Distribution Alignment:

Data-Driven Constraints for Representation Learning

David I. Inouye

PURDUE UNIVERSITY®
Elmore Family School of Electrical and Computer Engineering

Current ML performs well but lacks important desired qualities

Optimistic perspective

- The next generation of ML will need to exhibit new desired properties.
  - Fairness – Are the predictions fair w.r.t. age or race?
  - Robustness – Can the model predict accurately even under new environment conditions?
  - Causality – Can the model estimate interventional or counterfactual queries?

Pessimistic perspective

- The *unintentional misuse* of data by algorithms underpins many problems in ML.
  - Fair ML – Avoids misuse of age or race in predictions.
  - Robust ML – Avoids misuse of spurious signals that will only work in one environment.
  - Causal ML – Avoids misuse of factual information to infer erroneous interventional or counterfactual information.

David I. Inouye, Purdue University

How can we impose these desired properties on ML systems?

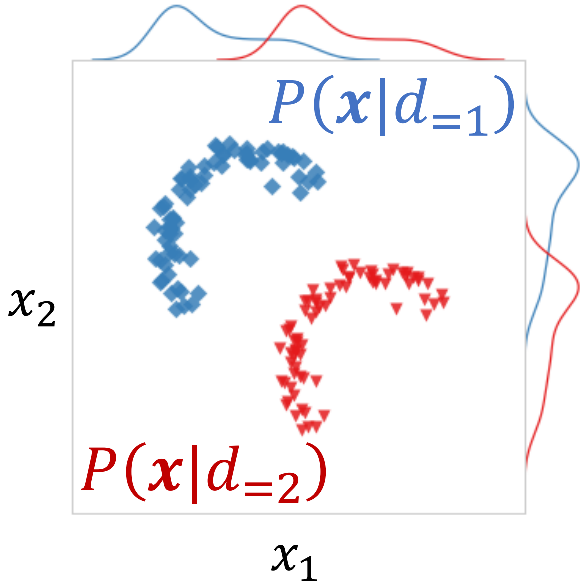
- Design model carefully (first wave of deep learning)
  - Improve inductive bias of model such as CNN and transformer architectures
  - Hand-design model to ensure specific property
    (e.g., graph models that are invariant to node permutations)

- Train bigger model with more data (second wave of deep learning)
  - Hope more data or computation will produce desired qualities
  - Yet, it is still unclear if this solves any of the prior problems or just hides them

- Explicitly enforce desired properties via distribution alignment (this talk ☺)
  - Broadly applicable to a wide range of problems
  - Property is *implicitly* defined by *domain labels*, which can be elicited from application expert

Distribution alignment is representation learning with the *opposite* objective of classification

Original Space



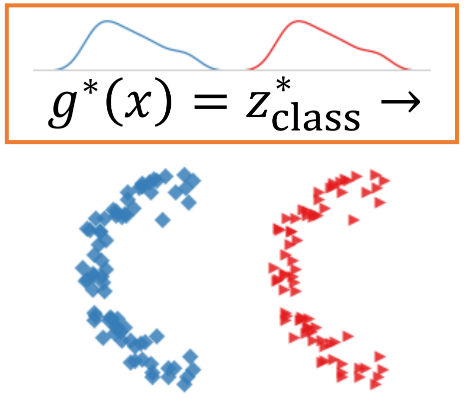
Representation Learning Objective

Classification

$$\max_{g \in \mathcal{G}} \phi\big(P(g(x)|d_{=1}), P(g(x)|d_{=2})\big)$$

where $g: \mathbb{R}^2 \to \mathbb{R}$ and ϕ is a distribution divergence (e.g., KL, JSD, W_2)

Latent Space

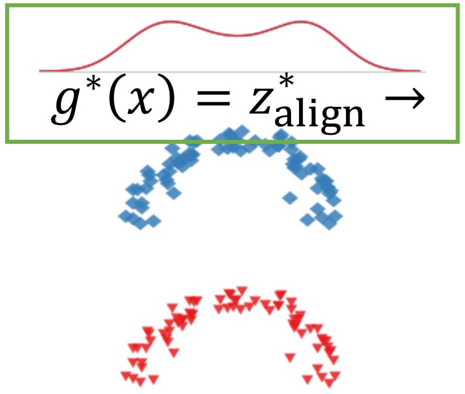


Distribution alignment

$$\min_{g \in \mathcal{G}} \phi\big(P(g(x)|d_{=1}), P(g(x)|d_{=2})\big)$$

Optimal solution

$$P(g^*(x)|d_{=1}) = P(g^*(x)|d_{=2})$$



Distribution alignment is NOT supervised alignment or spatial alignment

- Supervised/point-to-point alignment – Given pairs of points, learn a mapping between them
  - Example: (text, image) pairs to learn multi-modal alignment of text and images
  - Example: (English text, Spanish text) pairs to learn translation
  - In distribution alignment, no pairing information is available.

- Spatial alignment – Align images from two different perspectives
  - Example: Align satellite images of overlapping regions to combine information
  - Example: Align RGB-image with depth-image (e.g., remote sensing)
  - Example: Align pixels in frame 1 to frame 2 in video (e.g., video deblurring)
  - In distribution alignment, there is no notion of space, just distributions

Overview

Three representative applications of distribution alignment

- Fair Classification
- Domain Generalization (robustness to distribution shifts)
- Causal Representation Learning

Unified alignment framework

- Alignment definitions (what is it?)
- Alignment algorithms (how to optimize it?)
- Alignment evaluation metrics (how to evaluate it?)

Alignment applications can be unified as a task objective + ***(soft) alignment constraints***

Task objective

- Overall goal of learning
- "What we want"

Alignment constraints

- Ensures the desired property
- "What we want to avoid"

Application: Fair Classification

Background: Fair classification aims to correct historical or unintentional bias in ML systems

- In social ML applications such as loan approval, recidivism prediction (bail), or job applications, classification models can be unfair
  - Could be caused by bias in historical data
    e.g., bias against minorities in recidivism prediction
  - Could be caused by using sensitive attributes unintentionally
    e.g., even though gender is excluded from a loan approval application, other features such as name could be highly correlated with gender and used to predict
- **Demographic parity** – One notion of fairness that the prediction is independent of the sensitive attribute

$$\mathbb{E}[h(x)|d] = \mathbb{E}[h(x)]$$

  - Demographic parity gap: $|\max_d \mathbb{E}[h(x)|d] - \min_{d'} \mathbb{E}[h(x)|d']|$
- Approach: Learn a fair representation g and then classify using this representation f: $h(x) = f\big(g(x,d)\big)$

Fair classification aims to classify correctly while controlling for sensitive attributes

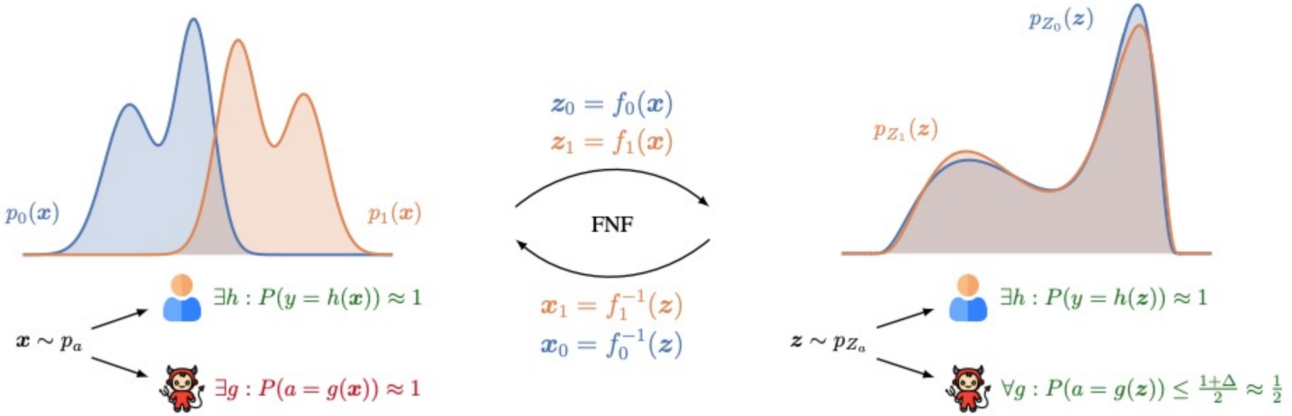
Task objective:
"What we want"

- Accurately predict whether a loan application should be approved
- Standard classification loss
  $\mathcal{L}_{\text{clf}}(f,g) = \mathbb{E}\left[\ell\left(f\left(g(x,d)\right),y\right)\right]$

(Soft) alignment constraints:
"What we want to avoid"

- The prediction must be *independent* of sensitive attribute $\boldsymbol{d}$
- Alignment constraint $\mathcal{L}_{\text{align}}(g) = \phi\left(P(g(\boldsymbol{x},d)|d_{=1}), P(g(\boldsymbol{x},d)|d_{=2})\right)$

Raw representation is good for task but sensitive attribute can be determined



Aligned representation is good for task but sensitive attribute **cannot** be determined

Illustration from: Balunovic, M., Ruoss, A., & Vechev, M. (2021, September). Fair normalizing flows. In *International Conference on Learning Representations*.

Approach 1: Fair normalizing flows use invertible models to provably learn a fair representation

- Assumption 1 – Aligner is **invertible**, i.e., $g(x, d)$ is invertible w.r.t. x
- Assumption 2 – Data distribution is known (or approximately known)
- Thus, distribution of $z = g(x, d)$ is known in closed-form as:

$$P(z|d) = \left|J_g(x|d)\right|^{-1} P(x|d)$$

- And the KL divergence can be directly estimated with samples

$$\mathcal{L}_{\text{align}}^{KL}(g) = \phi_{KL}\big(P(z|d_{=1}), P(z|d_{=2})\big) + \phi_{KL}\big(P(z|d_{=2}), P(z|d_{=1})\big)$$

- Final problem minimizes classification and alignment losses:

$$\min_{g,f} \mathcal{L}_{\text{clf}}(f, g) + \lambda \mathcal{L}_{\text{align}}^{KL}(g)$$

Balunovic, M., Ruoss, A., & Vechev, M. (2021, September). Fair normalizing flows. In *International Conference on Learning Representations*.

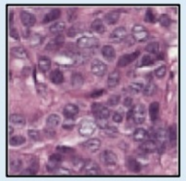
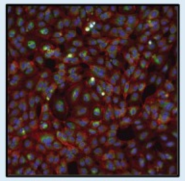
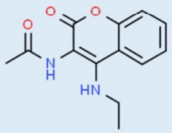
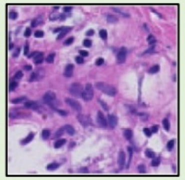
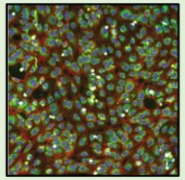
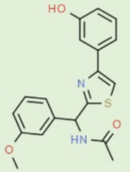
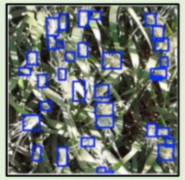
Approach 2: Fair variational autoencoders (VAE) leverage well-known upper bounds

- Assumption 1 – Encoder is probabilistic, $g(x, d) = P(z|x, d)$
- Fact 2 – Mutual information between the latent representation $z = g(x, d)$ and the domain label d is equivalent to Jensen-Shannon divergence
$$I(z = g(x, d), d) = \phi_{JSD}\big(P(z|d_{=1}), P(z|d_{=2})\big)$$
- Fair variational autoencoders upper bounds using a shared prior distribution
  - $\phi_{JSD}\big(P(z|d_{=1}), P(z|d_{=2})\big) \leq \min_{Q} \mathbb{E}_P \left[-\log \frac{Q(x|z,d)}{P(z|x,d)} Q(z)\right]$
- Others VAE-based works use other upper bounds on mutual information including via **contrastive estimation**

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. *ICLR, 2016.*
Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., & Ver Steeg, G. (2018). Invariant representations without adversarial training. *Advances in Neural Information Processing Systems*, *31*.
Gupta, U., Ferber, A. M., Dilkina, B., & Ver Steeg, G. (2021, May). Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 9, pp. 7610-7619).

Application: Domain Generalization

Background: A **distribution shift** means that the training and test distributions are different.

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000
Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	`import numpy as np … norm=np.___`
Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp p=sp.Popen() stdout=p.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... & Liang, P. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* (pp. 5637-5664). PMLR.

Background: **Distribution shifts** violate the standard assumptions in ML that train=test

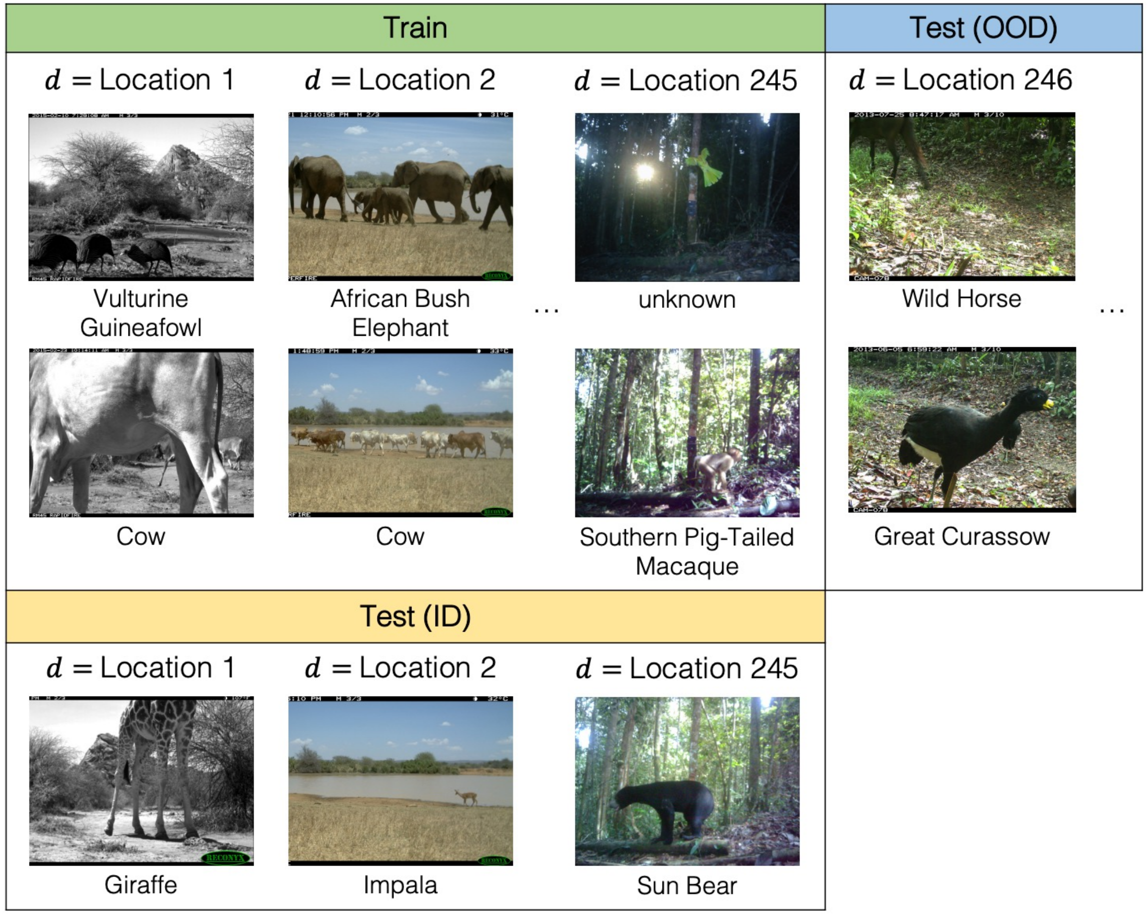
- Thus, the test accuracy very low even under benign shifts

Dataset	Metric	In-dist setting	In-dist	Out-of-dist	Gap
iWildCam2020-wilds	Macro F1	Train-to-train	47.0 (1.4)	31.0 (1.3)	16.0
Camelyon17-wilds	Average acc	Train-to-train	93.2 (5.2)	70.3 (6.4)	22.9
RxRx1-wilds	Average acc	Mixed-to-test	39.8 (0.2)	29.9 (0.4)	9.9
OGB-MolPCBA	Average AP	Random split	34.4 (0.9)	27.2 (0.3)	7.2
GlobalWheat-wilds	Average domain acc	Mixed-to-test	63.3 (1.7)	49.6 (1.9)	13.7
CivilComments-wilds	Worst-group acc	Average	92.2 (0.1)	56.0 (3.6)	36.2
FMoW-wilds	Worst-region acc	Mixed-to-test	48.6 (0.9)	32.3 (1.3)	16.3
PovertyMap-wilds	Worst-U/R Pearson R	Mixed-to-test	0.60 (0.06)	0.45 (0.06)	0.15
Amazon-wilds	10th percentile acc	Average	71.9 (0.1)	53.8 (0.8)	18.1
Py150-wilds	Method/class acc	Train-to-train	75.4 (0.4)	67.9 (0.1)	7.5

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... & Liang, P. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* (pp. 5637-5664). PMLR.

Domain generalization (DG) aims to predict accurately even under distribution shift

- Domain generalization seeks to reduce this gap caused by shifts
- A type of out-of-distribution generalization
- The test metric can be seen as a generalization of train-test split
  - Except the test split comes from an **unseen shifted** distribution
  - Given data from the training domains, find a model that performs well on a held-out **test domain dataset**



Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... & Liang, P. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* (pp. 5637-5664). PMLR.

DG Approach 1: **Domain-invariant** representation learning removes domain-specific features to aid DG performance

Task objective:
"What we want"

- Accurate prediction
- Standard classification loss on training domains
  $$\mathcal{L}_{\text{clf}}(f, g) = \mathbb{E}\left[\ell\left(f\left(g(\boldsymbol{x})\right), y\right)\right]$$

(Soft) alignment constraints:
"What we want to avoid"

- We want the features to be independent of the domain
- Feature-based alignment constraint

$$\mathcal{L}_{\text{align}}^{\text{feature}}(g) = \phi\left(P(g(\boldsymbol{x})|d_{=1}), P(g(\boldsymbol{x})|d_{=2})\right)$$

- Notice that aligner is shared across domains

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096-2030.

DG Approach 1: Domain adversarial neural networks aim to align latent features

- Intuition – Competitive game
  - Counterfeiter is trying to avoid getting caught
  - Police is trying to catch counterfeiter
- Algorithm – Usually alternating optimization between min and max
  - $h_{t+1} = \underset{h}{\operatorname{argmax}} -\mathbb{E}[\ell_{CE}(h(g_t(x)), d)]$
  - $g_{t+1}, f_{t+1} = \min_{g,f} \mathbb{E}[\ell(f(g(x)), y)] - \lambda \mathbb{E}[\ell_{CE}(h_{t+1}(g(x)), d)]$
- Drawbacks
  - **Unstable or poorly conditioned optimization**
  - Lacks domain-agnostic evaluation metrics (e.g., unable to check for overfitting)

Adversarial alignment problem

$$\min_{g,f} \mathcal{L}_{clf}(g, f) + \lambda \mathcal{L}_{align}^{adv}(g)$$

$$\min_{g,f} \mathbb{E}[\ell(f(g(x)), y)] + \lambda \left(\max_{h} -\mathbb{E}[\ell_{CE}(h(g(x)), d)] \right)$$

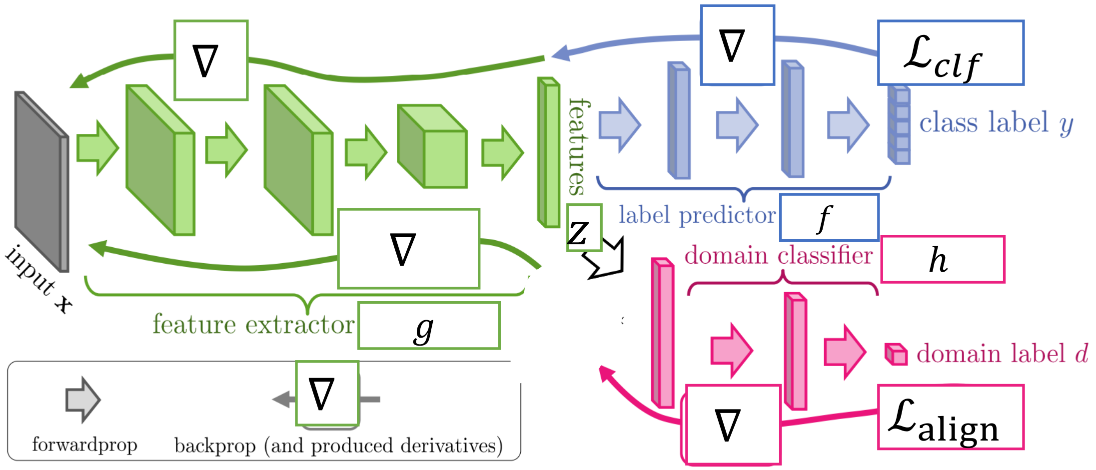


Illustration adapted from Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096-2030.

DG Approach 2: **Domain-invariant** predictors is an alternative approach to DG

Task objective:
"What we want"

- Accurate prediction
- Standard classification loss on training domains
  $$\mathcal{L}_{\text{clf}}(f, g) = \mathbb{E}\big[\ell\big(f\big(g(\boldsymbol{x})\big), y\big)\big]$$

(Soft) alignment constraints:
"What we want to avoid"

- We want the **predictors** to be independent of the domain
- Predictor-based alignment constraint

$$\mathcal{L}_{\text{align}}^{\text{predictor}}(g) = \phi\big(P(y|g(\boldsymbol{x}), d_{=1}), P(y|g(\boldsymbol{x}), d_{=1})\big)$$

  - This aligns the *conditional distribution* of label y given the features $z = g(x)$

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

DG Approach 2: **Invariant Risk Minimization (IRM)** attempts to align conditional distributions via bi-level optimization

- Minimize training error subject to the constraint that the predictive distribution is the same across all domains

$$\min_{g,f} \Sigma_d \mathcal{L}_{clf}^d(g, f)$$
$$\text{s.t. } P(y|g(x)) = P(y|g(x), d), \forall d$$

- Assumption 1: Assume that an optimal probabilistic classifier can approximate the true predictive distribution for each domain

$$f_d^* = \operatorname*{argmin}_{f'} \mathcal{L}_{clf}^d(g, f') \approx P(y|g(x))$$

- Minimize training error such that f is optimal classifier across domains

$$\min_{g,f} \Sigma_d \mathcal{L}_{clf}^d(g, f)$$
$$\text{s.t. } f = f_d^* = \operatorname*{argmin}_{f'} \mathcal{L}_{clf}^d(g, f'), \forall d$$

  - This is called a **bi-level optimization** problem

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

DG Approach 2: **Invariant Risk Minimization (IRM)** attempts to align conditional distributions via bi-level optimization

- Minimize training error such that f is optimal classifier across domains

$$\min_{g,f} \Sigma_d \mathcal{L}_{clf}^d(g,f)$$
$$\text{s.t. } f = f_d^* = \operatorname*{argmin}_{f'} \mathcal{L}_{clf}^d(g,f'), \forall d$$

  - This is called a **bi-level optimization** problem
- Bi-level optimization is very difficult similar to adversarial optimization
- Original paper proposes one approximation using gradients of classifier:

$$\min_{g,f} \Sigma_d \mathcal{L}_{clf}^d(g,f) + \lambda \left\| \nabla_f \mathcal{L}_{clf}^d(g,f) \right\|_2^2$$

  - If the gradients are zero across all domains, then f may be at an optimal point
  - Requires backpropagation through backpropagation (nested gradient computation)

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

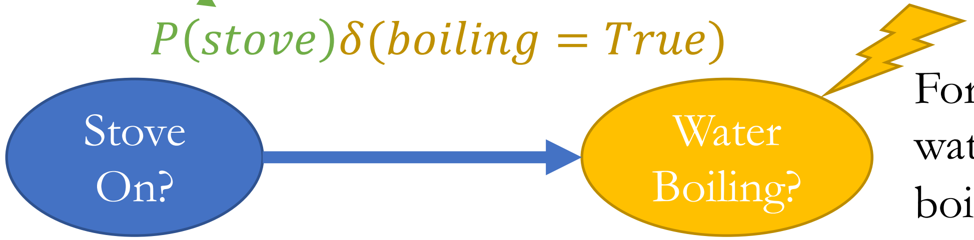
Application:
Causal Representation Learning

Background: Causal probabilistic models *implicitly* encode the effect of **interventions**

The stove distribution is the same, i.e., **aligned**!

Implied factorization
$P(stove)P(boiling|stove)$

Intervened distribution
$P(stove)\delta(boiling = True)$



Force the water to boil

Both are valid factorizations.
But which factorization is ***causal***?

One idea: The factorization that changes the least under an intervention.

Implied factorization
$P(boiling)P(stove|boiling)$

Force the water to boil

Intervened distribution
$\delta(boiling = True)\tilde{P}(stove|boiling)$

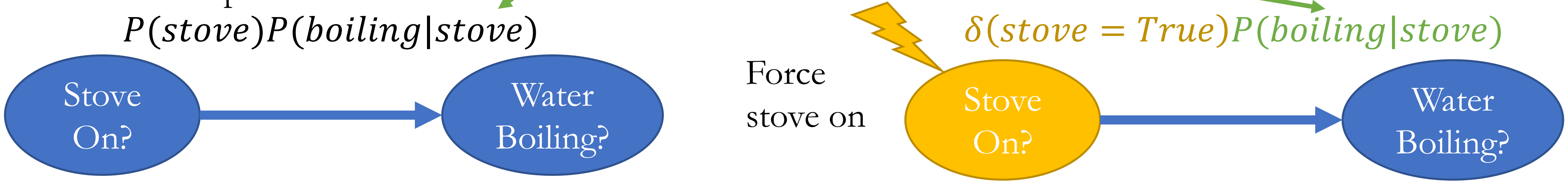
The stove distribution is different under intervention. $\tilde{P}(stove|boiling) \equiv P(stove)$

Background: Causal probabilistic models *implicitly* encode the effect of **interventions**

The boiling distribution is the same, i.e., **aligned**!

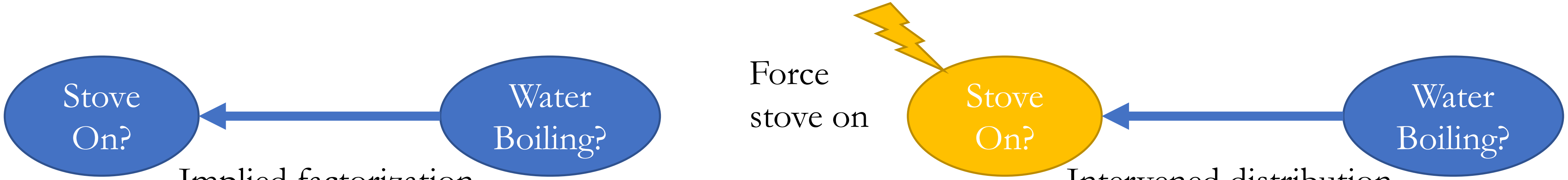
Implied factorization
$P(stove)P(boiling|stove)$

Intervened distribution
$\delta(stove = True)P(boiling|stove)$



Both are valid factorizations.
But which factorization is ***causal***?

One idea: The factorization that changes the least under an intervention.

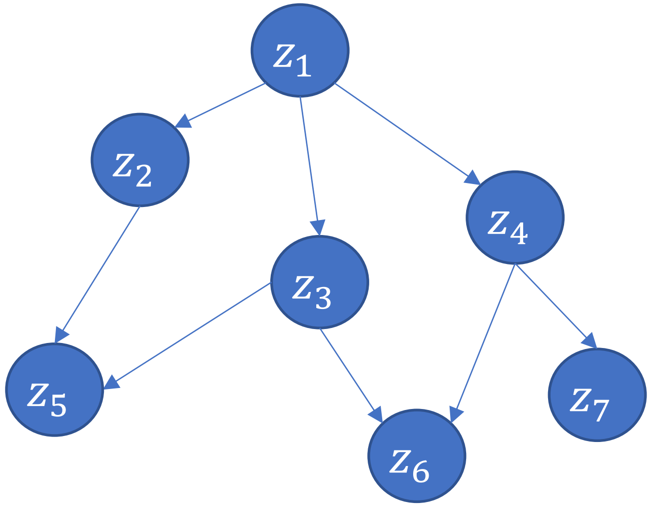


Implied factorization
$P(boiling)P(stove|boiling)$

Intervened distribution
$\tilde{P}(\text{boiling})\delta(stove = True)$

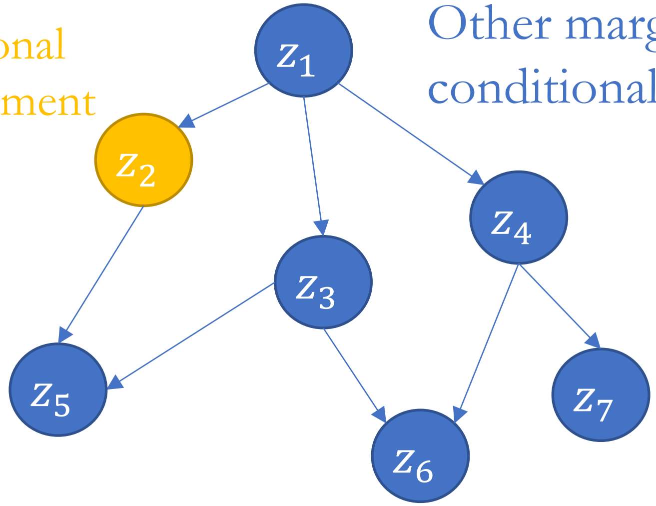
The boiling distribution is different under intervention.

Different domains can be viewed as ***unknown*** interventions in a ***latent*** causal space



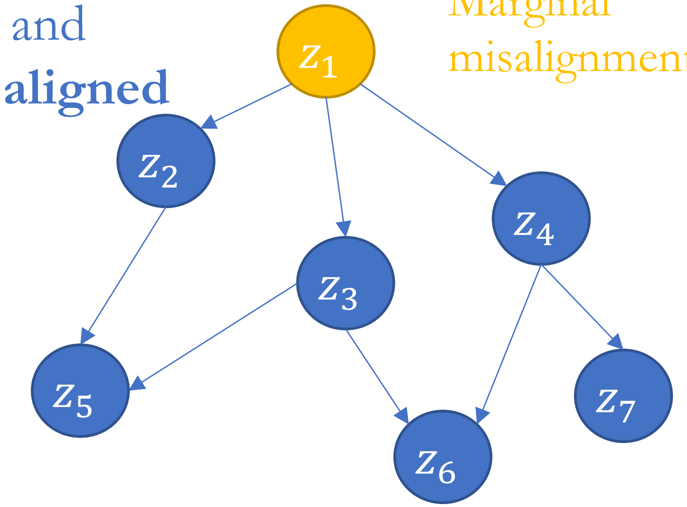
Latent space $\boldsymbol{z}|d_{=0} \sim \text{CausalModel}$

Conditional misalignment



Other marginals and conditionals are **aligned**

$\boldsymbol{z}|d_{=1} \sim \text{IntervenedCausalModel}$



Marginal misalignment

$\boldsymbol{z}|d_{=2} \sim \text{IntervenedCausalModel}$

Observed space $\boldsymbol{x} = g^{-1}(\boldsymbol{z})$

Image adapted from GlobalWheat dataset images from https://wilds.stanford.edu/datasets/.

David I. Inouye, Purdue University

Causal representation learning seeks a representation that can generate the training data but matches the true causal model

Task objective:
“What we want”

- Good generative model
- Standard generative model loss such as VAE loss
  $\mathcal{L}_{\text{gen}}(g, f) = \mathbb{E}\left[\ell_{VAE}\left(f\left(g(x)\right), x\right)\right]$

(Soft) alignment constraints:
“What we want to avoid”

- *Sparse Mechanism Shift Hypothesis* – Shifts are caused by a **sparse** change in the causal model.
- Thus, most conditional distributions should be aligned

$$\mathcal{L}_{\text{align}}^{SMS}(g)$$
$$= \Sigma_{j \notin I}\ \phi\left(P(g(x)_j | g(x)_{<j}, d_{=1}), P(g(x)_j | g(x)_{<j}, d_{=2})\right)$$

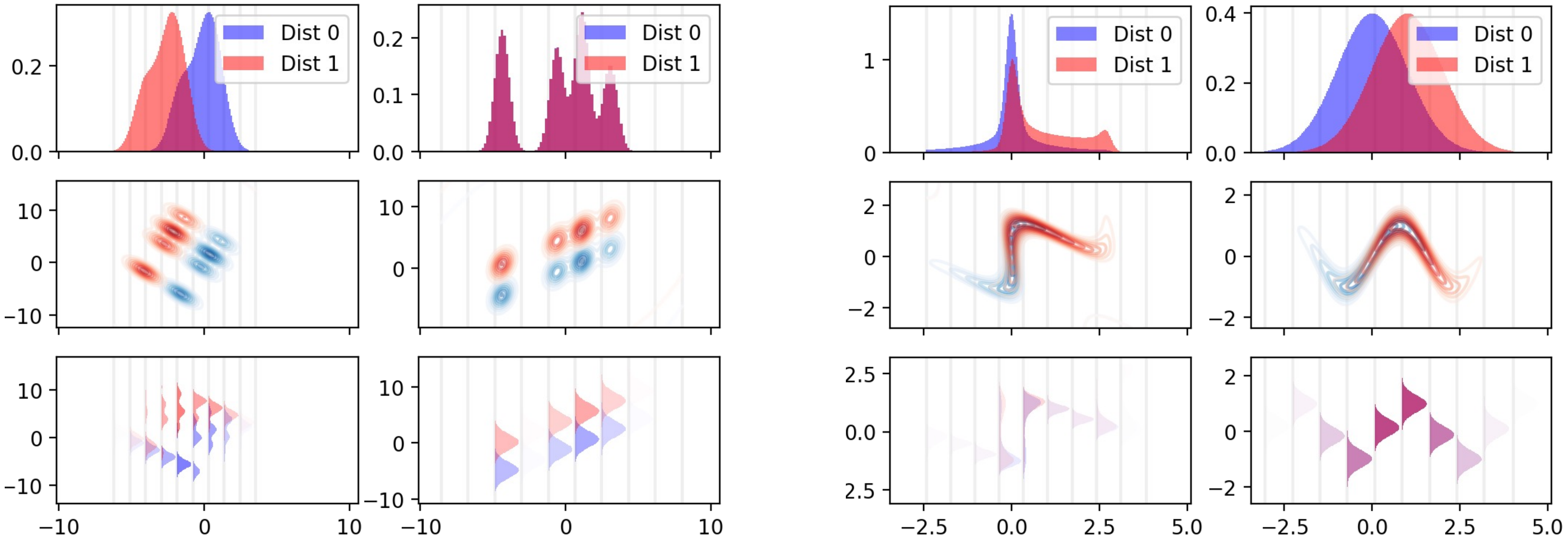
All dimensions NOT intervened should be aligned, where I is the intervention set.

Ideas inspired by Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Sparse intervention assumption => misalignment sparsity (Only a few conditionals are misaligned)

In 2D this means that either the marginal or conditionals are misaligned but **not both**.

Nothing is aligned (non-sparse) | Marginal $p(z_1)$ is aligned (sparse) | Nothing is aligned (non-sparse) | Conditional $p(z_2|z_1)$ is aligned (sparse)

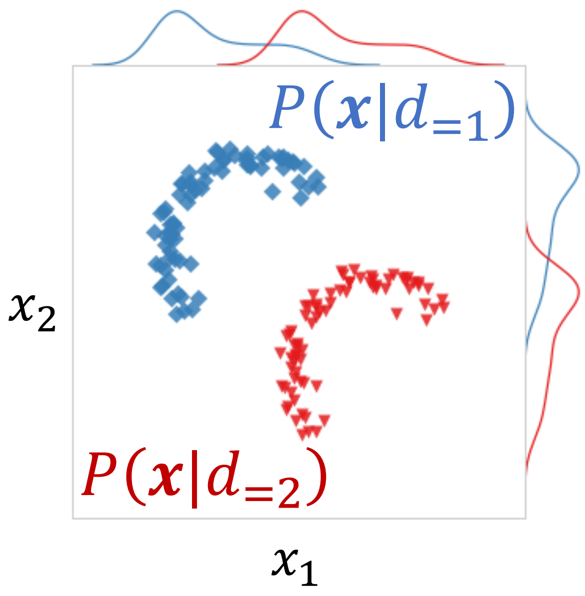


Ongoing work with Prof. Murat Kocaoglu, Sean Kulinski, Zeyu Zhou, and Ruqi Bai.

Alignment Definitions

Distribution alignment is the *opposite* objective of classification

Original Space



Optimization Objective

Classification

$$\max_{g} \phi\big(P(g(x)|d_{=1}), P(g(x)|d_{=2})\big)$$

where $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ and ϕ is a distribution divergence (e.g., KL, JSD, W_2)

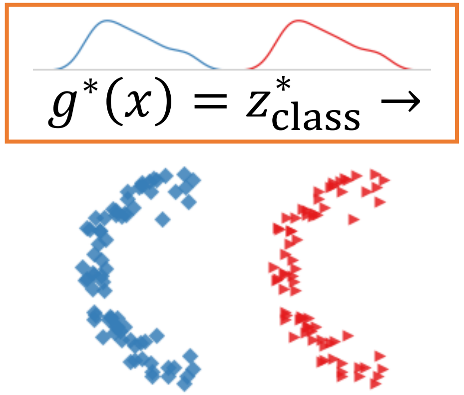
Distribution alignment

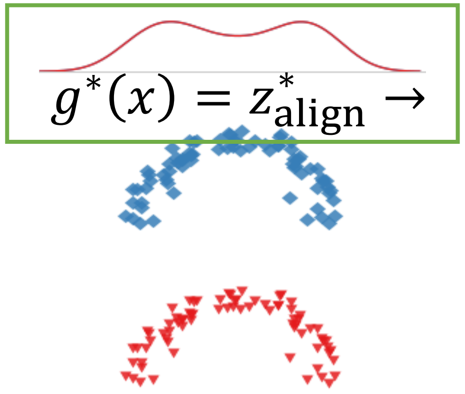
$$\min_{g} \phi\big(P(g(x)|d_{=1}), P(g(x)|d_{=2})\big)$$

Optimal solution

$$P(g^*(x)|d_{=1}) = P(g^*(x)|d_{=2})$$

Latent Space





Alignment can be with respect to the marginal, conditional, or joint distribution

Marginal alignment

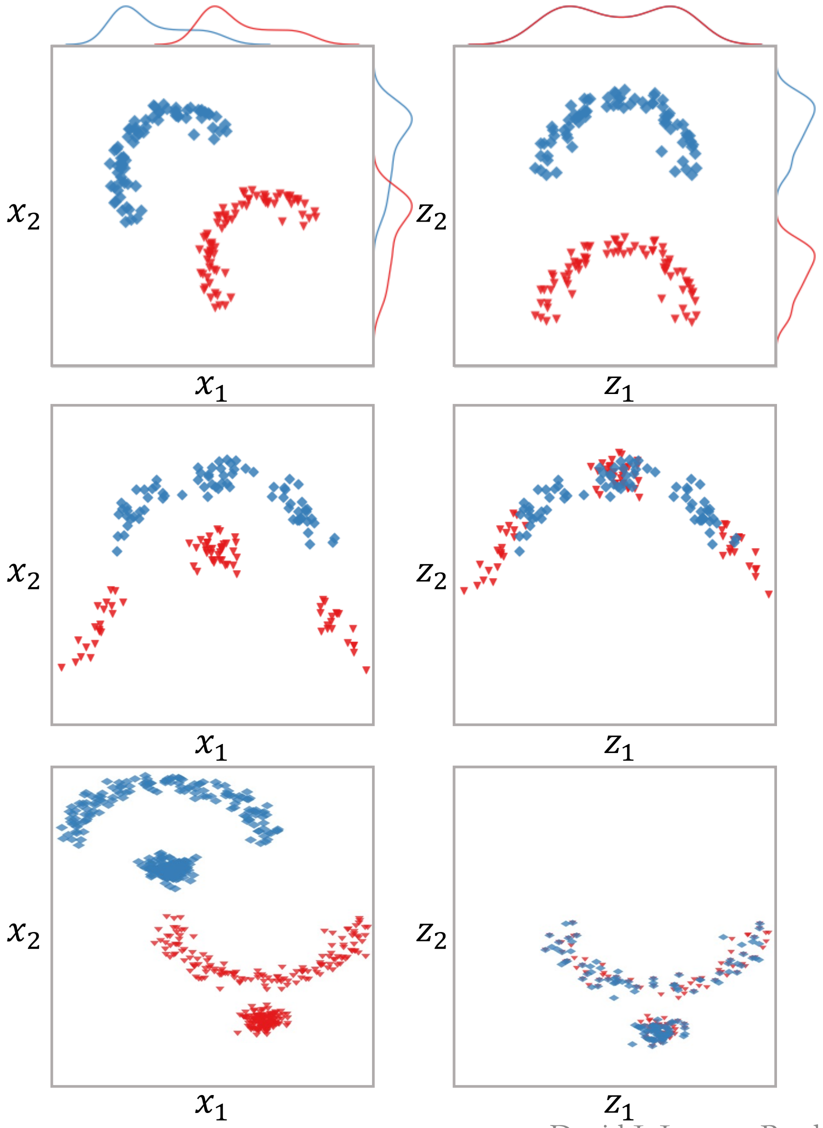
$P(z_1|d_{=1}) = P(z_1|d_{=2})$

Conditional alignment

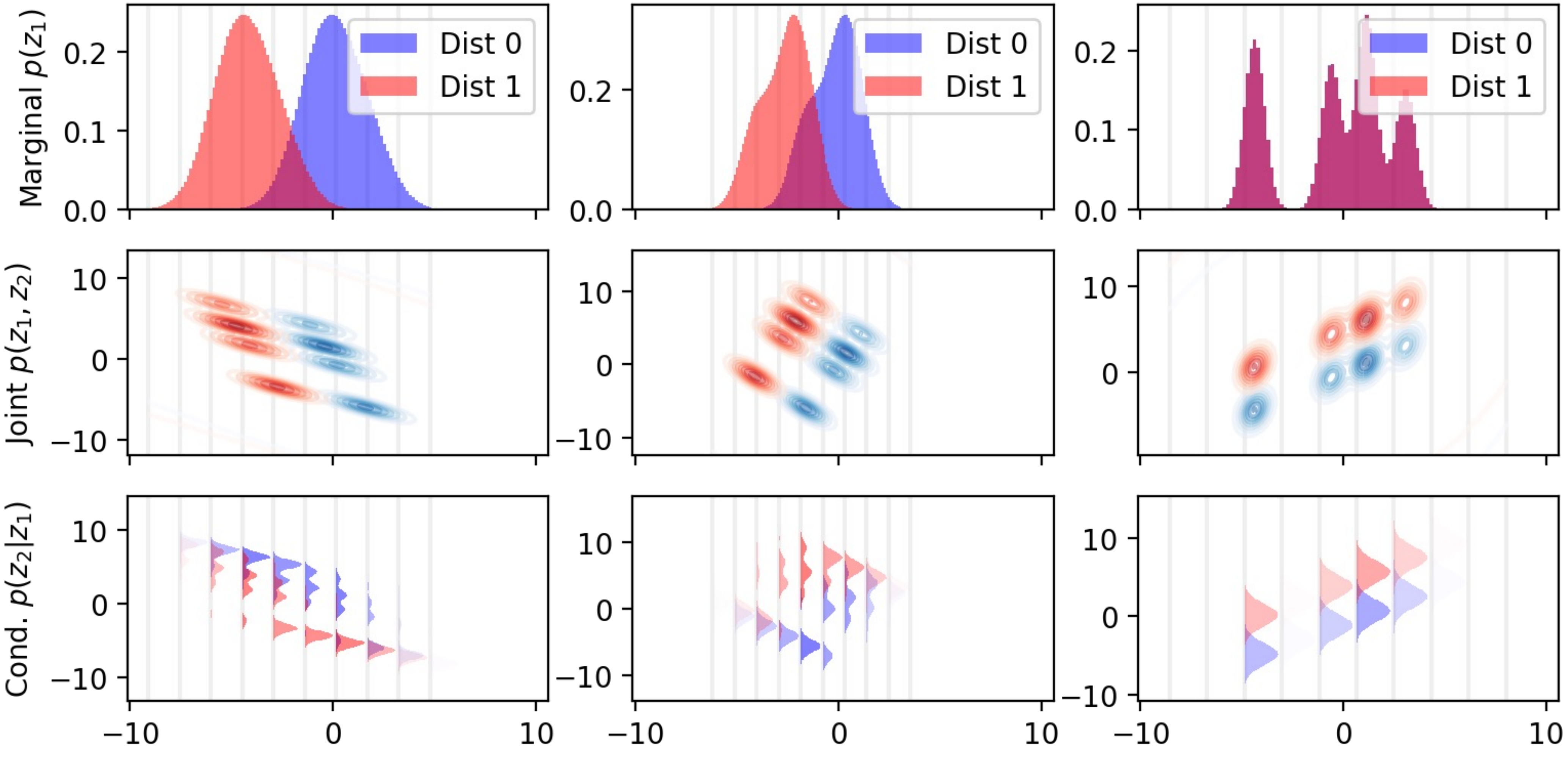
$P(z_2|z_1, d_{=1}) = P(z_2|z_1, d_{=2})$

Joint alignment

$P(z_1, z_2|d_{=1}) = P(z_1, z_2|d_{=2})$

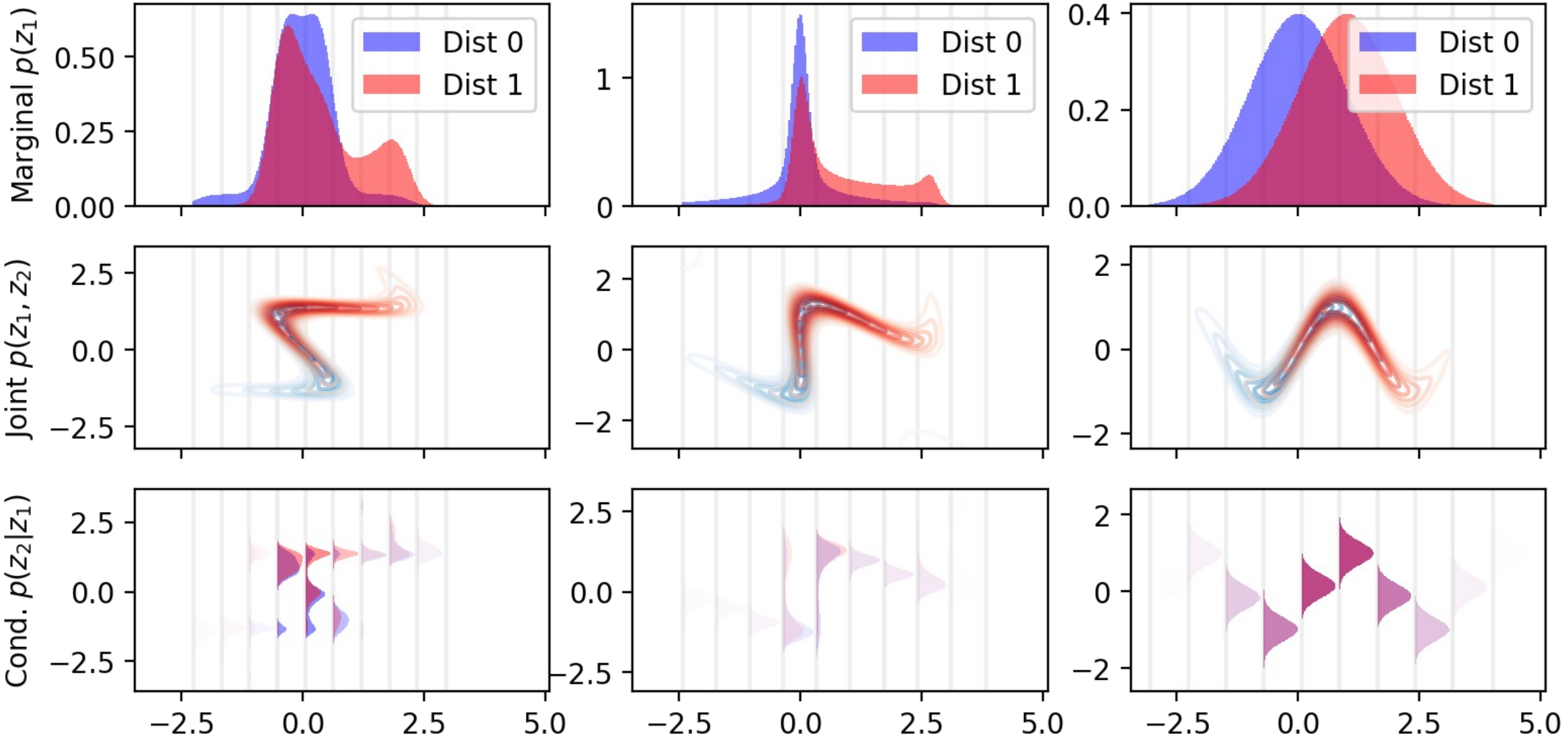


Example: Marginal alignment without conditional alignment



David I. Inouye, Purdue University

Example: Conditional alignment without marginal alignment



Distribution alignment minimizes the divergence between two distributions

Definition 1: Joint Distribution Alignment

Given samples from the joint distribution $P(\boldsymbol{x}, d)$, *distribution alignment* is the problem of finding an *aligner* $g\colon \mathcal{X} \times \mathcal{D} \to \mathcal{Z}$ that minimizes a distribution divergence $\phi\colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ between the domain-conditional distributions:

Any distribution divergence that satisfies non-negativity and $\phi(P, Q) = 0$ if and only if $P = Q$ (e.g., KL, JSD, W_2).

Aligner can depend on domain label d

$$\min_{g \in \mathcal{G}} \phi(P(\boldsymbol{z} \mid d_{=1}), P(\boldsymbol{z} \mid d_{=2})), \quad \text{where} \quad \boldsymbol{z} \equiv g(\boldsymbol{x}, d).$$

Definition 2: Conditional Distribution Alignment

Given two variable index sets $\mathcal{A}, \mathcal{B} \in \{1,2, \dots, m\}$, *conditional alignment* minimizes an aggregation, defined by an aggregator $\Omega_{\mathcal{Z}_\mathcal{B}}[\cdot]$, over all conditional divergences:

$$\min_{g \in \mathcal{G}} \Omega_{\mathcal{Z}_\mathcal{B}}[\, \phi(P(\boldsymbol{z}_\mathcal{A} \mid \boldsymbol{z}_\mathcal{B}, d_{=1}), P(\boldsymbol{z}_\mathcal{A} \mid \boldsymbol{z}_\mathcal{B}, d_{=2}))\,], \quad \text{where} \quad \boldsymbol{z} \equiv g(\boldsymbol{x}, d).$$

Usually this is merely the expectation over $\boldsymbol{z}_\mathcal{B}$, i.e., $\mathbb{E}_{P(\boldsymbol{z}_\mathcal{B})}[\cdot]$

David I. Inouye, Purdue University

Constraints on aligners can be explicit or implicit

- Explicit constraints
  - *Translation* aligner, i.e., $g(\boldsymbol{x}, d) = \begin{cases} \boldsymbol{x}, & \text{if } d = 1 \\ \tilde{g}(\boldsymbol{x}), & \text{otherwise} \end{cases}$
  - *Shared* aligner between domains, i.e., $g(\boldsymbol{x}, d) = \tilde{g}(\boldsymbol{x})$
  - *Invertible* aligner, i.e., $\exists g^{-1} \text{ s.t. } \forall \boldsymbol{x},\ g^{-1}(g(\boldsymbol{x}, d), d) = \boldsymbol{x}$
    - *Approximately invertible* via cycle consistency $\exists f \text{ s.t. } \forall \boldsymbol{x},\ f(g(\boldsymbol{x}, d), d) \approx \boldsymbol{x}$

- Implicit (soft-)constraints via other optimization terms
  - As in the **alignment applications**

Alignment Algorithms

Alignment algorithms fall into two broad categories: adversarial and non-adversarial

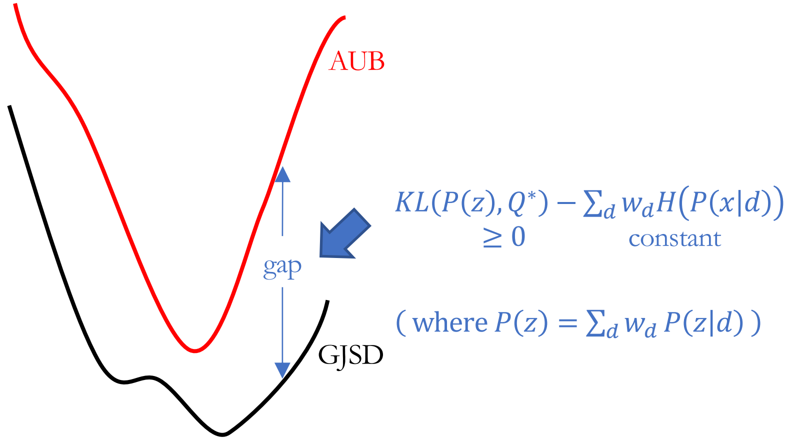
- Adversarial alignment was the first and continues to be the most popular approach to alignment
  - Good – Easy to implement, just add a discriminator for the domain
  - Good – No restriction on model architectures
  - Bad – Very challenging to optimize
  - Bad – Hard to evaluate solution
- Non-adversarial algorithms impose alignment via
  - Bi-level optimization
  - Likelihood-based (either **normalizing flows** or VAEs)
  - Input-convex models
  - Diffusion models
  - Optimal transport techniques
  - Good – Non-adversarial optimization is generally easier and more scalable
  - Bad – Sometimes tied to specific architectures (e.g., invertible or input-convex)

Alignment Upper Bound (AUB) forms an upper bound on JS divergence via ***invertible*** models

- A variational **upper** bound of JSD:

$$\phi_{AUB}(g) = \min_{Q \in \mathcal{Q}} \sum_{d=1}^{k} \mathbb{E}_{P(\boldsymbol{x}|d)}\left[-\log\left|J_{g_d}\right| Q\left(g(\boldsymbol{x}, d)\right)\right]$$

  - Q is a density model ***shared*** among domains
  - g is ***invertible*** and $\left|J_{g_d}\right|$ is the determinant Jacobian of $g(\cdot, d)$



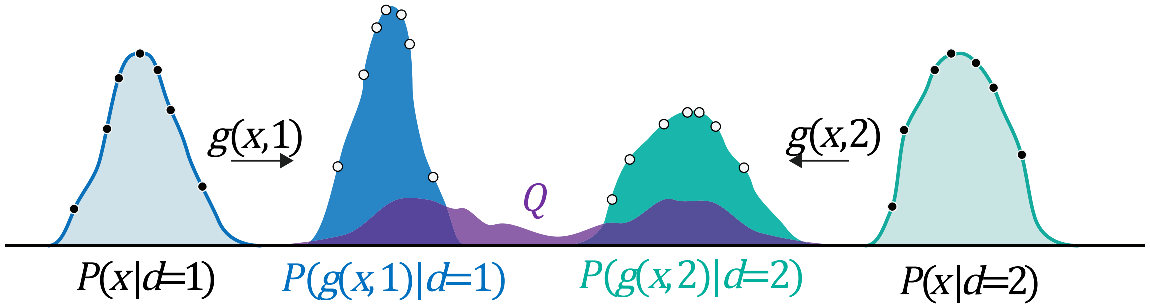
- ***Bound gap*** is exactly $KL\left(\sum_d w_d P(z|d),\ Q(\boldsymbol{z})\right)$
- ***Any*** Q provides an **upper** bound on JSD + const

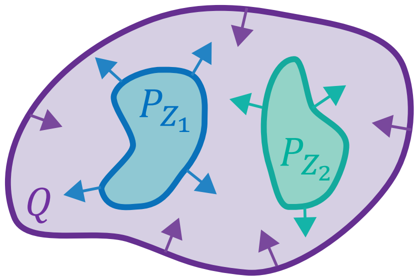
Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Alignment via JSD Upper Bound. Accepted to *Neural Information Processing Systems (NeurIPS)*. Preprint: https://arxiv.org/abs/2207.02286

AUB optimization provides a **cooperative** alternative to adversarial alignment

AUB <u>cooperative</u> alignment problem

$$\min_{g} \left(\min_{Q \in \mathcal{Q}} \sum_{j=1}^{k} \mathbb{E}_{P(\boldsymbol{x}|d)} \left[\log |J_{g_d}| Q(g(\boldsymbol{x}, d)) \right] \right)$$





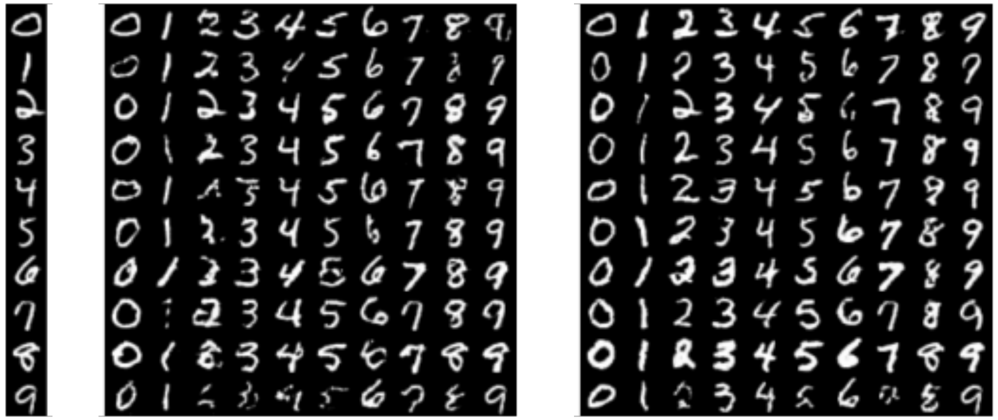
- Minimizing g makes distributions closer to current Q (left)
- Minimizing Q tightens bound by getting closer to the latent mixture, i.e., $\sum_d P(g(x,d)|d)$ (right)

Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Alignment via JSD Upper Bound. Accepted to *Neural Information Processing Systems (NeurIPS)*. Preprint: https://arxiv.org/abs/2207.02286

AUB can perform alignment on tabular data and between multiple domains

	MINIBOONE (42)	GAS (7)	HEPMASS (20)	POWER (5)
LRMF	12.79	-6.17	18.49	-0.93
AF (MLE)	14.08	-6.52	19.37	-0.77
AF (Adv. only)	18.18	-3.15	21.70	-0.39
AF (hybrid)	19.49	-3.76	21.42	-0.43
Ours	**12.11**	**-7.09**	**18.26**	**-1.19**

These results on 4 benchmark tabular datasets demonstrate that our algorithm can improve the AUB alignment measure on test data.



AlignFlow (MLE) Ours

Our AUB algorithm can translate between 10 domains (MNIST digits here) better than the closest competitor (AlignFlow) for invertible models. (Original real digits are far left and grid is translations to all other digits.)

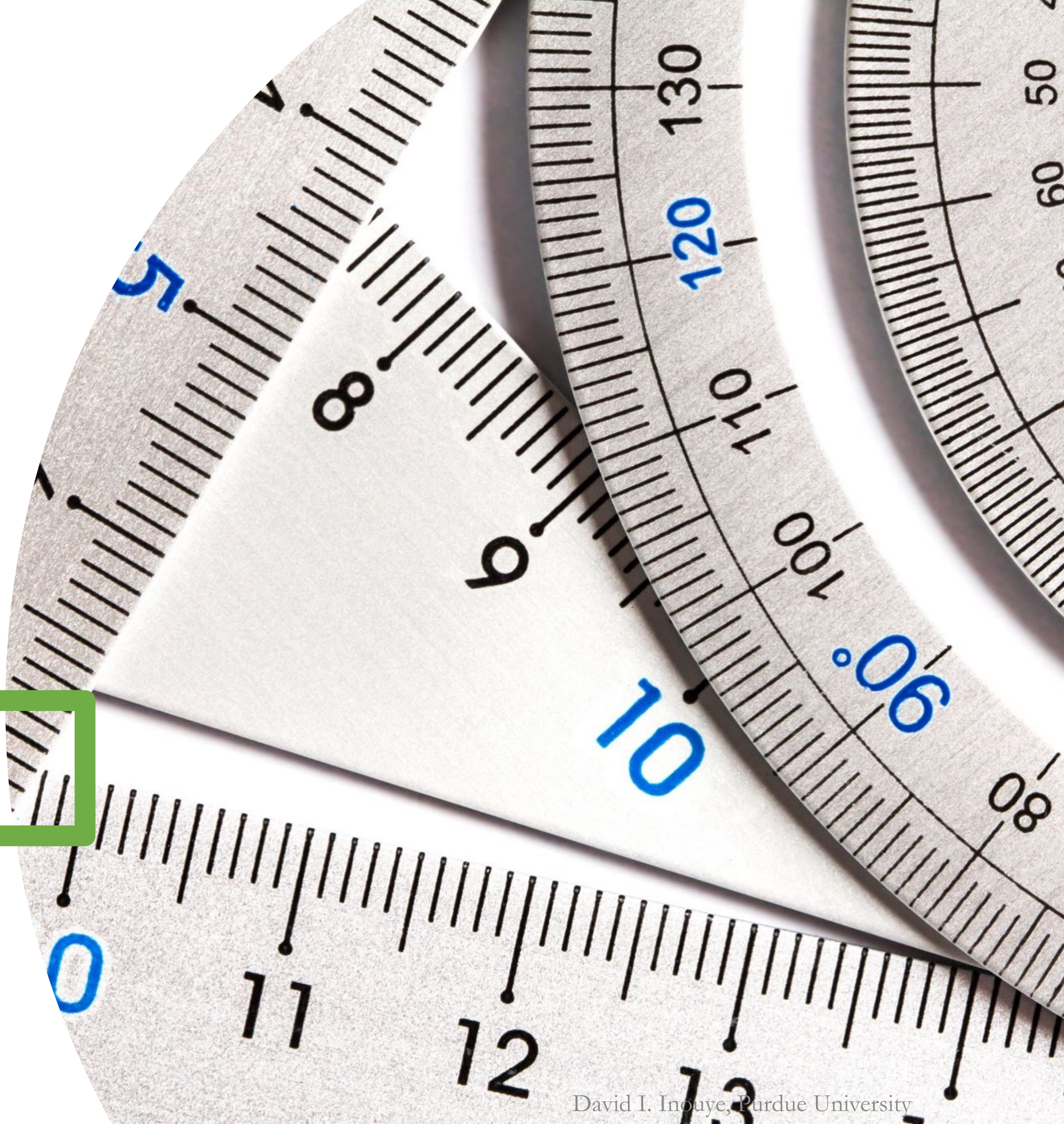
Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Alignment via JSD Upper Bound. Accepted to *Neural Information Processing Systems (NeurIPS)*. Preprint: https://arxiv.org/abs/2207.02286

Alignment problems can be formulated under a unified alignment framework

Name	Kind	$g(\boldsymbol{x}, d{=}1)$	$g(\boldsymbol{x}, d{=}2)$	$\boldsymbol{z}_{\mathcal{A}}$	$\boldsymbol{z}_{\mathcal{B}}$	Method	Other Objectives
Generative models (GAN) [37]	Marg.	$\tilde{g}(\boldsymbol{x})$	$\boldsymbol{x}$	$\boldsymbol{z}$	-	Adversarial	-
Unsupervised image-to-image translation [74, 91, 92]	Marg.	$\tilde{g}_1(\boldsymbol{x})$ s.t. $\tilde{g}_2(\tilde{g}_1(\boldsymbol{x})) \approx \boldsymbol{x}$	$\tilde{g}_2(\boldsymbol{x})$ s.t. $\tilde{g}_1(\tilde{g}_2(\boldsymbol{x})) \approx \boldsymbol{x}$	$\boldsymbol{z}$	-	Adversarial	Identity regularization, cycle consistency
Domain adversarial NN (DANN) [87]	Marg.	$\tilde{g}(\boldsymbol{x})$	$\tilde{g}(\boldsymbol{x})$	$\boldsymbol{z}$	-	Adversarial	Classification
Conditional DANN [88, 90]	Cond.	$[\tilde{g}(\boldsymbol{x}), y]$	$[\tilde{g}(\boldsymbol{x}), y]$	$\tilde{g}(\boldsymbol{x})$	y	Adversarial	Classification
Invariant Risk Minimization [10]	Cond.	$[\tilde{g}(\boldsymbol{x}), y]$	$[\tilde{g}(\boldsymbol{x}), y]$	y	$\tilde{g}(\boldsymbol{x})$	Bi-level Opt.	Classification
Optimal transport (Monge map) [75]	Marg.	$\tilde{g}(\boldsymbol{x})$	$\boldsymbol{x}$	$\boldsymbol{z}$	-	Empirical OT	Transport cost
Conditional optimal transport [93]	Cond.	$\tilde{g}(\boldsymbol{x}_{\mathcal{A}} \mid \boldsymbol{x}_{\mathcal{B}})$	$\boldsymbol{x}_{\mathcal{A}}$	$\boldsymbol{z}_{\mathcal{A}}$	$\boldsymbol{z}_{\mathcal{B}}$	Adversarial	Transport cost
Flow-based generation or translation [39, 40, 81, 94, 95]	Marg.	$\tilde{g}_1(\boldsymbol{x})$ s.t. $\exists\, \tilde{g}_1^{-1}$	$\tilde{g}_2(\boldsymbol{x})$ s.t. $\exists\, \tilde{g}_2^{-1}$	$\boldsymbol{z}$	-	Likelihood	-
Fair Variational Autoencoders [31, 33, 34]	Marg.	$\tilde{g}_1(\boldsymbol{x})+\epsilon$	$\tilde{g}_2(\boldsymbol{x})+\epsilon$	$\boldsymbol{z}$	-	Likelihood	Classification

Table from manuscript in preparation by David I. Inouye.

Alignment Evaluation



Evaluating alignment is challenging because most divergences are intractable to estimate given only samples

- Most theoretic divergences ϕ cannot be computed with only samples
  - KL divergence
  - Jensen-Shannon divergence
  - Wasserstein distance
- In practice, papers evaluate using extrinsic and intrinsic metrics
  - *Extrinsic metrics* – These do not directly estimate the divergence ϕ but are consequences of alignment
  - *Intrinsic metrics* – These directly approximate the divergence ϕ to see if the algorithm reduced the divergence
- Often, these approximations will be ***upper or lower bounds*** on the divergence
- Finally, some divergences are scale-invariant
  - Informally, this means that changing the unit of the dimensions (e.g., from inches to feet) does not affect the divergence

Alignment metrics can be unified under this common framework

Name	Kind	Bound	Scale Inv.	Notes
FID [48]	Extr.	-	No	FID is the most common evaluation measure.
Inception Score (IS) [49]	Extr.	-	No	Another common evaluation measure.
External task metric	Extr.	-	No	Examples: fair classification [35] or domain generalization [10].
f-divergence adv. loss [99]	Intr.	Lower	Yes	Adversarial losses are rarely used for evaluation.
Wasserstein adv. loss [100]	Intr.	Lower	No	Adversarial losses are rarely used for evaluation.
Flow-based likelihood measures [39–41]	Intr.	Upper	Yes	In prior work [41], we unify and generalize AlignFlow [39] and LRMF [40] via alignment upper bound (AUB).
VAE-based likelihood measures [31, 33, 34]	Intr.	Upper	Yes	In **Task 2.2**, I propose an improved VAE-based alignment objective generalizing my prior work [41].
Empirical (discrete) Wasserstein [75, 98]	Intr.	-	No	Quadratic in the number of samples. Variants: Monge via linear program [75] and entropic via Sinkhorn [98]
Sliced Wassserstein Distance [101, 102]	Intr.	-	No	Only sorting required given 1D projection. Variants: Average SW [101, 102], max SW [53, 77, 103], tree SW [104]

Table from manuscript in preparation by David I. Inouye.

David I. Inouye, Purdue University

Future research opportunities in all areas of distribution alignment

Alignment concepts	• Conditional alignment in particular
Alignment algorithms	• Stable and scalable non-adversarial methods
Alignment evaluation	• More application-agnostic measures • Rigorous evaluation protocols
Alignment applications	• Causal discovery and inference