair comparison, hyperparameters were tuned on a validation set for each baseline. For the adversarial-training-based method, we optimized weighting parameter λ from $\{0.001, 0.01, 0.1, 1.0\}$, except for MNISTR, for which it was optimized from $\{0.01, 0.1, 1.0, 10.0\}$. The value of α for CrossGrad was selected from $\{0.1, 0.25, 0.5, 0.75, 0.9\}$. Unless mentioned otherwise, we set the decay rate γ to 0.7 for all experiments.

Evaluation: In all the experiments, we selected the data of one or several domains for the test set and used the data of a disjoint domain as the training/validation data. Accurately, we split the data of the disjoint domain into groupings of 80% and 20%. We denote the test domain by a suffix (e.g., MNISTR-M0 denotes that the model is trained with the data from M15, M30, M45, M60, and M75 and evaluated on M0). We conducted 20 validations during training at equal intervals. In each validation, we measured the label classification accuracy (Y-acc) and the level of invariance. We empirically measured the level of invariance by training a post-hoc classifier D_{eval} that tries to predict a over learned representations, following previous studies (Xie et al., 2017; Iwasawa et al., 2017). Specifically, we trained the classifier with 800 hidden units $1k$ iterations (by RMSprop optimizer, with a learning rate of 0.001 and a batch size of 128) with the data that are used to train the encoder and evaluate attribute classification accuracy on the validation dataset.

5.2 Results

User anonymization: Table 1 compares the user-anonymization performance. The value represents the lowest user-classification accuracy (the lower the better) with specific performance degradation compared to CNN on classification accuracy. For example, the columns with 0.01 represent the lowest user-classification accuracy with less than 0.01 point performance degradation. The best performance is underlined and highlighted in bold, and the second-best performance is only highlighted in bold. IIDM+ represents the variants that use the semantic alignment extension introduced in section 4. The results show the clear benefit of our proposal to inducing user invariance. Specifically, IIDM performs best on seven out of ten configurations. Note that the value with 'None' represents the method always reduce the label classification performance significantly.

Domain generalization: Table 2 summarizes the classification performance on two different datasets: MNISTR, and PACS. The top row of each table represents the test domain. We report the mean accuracy as well as the standard error of three different seeds (five seeds for MNISTR and three seeds for PACS). The best performance is underlined and highlighted in bold, and the second-best performance is only highlighted in bold. We can make the following observations. (1) IIDM and IIDM+ demonstrate the best or comparable performance on all conditions except sketch domain. Although the semantic alignment extension does not help the performance on a simpler

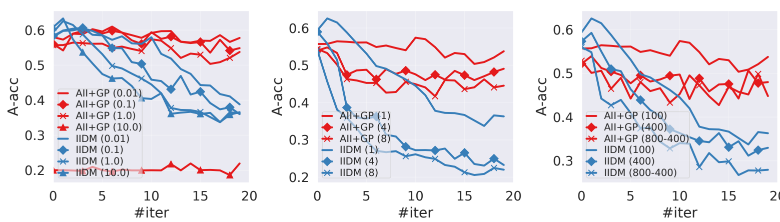
Table 2: Classification accuracies on unseen domains.

(a) MNISTR

	M0	M15	M30	M45	M60	M75	Avg
CNN	84.0± 1.7	**99.1± 0.5**	97.6± 0.9	91.9± 1.8	97.5± 0.5	87.7± 1.7	92.97
RevGrad	84.4± 1.6	98.8± 0.2	97.9± 0.8	92.1± 0.8	95.7± 2.2	85.9± 4.7	92.45
AII	83.8± 2.1	98.5± 0.4	97.4± 0.9	91.0± 1.4	97.0± 0.4	87.4± 2.4	92.52
AII+GP	86.2± 1.4	98.5± 0.2	97.9± 0.5	91.2± 0.7	97.0± 0.9	**87.9± 2.0**	93.11
CrossGrad	85.3± 0.9	**98.9± 0.5**	97.6± 0.8	90.9± 1.0	**98.2± 0.4**	87.5± 2.0	93.09
IIDM	**88.0± 1.6**	98.2± 1.0	**98.1± 0.7**	**94.3± 0.8**	98.0± 0.7	**88.9± 1.3**	**94.25**
IIDM+	**88.3± 0.9**	98.6± 0.5	**98.1± 0.6**	**93.0± 1.8**	**98.1± 0.9**	86.9± 2.5	**93.85**

(b) PACS

	photo	art	cartoon	sketch	Avg
CNN	80.8± 1.3	58.1± 2.6	62.7± 2.6	**60.6± 4.5**	65.57
RevGrad	82.9± 1.3	57.2± 1.9	61.6± 0.6	54.6± 4.6	64.06
AII	81.1± 0.7	59.1± 1.7	60.7± 3.1	**62.1± 3.0**	65.75
AII+GP	81.8± 0.4	60.7± 0.2	**64.0± 2.1**	60.6± 3.3	66.76
CrossGrad	81.4± 1.8	58.1± 4.7	60.5± 3.1	60.5± 1.3	65.15
IIDM	**82.9± 1.2**	**61.7± 1.5**	63.4± 0.7	59.5± 0.5	**66.89**
IIDM+	**84.8± 0.6**	**62.3± 1.6**	**64.8± 1.5**	60.2± 2.5	**68.04**



(a) λ (b) κ (c) architecture of ϕ

Figure 4: Comparison of AII and IIDM with different configurations on MNISTR dataset (M0 as test domain). The number in parenthesis represents the corresponding configuration.

task (MNISTR), it improves the performance on PACS dataset, giving approximately 1.0 point performance gain. (2) RevGrad and AII often fail to improve performance even when compared with a standard CNN. The score of AII+GP suggests that gradient penalty helps to improve the performance, but the improvements are lower than our proposal. (3) The Wilcoxon rank-sum test shows that IIDM is statistically better than CNN, RevGrad, AII, AII+GP, and CrossGrad with $p < 0.01$.

Figure 4 compares AII and IIDM on different (a) weighing parameter γ, (b) the number of the discriminator update κ, and (c) the network architecture of the discriminator. The dataset used is MNISTR with M0 as a test domain. In each figure, color represents a different method (red: AII, blue: IIDM) and marker denotes different configurations. The value represents the attribute classification accuracy (the lower the better invariant) by a post-hoc classifier $q_{eval}(a|z)$. For λ we used 1.0 by default. For κ and the architecture, we used default setting described in Section 5.1. The results show that our proposal is consistently to learn better invariant representations regardless of the choice of the hyperparameters. These results suggest that our proposal is better than searching such hyperparameters. Note that, $\lambda = 10.0$ for AII seems to attain better invariance, it was degenerated to the random representations and gives a random performance on the classification of y.

6 Conclusion

This paper presents a new method for invariance induction, called IIDM, by analyzing and extending current state-of-the-art adversarial invariance induction framework. This paper first examines the instability issue of AII both theoretically and empirically, indicating that AII has theoretical difficulty as it maximizes variational *upper* bound of the actual conditional entropy, and this fact leads AII to catastrophically fails even in simple cases. We then argue that a simple modification to AII can significantly stabilize the adversarial induction framework and achieve better invariant representations. The fundamental principle of our proposal is that a desirable invariance induction algorithm should also minimize the divergence between marginal distribution $p(z)$ between different attributes, as it is a requirement of maximum conditional entropy (Corollary 1) and missing in AII optimization. IIDM minimizes *discriminator matching* loss (Eq. 3), which is a proxy of the divergence between the marginals (Eq. 4). On toy dataset, we compare our proposal with the adversarial invariance induction framework, and show that our proposal significantly stabilizes the optimization (Figure 2 and Figure 3). Two real-world tasks (user-anonymization in Table 1 and domain generalization in Table 1) also supports that our proposal achieve better invariance induction.

References

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. In *Annual Conference on Neural Information Processing Systems (NIPS) Workshop on Transfer and Multi-Task Learning: Theory meets Practice*, 2014.

Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*, 2018.

David Barber, Mingtian Zhang, Raza Habib, and Thomas Bird. Spread divergences. *arXiv preprint arXiv:1811.08968*, 2018.

Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. *arXiv preprint arXiv:1811.00995*, 2018.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NIPS*, pp. 2178–2186, 2011.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *ICLR*, 2016.

Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 17(1):2096–2030, 2016.

Sebastien Gerchinovitz, Pierre Ménard, and Gilles Stoltz. Fano's inequality for random variables. *arXiv preprint arXiv:1702.05985*, 2017.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2551–2559, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 2672–2680, 2014.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Yusuke Iwasawa, Kotaro Nakayama, Ikuko Eguchi Yairi, and Yutaka Matsuo. Privacy issues regarding the application of dnns to activity-recognition using wearables and its countermeasures by use of adversarial training. In *Proceedings of the IJCAI*, pp. 1930–1936, 2017.

Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. *arXiv preprint arXiv:1802.07088*, 2018.

Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised adversarial invariance. In *Advances in Neural Information Processing Systems*, pp. 5092–5102, 2018.

Alexia Jolicoeur-Martineau. Gans beyond divergence minimization. *arXiv preprint arXiv:1809.02145*, 2018a.

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551. IEEE, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018a.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018b.

Yujia Li, Kevin Swersky, and Richard Zemel. Learning unbiased features. *arXiv preprint arXiv:1412.5244*, 2014.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair auto encoder. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, pp. 3478–3487, 2018.

Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-Shot Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems(NIPS)*, pp. 6673–6683. Curran Associates, Inc., 2017.

Hesam Sagha, Sundara Tejaswi Digumarti, José del R Millán, Ricardo Chavarriaga, Alberto Calatroni, Daniel Roggen, and Gerhard Tröster. Benchmarking classification techniques using the opportunity human activity dataset. In *Proceedings of the INSS*, 2011.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations (ICLR)*, 2018.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *International Joint Conference on Artficial Intelligence*, pp. 3995–4001, 2015.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 325–333, 2013.

Mi Zhang and Alexander A Sawchuk. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 1036–1043. ACM, 2012.

A PROOF OF THE PROPOSITION 1

Proof. Using the Lagrange multiplier method, the derivative of

$$L = -\sum_{a\in\mathcal{A}} p(a,z)\log p(a|z) + \lambda(1 - \sum_{a\in\mathcal{A}} p(a|z)) \tag{11}$$

is equal to zero for the maximum entropy $H(a|z)$. Solving the simultaneous equations, we can say $p(a_1|z) = p(a_2|z) = \cdots = p(a_K|z) = \frac{1}{K}$ for all $z \in \mathcal{Z}$ when the conditional entropy is maximized, and based on the definition, the conditional entropy becomes $-\log\frac{1}{K}$.

From Bayes' law and the assumption of uniform assumption of $p(a)$, $p(z|a_i) = p(z|a_j)$ holds $\forall a_i \neq a_j \in \mathcal{A}$ and $z \in \mathcal{Z}$. $\square$

B PROOF OF EQUATION 4

The proof use the data processing inequality for f-divergence and Jensen's inequality.

Theorem 1. *Consider a channel that produces y given x based on the $p(y|x)$. For any f-divergence $D_f(.||.)$*

$$D_f(p(y)||q(y)) \leq D_f(p(x)||q(x)) \tag{12}$$

As the KL divergence is also the family of the f-divergence, by replacing the $p(y|x)$ to $q_\phi(a|z)$, $p(x)$ to $p^i_\theta(z)$, $q(x)$ to $p^j_\theta(z)$, $p(y)$ to $q^i_\phi(a)$, and $p(y)$ to $q^j_\phi(a)$,

$$D_{KL}(p^i_\theta(z)||p^j_\theta(z)) \geq D_{KL}(q^i_\phi(a)||q^j_\phi(a)). \tag{13}$$

By expanding the KL divergence and uses $q^j_\phi(a) = \mathbb{E}_{p^j_\theta(z)}[q_\phi(a|z)]$,

$$\begin{aligned} D_{KL}(q^i_\phi(a)||q^j_\phi(a)) &= \sum q^i_\phi(a)\log q^i_\phi(a) - \sum q^i_\phi(a)\log \mathbb{E}_{p^j_\theta(z)}[q_\phi(a|z)] \\ &\geq \sum q^i_\phi(a)\log q^i_\phi(a) - \sum q^i_\phi(a)\mathbb{E}_{p^j_\theta(z)}\left[\log_\phi(a|z)\right] \\ &= \mathbb{E}_{z_j\sim p^j_\theta(z)}\left[D_{KL}(q^i_\phi(a)||q_\phi(a|z_j))\right]. \end{aligned}$$

Then,

$$D_{KL}(p^i_\theta(z)||p^j_\theta(z)) \geq E_{z_j\sim p^j_\theta(z)}\left[D_{KL}(q^i_\phi(a)||q_\phi(a|z_j))\right].$$