Towards Trustworthy Machine Learning via Distribution Matching

David I. Inouye

PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

While ML has made great strides in recent years, ML still has many issues

- Large language models could compound historical bias against minorities
- Autonomous driving systems could cause loss of life when exposed to unexpected conditions
- Medical ML systems could recommend fatal treatments based on false counterfactual prediction
- Scientists may make incorrect scientific conclusions based on black-box models

The next generation of **trustworthy ML** will need to exhibit properties beyond accuracy

- Group Fairness – Are the predictions fair w.r.t. age or race?
- Robustness – Are the predictions accurate even in new environments?
- Causality – Can counterfactual queries be correctly estimated?
- Explainability – Can distribution shifts be explained?

Fair classification example: Optimize performance given fairness constraint

- Let f be a classification model and let $x, y, d,$ and $\hat{y} \equiv f(x)$ be the input, label, sensitive attribute, and model's prediction respectively
- Let $p_f(x, y, d, \hat{y})$ denote the joint distribution over all variables
- **Demographic parity (DP) difference** measures the difference between the prediction probability between groups

$$\Delta_{DP}(f) = \left| p_f(\hat{y}_{=1} | d_{=1}) - p_f(\hat{y}_{=1} | d_{=2}) \right|$$

- Fair classification w.r.t. DP can be formalized as a task objective subject to a fairness constraint:

$$\begin{aligned} \min_f \quad & \mathcal{L}(f) \equiv \mathbb{E}_{p(x,y)}[\ell(f(x), y)] \\ \text{s.t.} \quad & \Delta_{DP}(f) \leq \delta \end{aligned}$$

David I. Inouye, Purdue University

The fairness constraint can be satisfied via distribution matching

- First, notice that $\Delta_{DP}(f)$ is total variation distance

$$\Delta_{DP}(f) = D_{TV}\left(p_f(\hat{y}|d_{=1}), p_f(\hat{y}|d_{=2})\right) \leq \delta$$

  - This is a distribution matching constraint on $p(\hat{y}|d)$!

- In practice, the model is often decomposed into feature extraction followed by a classifier head f_{cls}

$$f(x,d) = f_{cls}\big(g(x,d)\big)$$

  - where $z = g(x,d)$ is a latent representation
  - The task loss can be written in terms of g, i.e., $\mathcal{L}(g) = \min_{f_{cls}} \mathbb{E}_{p(x,y)}\left[\ell\big(f_{cls}(g(x,d)), y\big)\right]$

- To ensure DP fairness, a **sufficient** condition is to ensure the latent TV distance is small
(due to data processing inequality for *f*-divergences)

$$D_{TV}\left(p_f(\hat{y}|d_{=1}), p_f(\hat{y}|d_{=2})\right) \leq D_{TV}\left(p_g(z|d_{=1}), p_g(z|d_{=2})\right) \leq \delta$$

  - This is a distribution matching constraint on $p_g(z|d)$!

David I. Inouye, Purdue University

Distribution matching for trustworthy ML takes the form of (soft) task constraints

Definition 1: Given a task objective $\mathcal{L}(g)$, a distribution matching constraint imposes matching on the latent representation $z \coloneqq g(x, d, \epsilon)$, where $g \in \mathbb{G}$ is called a *matcher* and $D(p, q)$ is a divergence:

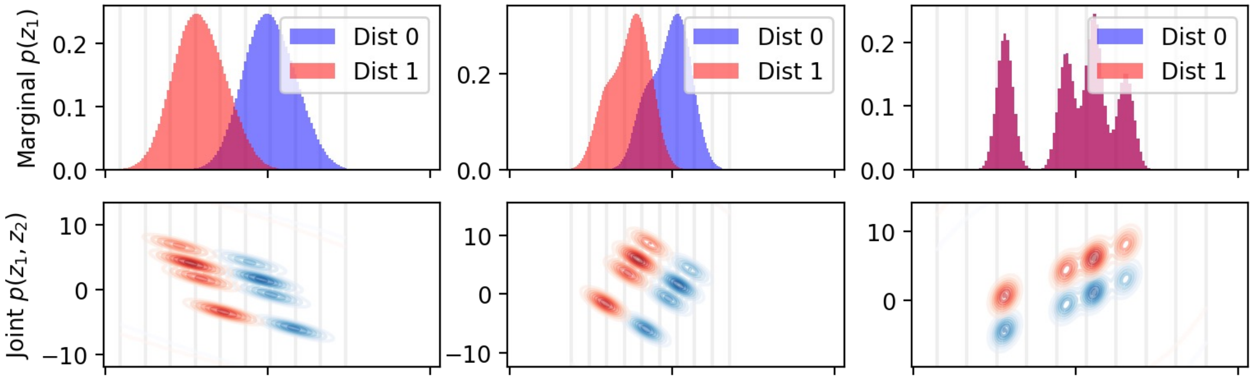
Hard DM constraint:

$$\min_{g \in \mathbb{G}} \ \mathcal{L}(g)$$
$$\text{s.t.} \ \ D\left(p_g(z|d_{=1}), p_g(z|d_{=2})\right) \leq \delta$$

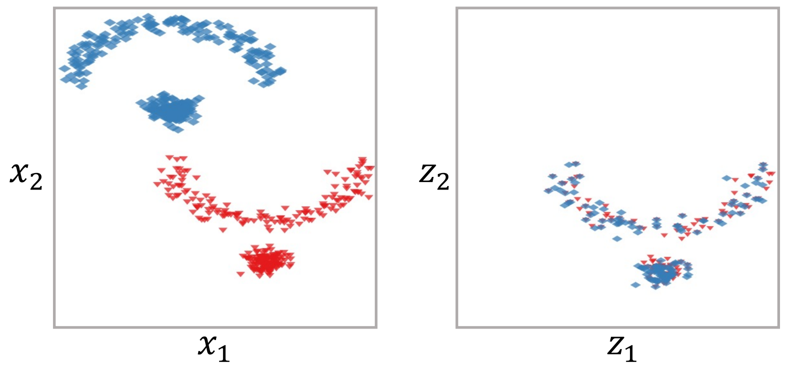
Soft DM constraint (i.e., regularization):

$$\min_{g \in \mathbb{G}} \ \mathcal{L}(g) + \lambda D\left(p_g(z|d_{=1}), p_g(z|d_{=2})\right)$$

Learn a rotation + 1D projection so that $p_g(z_1|d)$ is aligned



Learn 2D g such that $p(z_1, z_2|d)$ is aligned where the red distribution is fixed, i.e., $g(x, 2) = x$

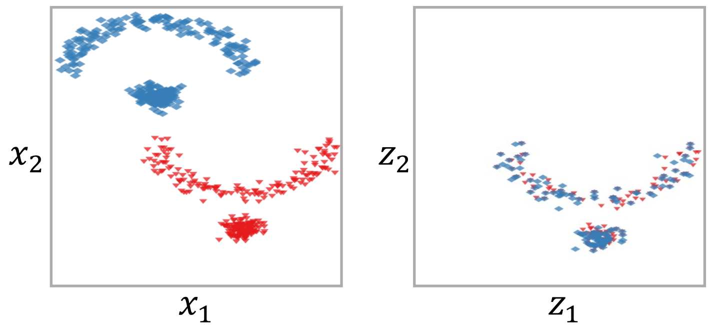


David I. Inouye, Purdue University

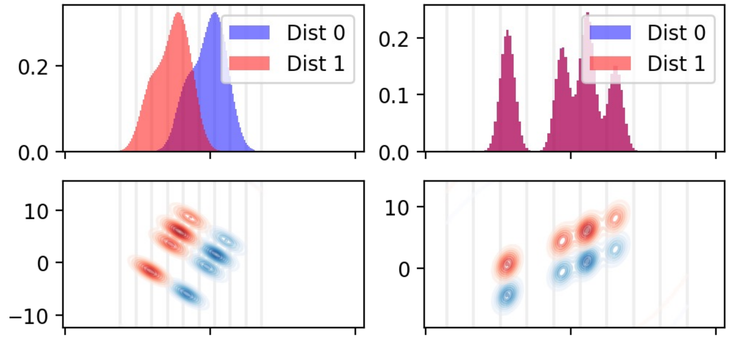
A matcher can have different structures $\mathbb{G}$ depending on the context

- *Translation* matcher, i.e., $g(\boldsymbol{x}, d) = \begin{cases} \boldsymbol{x}, & \text{if } d = 1 \\ \tilde{g}(\boldsymbol{x}), & \text{otherwise} \end{cases}$
- *Shared* matcher between domains, i.e., $g(\boldsymbol{x}, d) = \tilde{g}(\boldsymbol{x})$
- *Invertible* matcher, i.e., $\exists g^{-1}$ s.t. $\forall \boldsymbol{x},\ g^{-1}(g(\boldsymbol{x}, d), d) = \boldsymbol{x}$
  - *Approximately invertible* via cycle consistency $\exists f$ s.t. $\forall \boldsymbol{x}$, $f(g(\boldsymbol{x}, d), d) \approx \boldsymbol{x}$
- *Stochastic* matcher, i.e., $g(x, d, \epsilon)$, where ϵ is exogenous noise.

Translation matcher (unsupervised)



Shared matcher



Distribution matching has been known by other names

Distribution alignment

(Domain-)Invariant representation learning

Adversarial representation learning

Mutual information minimization

Can we generalize distribution matching to **conditional distributions?**

- Yes! But we need a notion of **conditional divergence**.

Definition 2: Given index sets $\mathcal{A}, \mathcal{B} \subseteq \{1,2,\dots,m\}$, a conditional divergence $D_{\mathcal{A}|\mathcal{B}}(p,q)$ is a function that satisfies two properties:

1. Non-negativity, i.e., $D_{\mathcal{A}|\mathcal{B}}(p,q) \geq 0$.
2. Conditional distribution equality, i.e., $D_{\mathcal{A}|\mathcal{B}}(p,q) = 0 \quad \Leftrightarrow \quad p(z_{\mathcal{A}}|z_{\mathcal{B}}) = q(z_{\mathcal{A}}|z_{\mathcal{B}}), \forall z_{\mathcal{B}}$

- Any divergence D can be extended via an expectation over some $\tilde{p}(z_{\mathcal{B}})$

$$D^{\mathbb{E}}_{\mathcal{A}|\mathcal{B}}(p,q) := \mathbb{E}_{\tilde{p}(z_{\mathcal{B}})}\left[D\left(p(z_{\mathcal{A}}|z_{\mathcal{B}}), q(z_{\mathcal{A}}|z_{\mathcal{B}})\right)\right]$$

- A conditional divergence allows the marginals of $z_{\mathcal{B}}$ to be different

Conditional distribution matching is much less explored but is more general

Definition 3: Given a task objective $\mathcal{L}(g)$, a conditional matching constraint imposes matching on latent conditional distributions given a conditional divergence $D_{\mathcal{A}|\mathcal{B}}$:

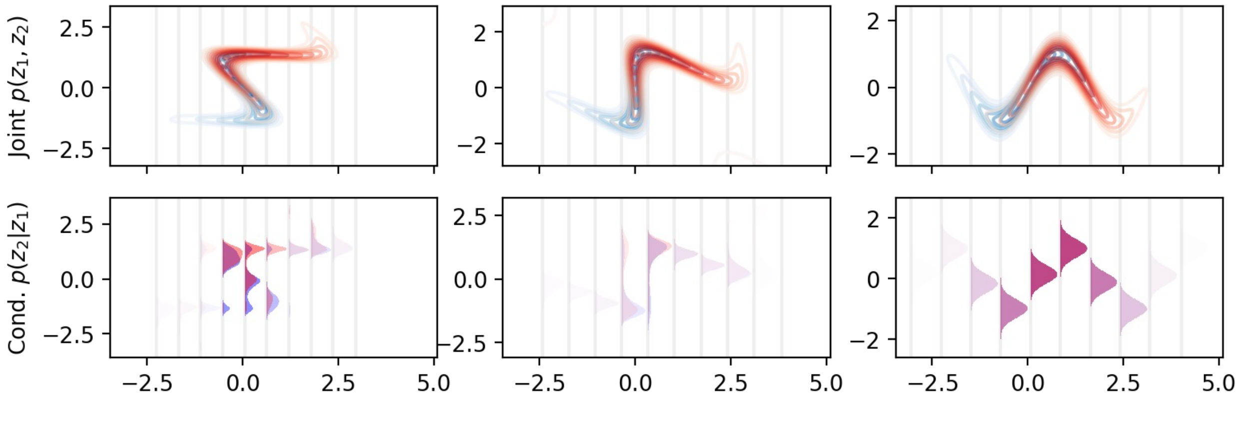
Hard DM constraint:

$$\min_{g\in\mathbb{G}} \mathcal{L}(g)$$
$$\text{s.t.} \quad D_{\mathcal{A}|\mathcal{B}}\left(p_g(z|d_{=1}), p_g(z|d_{=2})\right) \le \delta$$

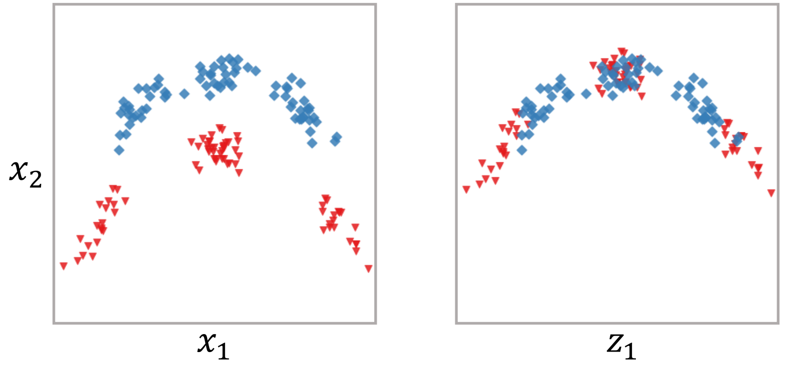
Soft DM constraint (i.e., regularization):

$$\min_{g\in\mathbb{G}} \mathcal{L}(g) + \lambda D_{\mathcal{A}|\mathcal{B}}\left(p_g(z|d_{=1}), p_g(z|d_{=2})\right)$$

Learn a rotation + 1D projection so that $p_g(z_2|z_1, d)$ is aligned



Learn 2D g such that $p(z_2|z_1, d)$ is aligned where the red distribution is fixed, i.e., $g(x, 2) = x$



Conditional distribution matching is much less explored but is more general

Definition 3: Given a task objective $\mathcal{L}(g)$, a conditional matching constraint imposes matching on latent conditional distributions given a conditional divergence $D_{\mathcal{A}|\mathcal{B}}$:

Hard DM constraint:

$$\min_{g\in\mathbb{G}} \ \mathcal{L}(g)$$
$$\text{s.t.} \ \ D_{\mathcal{A}|\mathcal{B}}\left(p_g(z|d_{=1}), p_g(z|d_{=2})\right) \le \delta$$

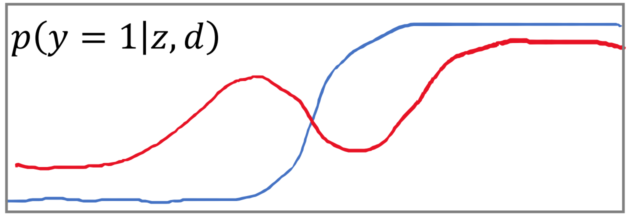
Soft DM constraint (i.e., regularization):

$$\min_{g\in\mathbb{G}} \ \mathcal{L}(g) + \lambda D_{\mathcal{A}|\mathcal{B}}\left(p_g(z|d_{=1}), p_g(z|d_{=2})\right)$$

Matching the binary class probability given a single latent feature $\boldsymbol{p(y|\tilde{z}, d)}$

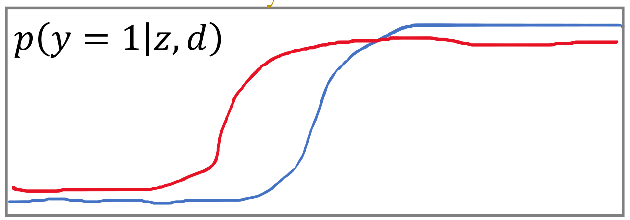
- In this case, $z \equiv (y, \tilde{z})$ and $\mathcal{A} = \{1\}, \mathcal{B} = \{2\}$
- For every value of $\tilde{z}$ the positive class probability $p(y = 1|\tilde{z}, d)$ must match
- (This is related to Invariant Risk Minimization, IRM)

Mismatched ☹



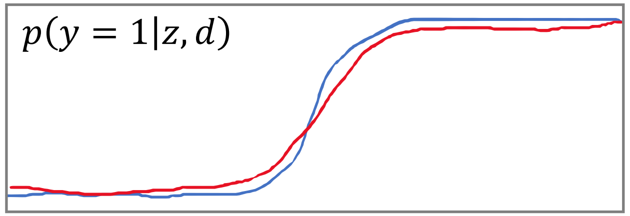
$\tilde{z}$

Partially matched 😐



$\tilde{z}$

Matched ☺



$\tilde{z}$

Many group fairness measures can be described as distribution matching constraints

Marginal Matching

Demographic parity

$p(\hat{y}|d)$

Conditional Matching

Equalized odds

$p(\hat{y}|y,d)$ — Conditional equalized odds: $p(\hat{y}|y,x_1,d)$ — $\cdots$ — $p(\hat{y}|y,x_1,\dots,x_m,d)$

Sufficiency

$p(y|\hat{y},d)$

Conditional Demographic Parity

$p(\hat{y}|x_1,d)$ — $p(\hat{y}|x_1,x_2,d)$ — $\cdots$ — $p(\hat{y}|x_1,\dots,x_m,d)$

- In this case, the latent representation is merely the prediction ($z \equiv \hat{y}$) and the matcher is merely the predictor ($g \equiv f$)

Fair classification is one of the trustworthy ML applications via distribution matching

Application	Method	Task Loss	Distribution to Match	Matcher Structure
Fair Classification	Fair VAE Adversarially Fair Fair Flows	Classification loss	$p_g(z\|d)$	Stochastic Shared Invertible

Many DG methods can be viewed as matching different latent distributions

Application	Method	Task Loss	Distribution to Match	Matcher Structure
Fair Classification	Fair VAE Adversarially Fair Fair Flows	Classification loss	$p_g(z\|d)$	Stochastic Shared Invertible
Domain Generalization ⭐ [Bai et al., 2023]	DANN	Classification loss	$p_g(z\|d)$	Shared
	CDANN		$p_g(z\|y, d)$	Shared
	IRM		$p_g(y\|z, d)$	Shared
	Fishr		$p_g(\nabla_\theta \mathcal{L}_\theta(x)\|d)$	Implicit

Causal ML methods have different task losses than classification

Application	Method	Task Loss	Distribution to Match	Matcher Structure
Fair Classification	Fair VAE Adversarially Fair Fair Flows	Classification loss	$p_g(z \mid d)$	Stochastic Shared Invertible
Domain Generalization ⭐ [Bai et al., 2023]	DANN CDANN IRM Fishr	Classification loss	$p_g(z \mid d)$ $p_g(z \mid y, d)$ $p_g(y \mid z, d)$ $p_g(\nabla_\theta \mathcal{L}_\theta(x) \mid d)$	Shared Shared Shared Implicit
Causality ⭐ [Kulinski et al., 2023]	CATE ICP Domain Counterfactuals	Factual risk n/a NLL	$p_g(z \mid d)$ $p_g(y \mid z_{Pa(y)}, d)$ $p_g(z_i \mid z_{<i}, d)$ $\forall i$ not intervened	Invertible Permutation Shared

Explaining distribution shifts can be cast as finding an interpretable matcher

Application	Method	Task Loss	Distribution to Match	Matcher Structure
Fair Classification	Fair VAE Adversarially Fair Fair Flows	Classification loss	$p_g(z\|d)$	Stochastic Shared Invertible
Domain Generalization ⭐ [Bai et al., 2023]	DANN	Classification loss	$p_g(z\|d)$	Shared
	CDANN		$p_g(z\|y, d)$	Shared
	IRM		$p_g(y\|z, d)$	Shared
	Fishr		$p_g(\nabla_\theta \mathcal{L}_\theta(x)\|d)$	Implicit
Causality ⭐ [Kulinski et al., 2023]	CATE	Factual risk	$p_g(z\|d)$	Invertible
	ICP	n/a	$p_g(y\|z_{Pa(y)}, d)$	Permutation
	Domain Counterfactuals	NLL	$p_g(z_i\|z_{<i}, d)$ $\forall i$ not intervened	Shared
Dist. Shift Explanations ⭐ [Kulinski & Inouye, 2023]	Sparse transport	Reg. Transport Cost	$p_g(z\|d)$	Sparse, translation
	Interpretable transport	Transport Cost	$p_g(z\|d)$	Sparse or cluster, translat.

Some analogies to summarize

Classification is to current ML
as
scaling data or models? is to trustworthy ML

No, doesn't solve fundamental issues

Classification is to current ML
as
<u>application-specific design?</u> is to trustworthy ML

No, it is not as broadly applicable as classification is to many different problems

Classification is to current ML
as
distribution matching is to trustworthy ML

In contrast to scaling, it is fundamentally different from classification

In contrast to application-specific design, it is broadly applicable to many tasks

Why could distribution matching be **an enabling tool** for trustworthy ML?

Two analogies: One fun, the other more technical

Classification is to Tigger as distribution matching is to Eeyore

Optimism – Pessimism

Classification is to task objective
as
distribution matching is to <u>task constraints</u>

Positive goal – Negative constraint

Performance – Safety

Why could distribution matching be **broadly applicable**?

Class labels are to classification as <u>domain labels</u> are to distribution matching

Data-driven

Easier to elicit from experts than formal definitions

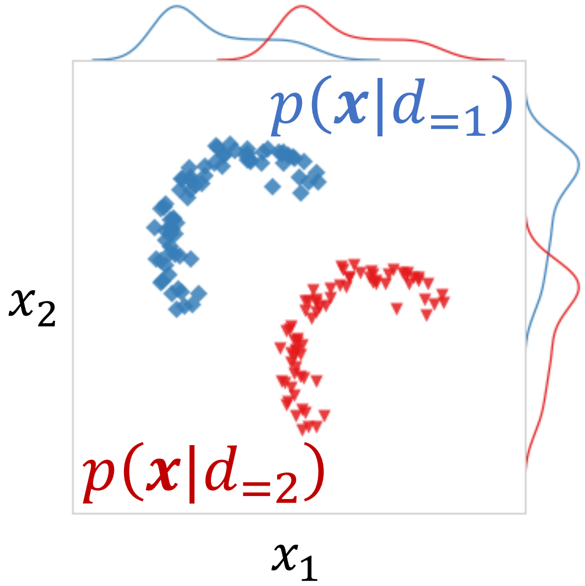
Generic algorithms

What is distribution matching?

Divergence maximization is to classification
as
divergence minimization is to distribution matching

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

Original Space



Representation Learning Objective

Classification

$$\max_{g \in \mathbb{G}} D\big(p(g(x)|d_{=1}), p(g(x)|d_{=2})\big)$$

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

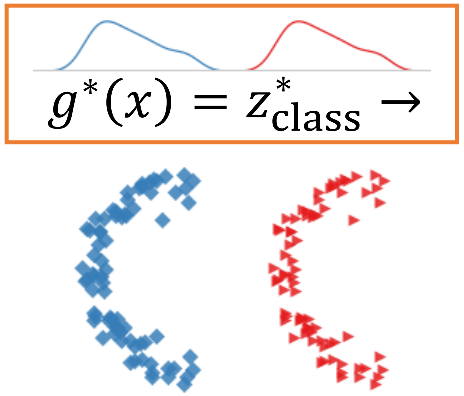
Distribution matching

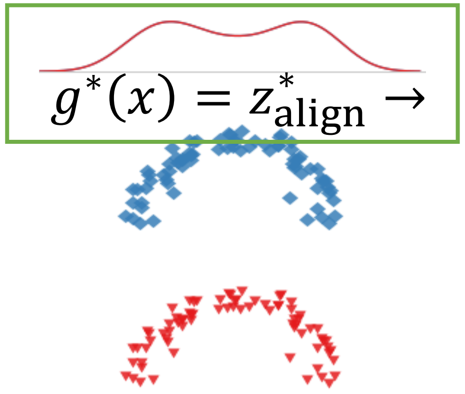
$$\min_{g \in \mathbb{G}} D\big(p(g(x)|d_{=1}), p(g(x)|d_{=2})\big)$$

Optimal solution

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

Latent Space





Yet, prior distribution matching research lacks **a unified scientific framework**

- Much prior work focuses specific applications (e.g., fairness or domain adaptation)
  - This stems from the fact that DM is rarely useful by itself
  - But DM is a much broader tool
- Other works only consider one algorithm (e.g., adversarial)
  - But there are diverse non-adversarial approaches
- DM has rarely been investigated in its own right

I aim to **unify distribution matching** under a common framework for trustworthy ML problems

Fundamentals (Already covered)

Applications (Already covered)

Algorithms

Evaluation

Algorithms:
How do we enforce distribution matching in practice?

David I. Inouye, Purdue University

Matching algorithms fall into three different categories depending on how they estimate the theoretic divergence

- Adversarial matching (GAN)
  - First and continues to be the most popular approach to matching
  - Easy to implement, just add a discriminator for the domain
  - No restriction on model architectures
  - Challenging to optimize in practice
  - Hard to evaluate solution
- Likelihood-based algorithms (flows ⭐ [Cho et al. 2022] and VAEs ⭐ [Gong et al. 2023])
  - Less well-known
  - Non-adversarial so more stable to optimize
  - Flow-based algorithms requires invertible model
  - VAE-based algorithms usually enforce fixed prior distribution
- Other algorithms
  - Optimal transport algorithms
  - Statistical conditional independence tests
  - Iterative matching ⭐ [Zhou et al., 2022a,b]

Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Matching via JSD Upper Bound. *Neural Information Processing Systems (NeurIPS)*.
Gong, Z., Usman, B., Zhao, H., & Inouye, D. I. (2023). Towards Practical Non-Adversarial Distribution Alignment via Variational Bounds. Preprint: https://arxiv.org/abs/2310.19690

Zhou, Z., Azam, S. S., Brinton, C., & Inouye, D. I. (2022, September). Efficient federated domain translation. In *The Eleventh International Conference on Learning Representations*.
Zhou, Z., Gong, Z., Ravikumar, P., & Inouye, D. I. (2022, May). Iterative matching flows. In *International Conference on Artificial Intelligence and Statistics* (pp. 6409-6444). PMLR.

Many matching algorithms form variational approximation of divergences

Adversarial / Lower Bound

$$\min_g \max_h \underline{D}(p_g(z|d_{=1}), p_g(z|d_{=2}); h)$$

- Variational maximization problem over critic $h(z)$ forms a lower bound on a divergence
- Different choices of approximation yield lower bounds on JSD, Wasserstein, f-divergences

Likelihood / Upper Bound

$$\min_g \min_q \overline{D}(p_g(z|d_{=1}), p_g(z|d_{=2}); q)$$

- Variational minimization problem over variational distribution q forms an upper bound on a divergence
- We generalize flow-based methods for upper bounds ⭐ [Cho et al., 2022]
- We revisit VAE-based methods for upper bounds ⭐ [Gong et al., 2023]

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

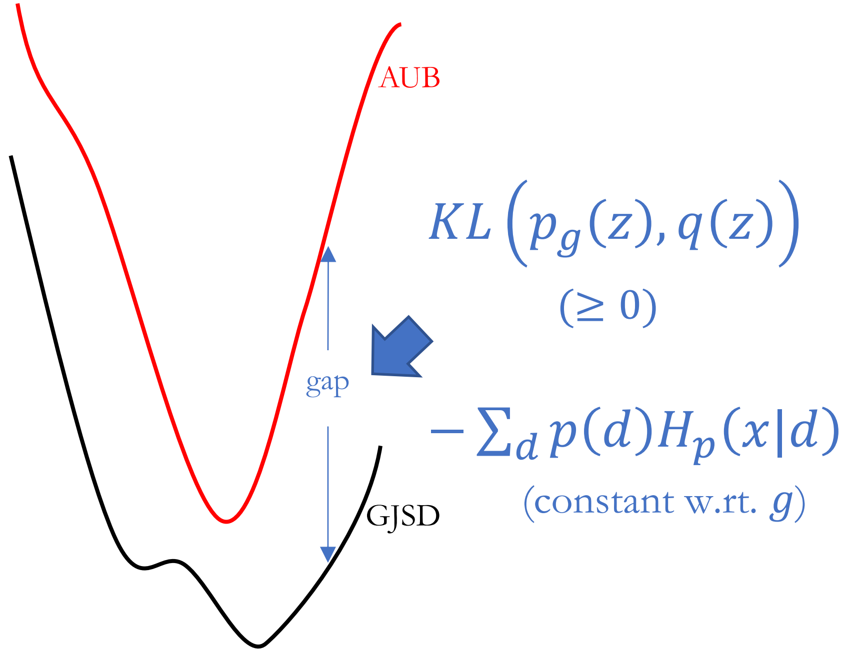
Gong, Z., Usman, B., Zhao, H., & Inouye, D. I. (2023). Towards Practical Non-Adversarial Distribution Alignment via Variational Bounds. Manuscript in preparation.

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

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

$$\overline{D}_{AUB}(g) = \min_{q(z)} \sum_{d=1}^{k} \mathbb{E}_{p(x|d)}\left[-\log\left|J_g(x,d)\right| q\left(g(x,d)\right)\right]$$

  - $q(z)$ is a prior density model ***shared*** among domains
  - $g(x,d)$ is ***invertible*** w.r.t x and $\left|J_g(x,d)\right|$ is the determinant Jacobian of g w.r.t. x



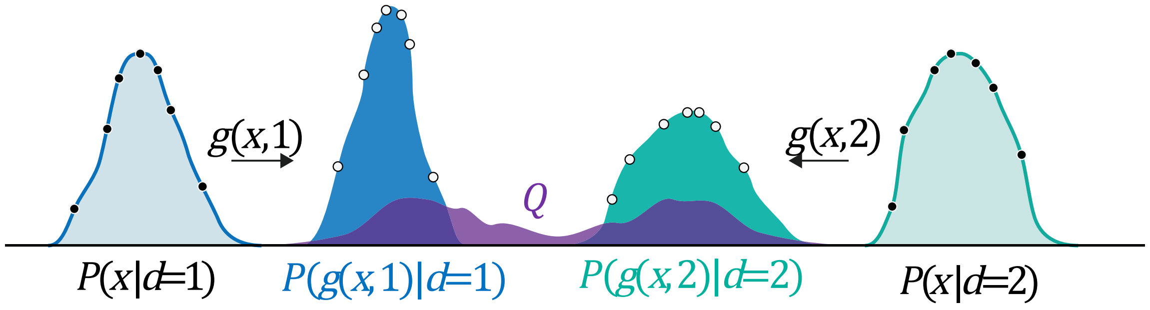
- ***Bound gap*** is exactly $KL\left(p_g(z),\ q(z)\right)$
- ***Any*** q provides an **upper** bound on JSD + const

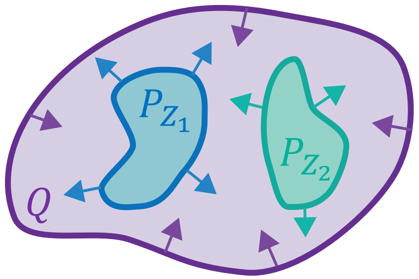
Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Alignment via JSD Upper Bound. Accepted to *Neural Information Processing Systems (NeurIPS)*. Preprint: https://arxiv.org/abs/2207.02286

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

AUB <u>cooperative</u> matching problem

$$\min_{g} \left(\min_{q(z)} \sum_{j=1}^{k} \mathbb{E}_{p(\boldsymbol{x}|d)}\left[\log|J_g(x,d)|q(g(x,d))\right] \right)$$





- Minimizing g makes distributions closer to current q (left)
- Minimizing $q(z)$ tightens bound by getting closer to the latent mixture $p(z) = \sum_d p(d)p(z|d)$ (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

The invertibility assumption can be relaxed via a decoder and reconstruction loss

Model	Jensen-Shannon Divergence Upper Bound
Flow $z = g(x, d)$	$\min_{q(z)} \; \mathbb{E}_p[-\log(\lvert J_g(x,d)\rvert \cdot q(z))] + C$
β-VAE $(\beta \leq 1)$ $z \sim p_g(z\|x, d) \equiv g(x, d, \epsilon)$	$\min_{\substack{q(z) \\ q(x\|z,d)}} \; \mathbb{E}_p\left[-\log\left(\frac{q(x\|z,d)}{p_g(z\|x,d)^\beta} \cdot q(z)^\beta\right)\right] + C$

- The Jacobian term can be relaxed by using the ratio of decoder to encoder

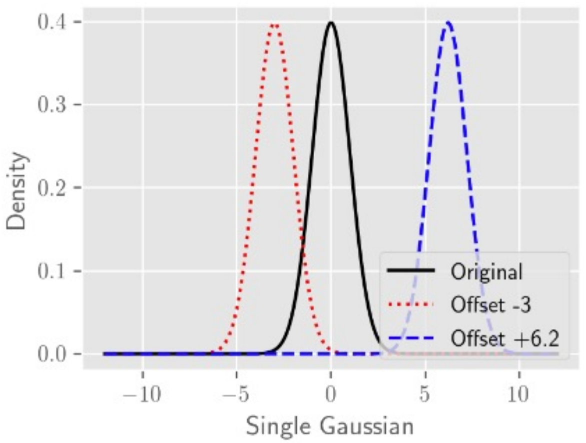
$$|J_g(x,d)| \Leftrightarrow \frac{q(x|z,d)}{p_g(z|x,d)}$$

- To encourage approximate invertibility of g, we can include a reconstruction regularization which simplifies to a β-VAE objective with $\beta \leq 1$

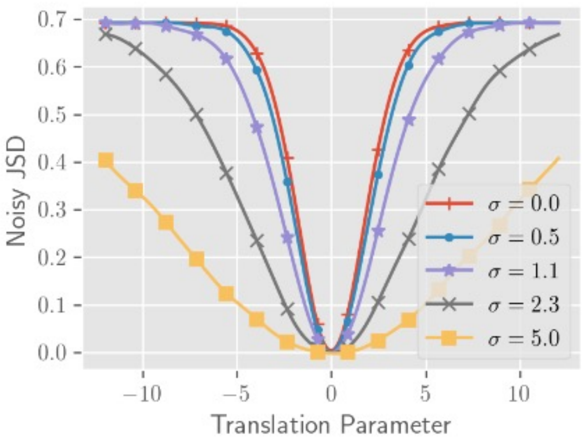
Gong, Z., Usman, B., Zhao, H., & Inouye, D. I. (2023). Towards Practical Non-Adversarial Distribution Alignment via Variational Bounds. *arXiv preprint arXiv:2310.19690.*

Adding noise to JSD can reduce vanishing gradient and local minimum problems

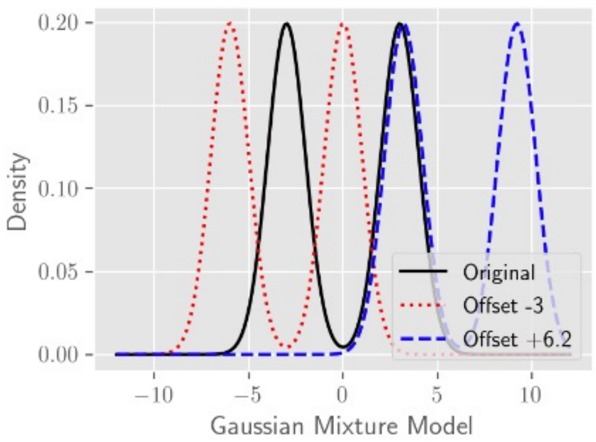
Model	Jensen-Shannon Divergence Upper Bound	Noisy JSD Upper Bound
Flow $z = g(x, d)$	$\min_{q(z)} \mathbb{E}_p[-\log(\lvert J_g(x,d)\rvert \cdot q(z))] + C$	$\min_{q(\tilde{z})} \mathbb{E}_p[-\log(\lvert J_g(x,d)\rvert \cdot q(z+\epsilon))] + C$
β-VAE $(\beta \leq 1)$ $z \sim p_g(z\|x,d) \equiv g(x,d,\epsilon)$	$\min_{q(z),\, q(x\|z,d)} \mathbb{E}_p\left[-\log\left(\frac{q(x\|z,d)}{p_g(z\|x,d)^\beta} \cdot q(z)^\beta\right)\right] + C$	$\min_{q(\tilde{z}),\, q(x\|z,d)} \mathbb{E}_p\left[-\log\left(\frac{q(x\|z,d)}{p_g(z\|x,d)^\beta} \cdot q(z+\epsilon)^\beta\right)\right] + C$



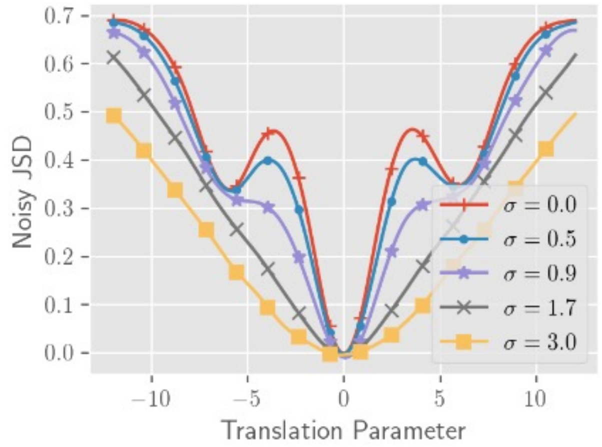
(a) Case 1: Gaussian distribution.



(b) Case 1: Opt. landscape for different noise levels.



(c) Case 2: Mixture of Gaussian distributions.



(d) Case 2: Opt. landscape for different noise levels

Gong, Z., Usman, B., Zhao, H., & Inouye, D. I. (2023). Towards Practical Non-Adversarial Distribution Alignment via Variational Bounds. *arXiv preprint arXiv:2310.19690.*

Iterative matching flows iteratively solve 1D matching problems to create deep matcher

1. Find 1D projection that is **maximally mismatched**

$$\max_{\theta} \; W_2\big(p(\theta^T x | d_{=1}), p(\theta^T x | d_{=2})\big)$$

2. Match along this 1D projection by mapping to barycenter distribution

$$\begin{aligned} \min_{\tilde{g}} \;& \mathbb{E}_{p(\tilde{x}=\theta^T x, d)}[\|\tilde{g}(\tilde{x}, d) - \tilde{x}\|^2] \\ \text{s.t.} \;& D\big(p(\tilde{g}(\tilde{x}, 1) | d_{=1}), p(\tilde{g}(\tilde{x}, 2) | d_{=2})\big) = 0 \end{aligned}$$

3. Update global matcher (add one layer) and repeat

$$g(x, d) = \tilde{g}(\theta^T x, d)\theta + x_\theta^\perp$$
$$g_{\text{global}}^{\text{new}} = g \circ g_{\text{global}}^{\text{old}}$$
$$x^{new} = g(x)$$

A projection-pursuit type of algorithm

Zhou, Z., Gong, Z., Ravikumar, P., & Inouye, D. I. (2022, May). Iterative Alignment Flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. https://proceedings.mlr.press/v151/zhou22b/zhou22b.pdf

Our framework shows the many different algorithms applied to distribution matching

Application	Method	Task Loss	Distribution to align	Aligner Structure	Algorithm
Fair Classific.	Fair VAE	ERM	$p_g(z\|d)$	Stochastic	VAE-based
	Adversarially Fair			Shared	Adversarial
	Fair Flows			Invertible	Flow-based
Domain Generaliz. [Bai et al., 2023]	DANN	ERM	$p_g(z\|d)$	Shared	Adversarial
	CDANN		$p_g(z\|y,d)$	Shared	Adversarial
	IRM		$p_g(y\|z,d)$	Shared	Bi-level optimization
	Fishr		$p_g(\nabla_\theta \mathcal{L}_\theta(x)\|d)$	Implicit	Gradient variance regularization
Causality [Kulinski et al., 2023]	CATE	Factual risk	$p_g(z\|d)$	Invertible	Integral prob. metric minimization
	ICP	n/a	$p_g(y\|z_{Pa(y)},d)$	Permutation	Statistical Indep. Tests
	Domain Counterfactuals	NLL	$p_g(z_i\|z_{<i},d)$ $\forall i$ not intervened	Shared	Generative model
Dist. Shift Explanations [Kulinski & Inouye, 2023]	Sparse transport	Reg. Transport Cost	$p_g(z\|d)$	Sparse	Sinkhorn algorithm
	Interpretable transport	Transport Cost	$p_g(z\|d)$	Sparse or cluster	Post-process empirical OT

Classification is to complexity **upper bounds**
as
distribution matching is to complexity **lower bounds**

$P \neq NP$? - Easy to construct exponential alg. but nearly impossible to prove lower bound.

Classification - Constructing one classifier that works is like proving a complexity upper bound

DM - Ensuring that **no** classifier can work is like proving a complexity lower bound

Evaluation:
How do we evaluate DM?

Distribution matching evaluation requires comparing sets in high dimensions

- Classification metric - Empirical average over **point-wise** distances/losses $\ell(\hat{y}, y)$

$$L_{\text{cls}}\big(f, p(x, y)\big) \approx \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- Distribution matching objective – Must compute **set-wise** distance

$$L_{\text{DM}}\big(g, p(x, d)\big) = D\big(p(g(x)|d_{=1}), p(g(x)|d_{=2})\big) \approx \widehat{D}\left(\left\{z_i^{(1)}\right\}_{i=1}^{n_1}, \left\{z_i^{(2)}\right\}_{i=1}^{n_2}\right) = ?$$

  - Where $\widehat{D}$ is a **set-wise** distance function that approximates the true divergence D given only samples

Current DM evaluation methods are diverse with varied strengths and weaknesses

1. Qualitative
   - Easy to inspect images or 2D visualization
   - Subjective and unsystematic
2. Two-sample test statistics
   - Examples: Demographic parity in fairness, FID in image generation models
   - Often simple to compute
   - Usually, necessary but not sufficient condition for match
     (e.g., two distributions can have the same mean and covariance but be quite different)
3. Empirical optimal transport
   - Wasserstein divergence is well-defined for empirical distributions (i.e., comparing samples directly)
   - Can be computed efficiently for empirical distributions via Sinkhorn algorithm or for 1D distributions via sorting
   - Strongly depends on the geometry of latent space
   - May not scale to high dimensions since non-parametric
4. Variational bounds on divergences
   - Lower bounds – Inner maximization of adversarial methods
   - Upper bounds – Inner minimization of likelihood-based methods
   - May scale better in high dimensions since *parametric*
   - Looseness of bounds is hard to estimate or quantify

DM metrics can be unified under these four categories

Distribution Matching Metric	Category	Bound
Compare images	Qualitative	
t-SNE plot	Qualitative	
Demographic parity in fairness	Two-sample statistic	
Frechet Inception Distance (FID)	Two-sample statistic	
Maximum Mean Discrepancy (MMD)	Two-sample statistic	
Entropic-regularized discrete OT	Empirical OT	
Sliced Wasserstein distance	Empirical OT	
f-divergence adversarial loss	Variational	Lower
Wasserstein adversarial loss	Variational	Lower
Flow-based likelihood loss	Variational	Upper
VAE-based likelihood loss	Variational	Upper

Classifier metrics are to comparing points
as
distribution matching metrics are to comparing sets

Easy - Classification accuracy is an average over **point-wise** distances

Hard - DM evaluation must compute a **set-wise** distance in high dimensional space

Conjecture: Variational bounds hold the most promise long-term.

Optimization-based metrics can leverage advances in (1) model architectures, (2) computational power, and (2) optimization, while other metrics do not benefit from these advancements.

Future research opportunities in all areas of distribution matching

Matching fundamentals	• Conditional matching in particular
Matching applications	• Causal representation learning • Domain generalization • Fair clustering?
Matching algorithms	• Stable and scalable non-adversarial methods
Matching evaluation	• More application-agnostic measures • Rigorous evaluation protocols

Thanks for listening! Any questions?

(Many thanks to helpful feedback from previous audiences.)

Application	Method	Task Loss	Distribution to align	Aligner Structure	Algorithm
Fair Classific.	Fair VAE	ERM	$p_g(z\|d)$	Stochastic	VAE-based
	Adversarially Fair			Shared	Adversarial
	Fair Flows			Invertible	Flow-based
Domain Generaliz. ⭐ [Bai et al., 2023]	DANN	ERM	$p_g(z\|d)$	Shared	Adversarial
	CDANN		$p_g(z\|y,d)$	Shared	Adversarial
	IRM		$p_g(y\|z,d)$	Shared	Bi-level optimization
	Fishr		$p_g(\nabla_\theta \mathcal{L}_\theta(x)\|d)$	Implicit	Gradient variance regularization
Causality ⭐ [Kulinski et al., 2023]	CATE	Factual risk	$p_g(z\|d)$	Invertible	Integral prob. metric minimization
	ICP	n/a	$p_g(y\|z_{Pa(y)},d)$	Permutation	Statistical Indep. Tests
	Domain Counterfactuals	NLL	$p_g(z_i\|z_{<i},d)$ $\forall i$ not intervened	Shared	Generative model
Dist. Shift Explanations ⭐ [Kulinski & Inouye, 2023]	Sparse transport	Reg. Transport Cost	$p_g(z\|d)$	Sparse	Sinkhorn algorithm
	Interpretable transport	Transport Cost	$p_g(z\|d)$	Sparse or cluster	Post-process empirical OT

This was work supported in part by NSF (IIS-2212097), ARL (W911NF-2020-221), and ONR (N00014-23-C-1016). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.