Stabilizing Adversarial Invariance Induction from Divergence Minimization Perspective

Yusuke Iwasawa , **Kei Akuzawa** and **Yutaka Matsuo**
The University of Tokyo, Japan
{iwasawa, akuzawa-kei, matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

Adversarial invariance induction (AII) is a generic and powerful framework for enforcing an invariance to nuisance attributes into neural network representations. However, its optimization is often unstable and little is known about its practical behavior. This paper presents an analysis of the reasons for the optimization difficulties and provides a better optimization procedure by rethinking AII from a divergence minimization perspective. Interestingly, this perspective indicates a cause of the optimization difficulties: it does not ensure proper divergence minimization, which is a requirement of the invariant representations. We then propose a simple variant of AII, called invariance induction by discriminator matching, which takes into account the divergence minimization interpretation of the invariant representations. Our method consistently achieves near-optimal invariance in toy datasets with various configurations in which the original AII is catastrophically unstable. Extensive experiments on four real-world datasets also support the superior performance of the proposed method, leading to improved user anonymization and domain generalization.

1 Introduction

Invariance to nuisance attributes is a desirable property of many representation learning tasks. A domain-invariant representation is key to building classifiers robust to domain changes [Muandet *et al.*, 2013]. When practitioners apply DNN to data that includes a large amount of user information, the desired representations should be invariant to it [Edwards and Storkey, 2016; Iwasawa *et al.*, 2017]. 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].

Adversarial invariance induction (AII) is a generic and powerful framework for enforcing invariance to nuisance attributes a into neural networks representations z [Xie *et al.*, 2017]. At the core of AII is the idea of using an external attribute classifier q_ϕ to measure the level of invariance, more specifically to approximate the conditional entropy $H(a|z)$. This approximated conditional entropy is then used to update the feature extractor f_θ, which corresponds to updating f_θ to deceive the external classifier. [Xie *et al.*, 2017] proved that, under the assumption of optimal q_ϕ, alternatively optimizing the attribute classifier and the feature extractor converges to the equilibrium where the feature extractor maximizes the true conditional entropy. As the true conditional entropy is maximized if and only if the representations are invariant against nuisance attributes, the procedure naturally ensures the invariance of the representations. A similar approach was extensively used in domain generalization, fairness-aware, and privacy-protection contexts [Edwards and Storkey, 2016; Motiian *et al.*, 2017; Xie *et al.*, 2017; Iwasawa *et al.*, 2017].

Despite the theoretical justification, the practical behavior of AII is still unclear and optimization of AII is often unstable. For example, consider the example of learning lighting invariant human identification tasks used by [Xie *et al.*, 2017]. They showed that adversary learned representations still contain significant information on lighting conditions (specifically, 0.53% prediction accuracy for five scenarios). [Moyer *et al.*, 2018] reported that AII is often overfitted; even if the feature extractor utterly defeats an attribute classifier, a post-hoc classifier can predict the attributes from the representations. Section 2.2 further highlights the instability issue using a custom designed toy dataset.

This paper approaches the instability of AII by rethinking it from a divergence minimization perspective. Here, this divergence minimization perspective refers to the approach of ensuring the invariance by aligning the representations associated with different attributes [Zemel *et al.*, 2013; Louizos *et al.*, 2016]. By formally connecting the goal of AII (maximizing the conditional entropy) and the divergence minimization perspective, we identify a fundamental misconception of the AII formulation; i.e.,*it does not ensure proper divergence minimization*, which is a requirement of the invariant representations. The lack of divergence minimization considerations explains several practical instabilities of AII. While this paper primarily focuses on the original version of AII, our findings are applicable to several extensions of it [Wang *et al.*, 2018; Jaiswal *et al.*, 2019].

We then present a modification to ensure divergence minimization under the adversarial invariance induction framework. As with AII, the proposed method leverages the power of the adversarial game between a feature extractor and an

external attribute classifier, but it deceives the external classifier differently; while AII directly maximizes the approximated conditional entropy, the proposed method implicitly maximizes it by forcing the representations with different attributes to be recognized similarly by the external classifier. We refer to the proposed method as *Invariance Induction by Discriminator Matching* as it induces invariance by matching the discriminator's output between different attributes. While the modification is simple and easy to implement, it attains better properties from the divergence minimization perspective and therefore gives significant performance gains. For example, our method consistently achieves near-optimal invariance in toy datasets with various configurations, in which AII catastrophically fails. Experiments on four real-world datasets (Opp and USC for user anonymization, and MNISTR and PACS for domain generalization) also support the superior performance of this proposal.

2 Instability Issue of Adversarial Invariance

2.1 Preliminary: Adversarial Invariance Induction

Assume we have a training dataset made of the tuples of (x, y, a), where $x \in \mathcal{X}$ is an observation, $y \in \mathcal{Y}$ is a target of x, $a \in \mathcal{A}$ is a nuisance attribute associated with x, and x, y, z are drawn from the true data distribution $p(x, y, a)$. For example, in the context of privacy-preserving activity recognition using data from wearables, the attribute a corresponds to some sensitive user information, and y is the activity label. In the context of domain generalization, learning representations invariant to domain shifts is a popular approach (the domain label is the attribute in that case).

Assume f_θ is a feature extractor parameterized by deep neural networks θ. Adversarial invariance induction (AII) optimizes the following min–max game:

$$\min_{\theta,\psi} \max_{\phi} \mathbb{E}_{p(x,y,a)}[-\log q_\psi(y|f_\theta(x)) + \lambda \log q_\phi(a|f_\theta(x))], \tag{1}$$

where $q_\psi(y|.)$ and $q_\phi(a|.)$ is a conditional probability distribution approximated by deep neural networks parameterized by ψ (for y) and ϕ (for a) respectively. The categorical attribute classifier $q_\phi(a|.)$ is often called a discriminator or adversary in this context. λ is the weighting parameter.

The above min–max game is closely related to the maximization of the conditional entropy of attributes given representations, $H(a|z)$. This is a good property of AII because $H(a|z)$ is a natural measurements of the invariance, as it is maximized if and only if z is invariant to a. Formally,

$$\begin{aligned} H(a|z) &= \mathbb{E}_{p(x,a)}\left[-\log p(a|f_\theta(x))\right] \\ &\leq \mathbb{E}_{p(x,a)}\left[-\log q_\phi(a|f_\theta(x))\right] = H_{p,q}(a|z), \end{aligned} \tag{2}$$

where $H_{p,q}(a|z)$ is the cross entropy between p and q. The bound is tight when $p(a|z) = q_\phi(a|z)$. Since $H_{p,q}(a|z)$ is equal to the second term of eq. 1 excluding the sign, minimizing eq. 1 with respect to θ is equivalent to maximizing the variational upper bound of the conditional entropy. By alternately (or jointly with the gradient reversal layer [Gan *et al.*, 2016]) optimizing θ and ϕ, this framework ensures the removal of nuisance information from the representations. 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 attribute classifier κ times by maximizing eq. 1 while fixing f_θ. AII then updates the feature extractor and label classifier by minimizing the eq.1 while fixing q_ϕ.

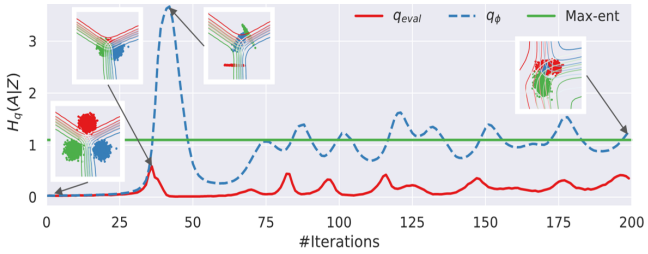


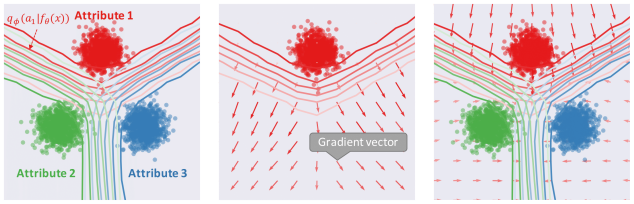
Figure 1: Conditional entropy estimates for a post-hoc classifier (red line) and the classifier used during training (blue dashed line). The gap suggests that AII uses inaccurate estimates during training. The learned representations are far from optimal invariance.

The relationship (eq. 2) is also used to empirically evaluate the level of invariance [Xie *et al.*, 2017]; i.e., $H_{p,q_{eval}}$ is used as an estimate of $H(a|z)$ where q_{eval} is a post-hoc classifier that is trained to predict a over the learned representations as correctly as possible. We uses same procedure to empirically evaluate the conditional entropy.

2.2 Instability of AII: Case Study

Before the main analysis, here we highlight the unstable behavior of AII using a toy dataset. The dataset comprises samples from K pieces of 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 apply AII on this dataset; q_ϕ, which predicts the distribution ID from the representations, is first updated 100 times with a batch size of 128, and q_ϕ and f_θ are alternately updated using stochastic gradient descent with a learning rate of 0.1. Figure 1 compares three values: (1) $H_{p,q_{eval}}(a|z)$ using a post-hoc classifier q_{eval}, (2) $H_{p,q}(a|z)$ using q_ϕ which is used during the training, and (3) the theoretical maximum value of the conditional entropy (green line). The post-hoc classifier q_{eval} is parameterized by a neural network with the same architecture as that of q_ϕ. For reference, the learned representations after several iterations (1, 35, 40, and 200 steps respectively) are also visualized. For simplicity, here we only show the results of $K = 3$. We later compare AII and our proposed method on various configurations.

The results indicate a significant mismatch between $H_{p,q_{eval}}(a|z)$ (red line) and $H_{p,q}(a|z)$ (dashed blue line). For example, around 40 iterations, the estimates using q_ϕ become significantly large but the estimates with a post-hoc classifier remain constant or even decrease. As $H_{p,q_{eval}}(a|z)$ is a more reliable estimate of the conditional entropy, the mismatch means that AII uses inaccurate estimates to update f_θ. Moreover, the estimates of conditional entropy using q_ϕ are often larger than the theoretical maximum value of $H(a|z)$. As such, the optimization of AII is unstable and does not maximize the conditional entropy $H(a|z)$ as expected.



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

Figure 2: Visualizing the gradient vector fields (the arrow in b and c) of AII and proposed method. (a) Both methods utilize $q_\phi(a|z)$, which represented as counterplot in the figure. (b) AII has an incentive to move the distribution away from decision boundary. (c) IIDM prevent the problem by considering both the decision boundary and information from $p_\theta(z|a_j)$ of different attributes.

2.3 Divergence Minimization Perspective

This paper examines the above instability issue from a divergence minimization perspective. Specifically, we consider the pairwise divergence among the conditional distribution of representations given different attributes. In the toy dataset used in Sec. 2.2, it corresponds to the distributional difference among red, green, and blue data points. We first formally link the divergence minimization perspective and the conditional entropy maximization problem, which is the goal of the formulation of AII. Assume D is the divergence over a space of probability distributions.

Proposition 1. *We assume that a is a uniform categorical random variable, and z is a random variable of the representations. If f_θ gives a representation that maximizes the conditional entropy $H(a|z)$, then $D(p_\theta(z|a_i)||p_\theta(z|a_j)) = 0$ for all a_i, a_j and vice versa. Here, the subscription denotes that the distribution depends on the feature extractor.*

Proof. The proposition can be derived by the following property of the conditional entropy: $H(a|z)$ 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 $z \in \mathcal{Z}$. Using the Lagrange multiplier, the derivative of

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

is equal to zero for the maximum entropy $H(a|z)$. Solving both 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 uniformity of $p(a)$, $p(z|a_i) = p(z|a_j)$ holds $\forall a_i \neq a_j \in \mathcal{A}$ and $z \in \mathcal{Z}$. □

This proposition means that maximizing conditional entropy is asymptotically equal to minimizing pairwise divergence. Unfortunately, the connection challenges AII, because it only pushes the distribution away from the decision boundary, without considering the divergence minimization requirement. Figure 2-(a, b) visualizes how AII updates the feature extractor using the toy dataset used in the previous section. The arrows in the figure represent the direction of the gradient when updating the f_θ using the data from $p_\theta(z|a = red)$ (updating the distribution $p_\theta(z|a = red)$). The gradient vector suggests that AII has an incentive to move the distribution far away, regardless of whether it aligns marginal distributions of different attributes. In other words, AII keeps the $p_\theta(z|a = red)$ away from the non-desired point where a discriminator correctly predicts the attribute, but does not ensure that it approaches some target distribution, such as $p_\theta(z|a = blue)$ or $p_\theta(z|a = green)$.

The lack of divergence minimization explains the undesired behavior of AII described in the previous section. For example, $H_{p,q}(a|z)$ reaches values larger than the theoretical maximum of the conditional entropy when AII moves the distribution without minimizing the divergence. Also, a significant mismatch between the approximated entropy $H_{p,q}(a|z)$ and $H_{p,q_{eval}}(a|z)$ can happen when the update of the feature extractor deceives the current discriminator without minimizing the divergence, or even increasing it.

3 Proposed Method

Based on this analysis, we hypothesize the lack of the divergence minimization perspective is a major cause of unstable behavior. Here, we describe a way to effectively incorporate the divergence minimization requirement into the adversarial invariance framework. Similar to the original AII, our method employs adversarial training between a feature extractor f_θ and a discriminator q_ϕ, but we deceive the discriminator differently. Specifically, we update f_θ by minimizing the expectation of the following discriminator matching loss:

$$\mathbb{E}_{z_j\sim p_\theta(z|a_j)} \sum_{i:a_i\neq a_j} \left[D_{KL}(q^i_{\theta,\phi}(a)||q_\phi(a|z_j))\right], \quad (3)$$

where $q^i_{\theta,\phi}(a) = \int p_\theta(z|a_i)q_\phi(a|z)dz$, which represents the average output of the discriminator given representations associated with attribute a_i. D_{KL} is the KL-divergence defined over the output of the discriminator:

$$D_{KL}(q^i_{\theta,\phi}(a)||q_\phi(a|z_j)) = \sum_{a\in\mathcal{A}} q^i_{\theta,\phi}(a)\log\frac{q^i_{\theta,\phi}(a)}{q_\phi(a|z_j)}. \quad (4)$$

We refer to this method as invariance minimization by discriminator matching (IIDM).

While the modification is simple, the objective has better interpretation from the perspective of divergence minimization. In contrast to AII, which updates the feature extractor by considering only the decision boundary, the discriminator matching loss also considers the location of the representations of different attributes. Figure 2 compares the gradient vector fields between AII and IIDM. The gradient vector fields indicate that our proposed method has no incentive to move the distribution far away from the decision boundary, as it violates the divergence minimization constraint. In other words, IIDM prevents the feature extractor to move the distribution regardless of whether it aligns marginal distributions of different attributes or not.

More formally, the objective of IIDM is related to the divergence $D_{KL}(p_\theta(z|a_i)||p_\theta(z|a_j))$ through the divergence between $D_{KL}(q^i_{\theta,\phi}(a)||q^j_{\theta,\phi}(a))$. Firstly, we can derive

$$D_{KL}(q^i_{\theta,\phi}(a)||q^j_{\theta,\phi}(a)) \leq \mathbb{E}_{z_j\sim p^j_\theta(z)}\left[D_{KL}(q^i_\phi(a)||q_\phi(a|z_j))\right],$$

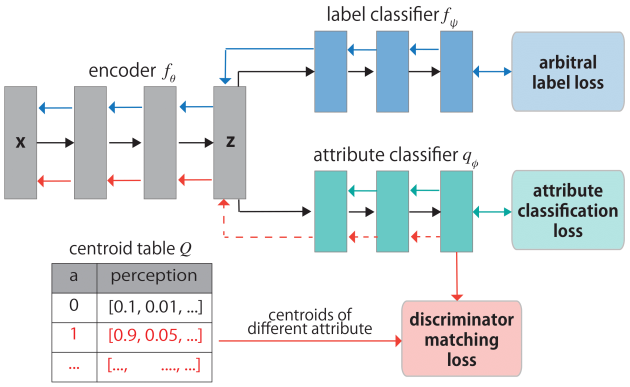


Figure 3: Overview of the proposed method (IIDM). IIDM force the representations with different attributes to be recognized similarly by the external classifier. Moving centroids techniques is used to reduce computational costs.

using the Jensen's inequality. Also, based on the data processing inequality of the f-divergence [Gerchinovitz *et al.*, 2017; Barber *et al.*, 2018], the following inequality holds:

$$D_{KL}(p_\theta(z|a_i)||p_\theta(z|a_j)) \geq D_{KL}(q^i_{\theta,\phi}(a)||q^j_{\theta,\phi}(a)).$$

The equality holds if the attribute classifier is invertible [Barber *et al.*, 2018]. In the special case, minimizing the discriminator matching loss ensures the divergence minimization as it is the upper bound of $D_{KL}(p_\theta(z|a_i)||p_\theta(z|a_i))$. Note that, the invertibility is similar to the optimality of the attribute classifier, which is often assumed in the analysis of the adversarial training. Ensuring inevitability is difficult in general and an open research areas in neural networks community [Behrmann *et al.*, 2018]. Instead, we empirically validate that the proposed method reliably learns invariant representations even without such a regularization.

One implementation issue is how to calculate $q^i_\phi(a)$ in Eq. 3. The straightforward approach is through Monte Carlo approximation: $q^i_\phi(a) = \mathbb{E}_{p_\theta(z|a_i)}[q_\phi(a|z)]$. 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. However, it requires additional computation other than standard mini-batch estimation. Moreover, the computation becomes intractable ($O(K^2)$) as the number of attribute values grows.

We address these issues by a moving centroid mechanism. Instead of estimating $q^i_{\theta,\phi}(a)$ every time with sufficiently large samples, IIDM uses the moving average of $q^i_{\theta,\phi}(a)$:

$$Q^i_t(a) = \gamma Q^i_{t-1}(a) + (1-\gamma)q^i_t(a), \tag{5}$$

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 using all training data points.

Then we can use the standard mini-batch method to calculate the discriminator matching loss. As with AII, IIDM incorporates alternating optimization. Specifically, IIDM firstly updates the attribute classifier q_ϕ by eq. 1, and updates the feature extractor and the classifier by

$$\min_{\theta,\psi} \mathbb{E}_{p(x,a,y)} \Bigg[-\log q_\psi(y|f_\theta(x)) + \lambda \left[\sum_{i:a_i \neq a} D_{KL}(Q^i(a')||q_\phi(a'|f_\theta(x))) \right] \Bigg], \tag{6}$$

where λ is a weighting parameter. Figure 3 shows a pictorial illustration of the proposed method.

4 Related Work

As briefly mentioned in the introduction of the paper, the formulation of AII (eq. 1) has theoretical groundings as a way to maximize the conditional entropy [Xie *et al.*, 2017]. While our analysis reveals the unstable behavior of AII, we do not intend to challenge the previous theoretical analysis. Building upon the theoretical grounding under ideal conditions, we examine the practical problem of AII by rethinking it from a divergence minimization perspective of invariance induction. We admit that the perspective itself is not brand new in these fields [Zemel *et al.*, 2013; Louizos *et al.*, 2016]; nevertheless, none of the studies has been linked AII with the divergence minimization perspective.

Similar to our work, several studies have proposed extensions of AII. For example, [Jaiswal *et al.*, 2019] proposes an adversarial forgetting mechanism to ensure the invariance, by introducing another network to produce forgetting masks over the representations. [Wang *et al.*, 2018] combines the adversarial invariance objective and a variational autoencoder to further enforce the invariance. As we have not introduced any architectural modifications, our method could be incorporated to the other extensions. Besides, our findings regarding the original formulation are applicable to several extensions that have the same objective.

As a extention of AII, several studies focus on the semantic alignment problem, i.e., how to align only the pair of samples that have the same semantics (the target label) [Li *et al.*, 2018b; Li *et al.*, 2018c]. Our method can be extended to consider semantic alignment without additional computational costs; semantic alignment can be carried out by computing the centroids for each (attribute, label) tuple and aligning the $q_\phi(a|z)$ of $\{x, y, a\}$ between only centroids of the same label $y' = y$ but different attributes $a' \neq a$. We test this variant in experiments below and denote it as IIDM+.

It is noteworthy that the above formulation resembles the original formulation of GAN [Goodfellow *et al.*, 2014] and domain adversarial networks (DAN) [Ajakan *et al.*, 2014; Gan *et al.*, 2016]. However, it is never used practically as it is known to be unstable and hard to optimize. This fact motivates us to replace the min–max game of the adversarial invariance induction problem. Although no prior work in the invariance induction community has been explicitly considered yet, one can transfer the non-saturating (NS) heuristic used in the GAN via label flipping. The NS-objective is similar to our method in the sense that it uses an asymmetric

objective. However, NS has same issue with AII as it only considers the current decision boundary and does not ensure the divergence minimization. We refer to this version as the *non-saturating version* and denote it by NS.

5 Experiments

5.1 Experimental Settings

Datasets

We provide experimental results on the synthesized dataset and two real world 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 either a GTX 1080 or Tesla V100.

For user anonymization tasks, Opportunity and USC datasets were used. This task is to learn anonymized representations (z that do not contain user-identifiable information) while maintaining classification performance. The **Opp** 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. 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. The feature extractor is parameterized by a convolutional neural network (CNN) with three convolution-ReLU-pooling repetitions followed by one fully connected layer and classification by logistic regression. 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 and 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 same sliding window procedure 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]. It comprises 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; Li *et al.*, 2018a]. A discriminator with 1024 hidden units is connected to the output of the last fully connected layer.

[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] We used the dataset distributed at https://github.com/ghif/mtae.

Baselines

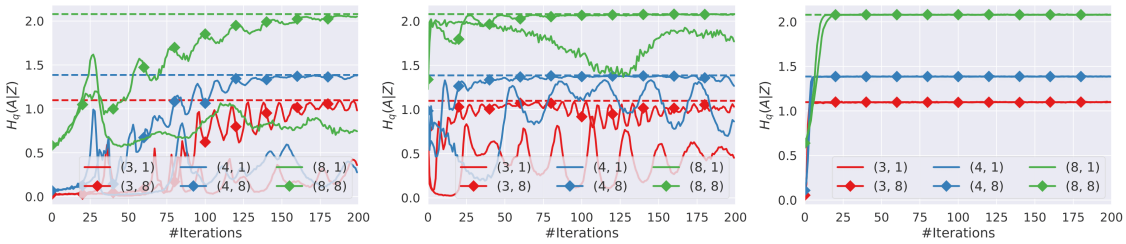
CNN trained on the aggregation of data from all source domains. Although there are special treatments for domain generalization, [Li *et al.*, 2017] reports that CNN outperforms many domain generalization methods on the PACS dataset. **AII** [Xie *et al.*, 2017] is the main baseline. **AII+GP** uses a variant of AII with an additional gradient penalty regularization used in GAN [Mescheder *et al.*, 2018]. **RevGrad** is a slightly modified version of AII, which uses the gradient reversal layer [Ganin *et al.*, 2016] to train all the networks (feature extractor, classifier, and discriminator) at the same time. **NS** is the non-saturating version of AII described in section 4 of this paper. **CrossGrad** [Shankar *et al.*, 2018], 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. **IIDM** is our proposal. We used the gradient penalty as well. We also tested the semantic alignment version and denoted it as **IIDM+**.

Optimization

For all datasets and methods, we used RMSprop for 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 $10k$, $5k$, $20k$, and $30k$ 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.

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. 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. For measuring the level of invariance, we trained a post-hoc classifier q_{eval} with 800 hidden units $1k$ iterations (by RMSprop optimizer, with a learning rate of 0.001 and a



(a) AII (b) NS (c) Proposal

Figure 4: Quantitative comparison of AII, NS and IIDM (proposed method) on the toy datasets with various configurations. IIDM consistently achieves near-optimal invariance.

batch size of 128) with the data that is used to train the feature extractor.

5.2 Results

Simulation

Figure 4 compares the performance of (a) AII (b) NS, and (c) IIDM on different configurations of $\{K, \kappa\}$, where color denotes different K, marker denotes different κ. The dashed line denotes optimal values. The experimental settings are the same as in Section 2.2. The results clearly support the benefits of the proposed method. (1) In all configurations, IIDM reaches the theoretical maximum values within the first few iterations. (2) In all configurations, AII and NS are unstable as shown in the vibration of the estimated conditional entropy. When $\kappa = 1$, AII and NS are far from the maximum conditional entropy. AII and NS with $\kappa = 8$ give much better performance, but its behabior is still unstable and convergence is significantly slower than the IIDM.

User Anonymization

Table 1 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 highlighted in bold. Note that the value 'None' represents the method always reducing the label classification performance significantly. As a results, IIDM performs best on seven out of ten configurations, suggesting the clear benefit of our proposal on the user anonymization task.

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 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+ show the best or comparable performance on all conditions except the sketch domain. Although the semantic alignment extension does not help with a simpler task (MNISTR), it improves the performance on the 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 the gradient penalty helps to improve 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 $p < 0.01$.

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	**0.179**
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	**0.462**	**0.417**	**0.415**	**0.415**	**0.409**	**0.409**	**0.486**	0.486	**0.499**	0.499
IIDM+	**0.502**	**0.433**	**0.474**	**0.474**	0.495	0.495	0.631	**0.461**	**0.478**	**0.478**

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.

Figure 5 compares AII and IIDM on different (a) weighing parameter γ, (b) the number of the discriminator updates 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) of a post-hoc classifier $q_{eval}(a|z)$. For λ we used 1.0 by default. For κ and the architecture, we used the default settings described in Section 5.1. The results show that our proposal consistently learns better invariant representations regardless of the choice of hyperparameters. These results suggest that our proposal is better than searching for such hyperparameters. Note that, $\lambda = 10.0$ for AII seems to attain better invariance, but it was degenerated to the random representations and gives a random performance on the classification of y.

6 Conclusion

This paper examines the optimization difficulty of AII (highlighted in Figure 1) and proposes a new method to attain invariance to nuisance attributes, by rethinking the AII's objective from a divergence minimization perspective. By formally linking the goal of AII with the pairwise divergence mini-

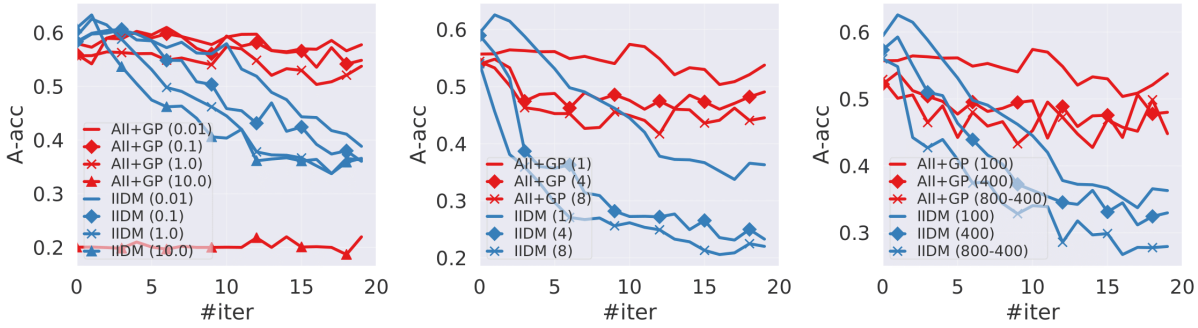
	M0	M15	M30	M45	M60	M75	Avg
CNN	84.0± 1.7	**99.1± 0.5**	97.6± 0.9	91.9± 1.8	97.5± 0.5	87.7± 1.7	92.97
RevGrad	84.4± 1.6	98.8± 0.2	97.9± 0.8	92.1± 0.8	95.7± 2.2	85.9± 4.7	92.45
AII	83.8± 2.1	98.5± 0.4	97.4± 0.9	91.0± 1.4	97.0± 0.4	87.4± 2.4	92.52
AII+GP	86.2± 1.4	98.5± 0.2	97.9± 0.5	91.2± 0.7	97.0± 0.9	**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**

(a) MNISTR

	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**

(b) PACS

Table 2: Classification accuracies on unseen domains.



(a) λ (b) κ (c) architecture of ϕ

Figure 5: Comparison of AII and IIDM with different configurations on MNISTR dataset (M0 as test domain). The number in parenthesis represents the corresponding configuration.

mization of conditional distribution of representations given attributes (Proposition 1), we identify a cause of its optimization difficulty; *it does not ensure proper divergence minimization*, which is a requirement of the invariant representations. We propose a simple way to effectively incorporate this requirement into the adversarial invariance framework, which leverages the power of the adversarial game but solves it more stably. Namely, the proposed method minimizes KL divergence defined over a space of attribute classifier's output (eq. 3), which is closely related to the divergence over the representation space, which we want to minimize. While the modification is easy to implement, it gives a significant performance gain; Our proposal consistently achieves near-optimal invariance in a toy dataset (Figure 4), where AII results in significantly unstable behavior. Our method is also good at user anonymization tasks (Table 1), and domain generalization tasks (Table 2). All these results suggest that the proposed method works well, and the divergence minimization interpretation introduced in this paper is significant.

References

[Ajakan *et al.*, 2014] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. In *NIPS Workshop on Transfer and Multi-Task Learning: Theory meets Practice*, 2014.

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

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

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

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

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMRL*, 17(1):2096–2030, 2016.

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

[Ghifary *et al.*, 2015] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559, 2015.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley,

Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Iwasawa *et al.*, 2017] 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 *IJCAI*, pages 1930–1936, 2017.

[Jaiswal *et al.*, 2019] Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. Invariant representations through adversarial forgetting. In *AAAI*, 2019.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5543–5551. IEEE, 2017.

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

[Li *et al.*, 2018b] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI*, 2018.

[Li *et al.*, 2018c] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018.

[Louizos *et al.*, 2016] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair auto encoder. In *ICLR*, 2016.

[Mescheder *et al.*, 2018] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, pages 3478–3487, 2018.

[Motiian *et al.*, 2017] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-Shot Adversarial Domain Adaptation. In *NIPS*, pages 6673–6683. Curran Associates, Inc., 2017.

[Moyer *et al.*, 2018] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Invariant representations without adversarial training. In *NIPS*, 2018.

[Muandet *et al.*, 2013] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18, 2013.

[Sagha *et al.*, 2011] 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 *INSS*, 2011.

[Shankar *et al.*, 2018] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018.

[Wang *et al.*, 2018] Ye Wang, Toshiaki Koike-Akino, and Deniz Erdogmus. Invariant representations from adversarially censored autoencoders. *arXiv preprint arXiv:1805.08097*, 2018.

[Xie *et al.*, 2017] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *NIPS*, 2017.

[Yang *et al.*, 2015] 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 *IJCAI*, pages 3995–4001, 2015.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, pages 325–333, 2013.

[Zhang and Sawchuk, 2012] Mi Zhang and Alexander A Sawchuk. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *UbiComp*, pages 1036–1043. ACM, 2012.