NEURAL INFORMATION PROCESSING SYSTEMS

Scalable Infomin Learning

Yanzhi Chen[1], **Weihao Sun**[2], **Yingzhen Li***[3], **Adrian Weller***[1,4]

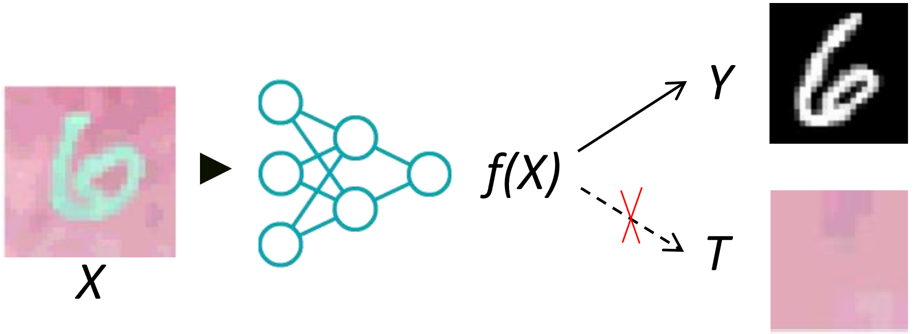
[1]Cambridge University, [2]Apple Inc., [3]Imperial College London, [4]Alan Turing Institute

Agenda

- **Background**
- **Method**
- **Results**
- **Conclusion**

Background

Infomin representation learning



$$\min_f \underbrace{\mathcal{L}(f(X), Y)}_{\text{utility}} + \beta \cdot \underbrace{I(f(X); T)}_{\text{mutual info}}$$

Applications:

(a) domain adaptation; (b) disentanglement; (c) fairness, ...

Background

Why infomin learning is hard?

$$\min_f \underbrace{\mathcal{L}(f(X), Y)}_{\text{utility}} + \beta \cdot \underbrace{I(f(X); T)}_{\text{mutual info}}$$

intractable

estimate MI by a neural net t*

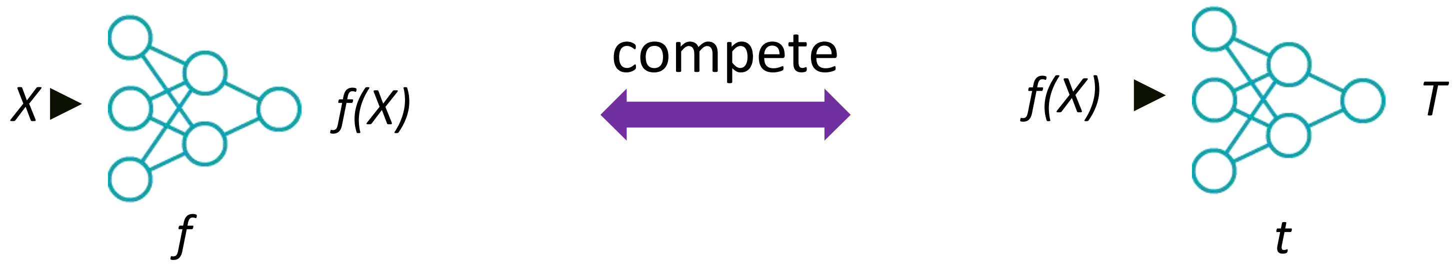
$$\min_f \max_t \mathcal{L}(f(X); Y) + \beta \cdot \hat{I}_t(f(X); T)$$

*see e.g. DANN, LAFTR, Factor-VAE, CLUB, Learning-not-to-learn...

Background

Why infomin learning is hard?

$$\min_f \max_t \mathcal{L}(f(X);Y) + \beta \cdot \hat{I}_t(f(X);T)$$



- time consuming
- hard to optimise

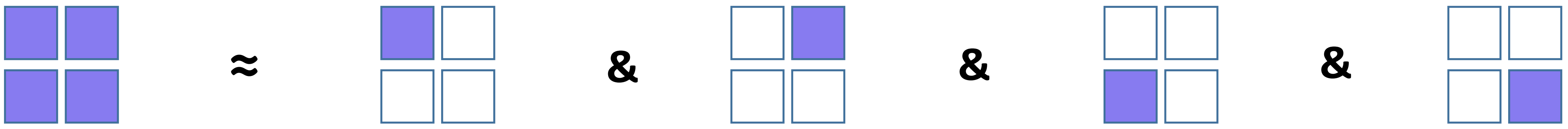
Method

Goal: $\cancel{\min_{f} \max_{t} \mathcal{L}(f(X);Y) + \beta \cdot \hat{I}_t(f(X);T)}$

Idea: find a proxy to *I*(Z; T) whose estimate is easy to compute

▼

sliced mutual information (*SI*)

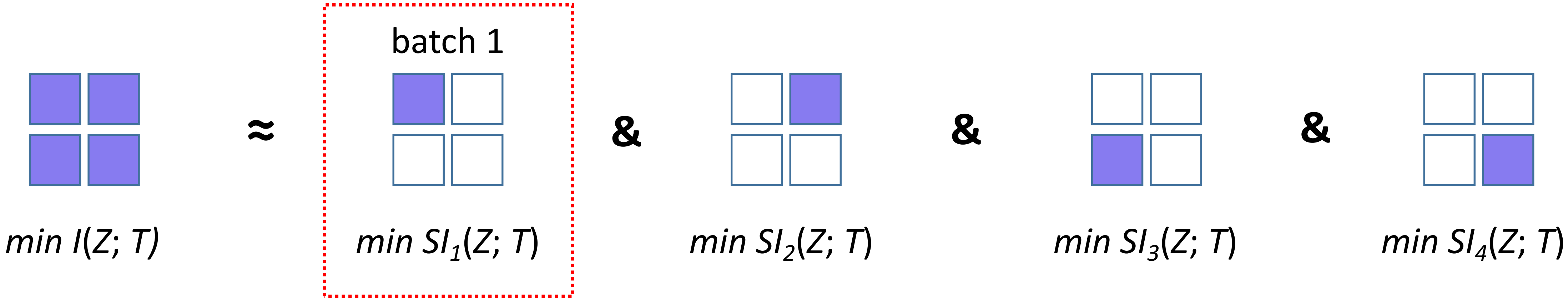


min I(Z; T) *min* SI_1*(Z; T)* *min* SI_2*(Z; T)* *min* SI_3*(Z; T)* *min* SI_4*(Z; T)*

Method

$$\min_f \mathcal{L}(f(X);Y) + \beta \cdot \underline{SI(f(X);T)},$$

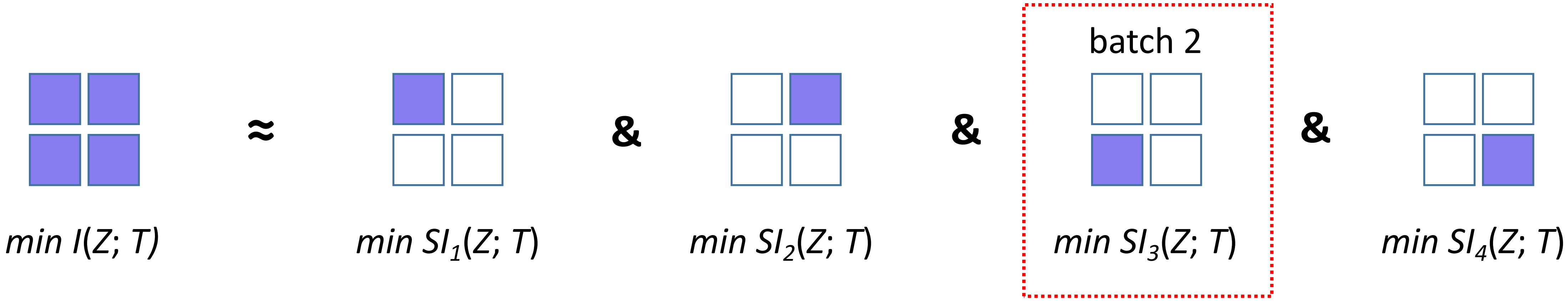
minimise randomly chosen *SI* in different mini-batches



Method

$$\min_{f} \mathcal{L}(f(X);Y) + \beta \cdot \underline{SI(f(X);T)},$$

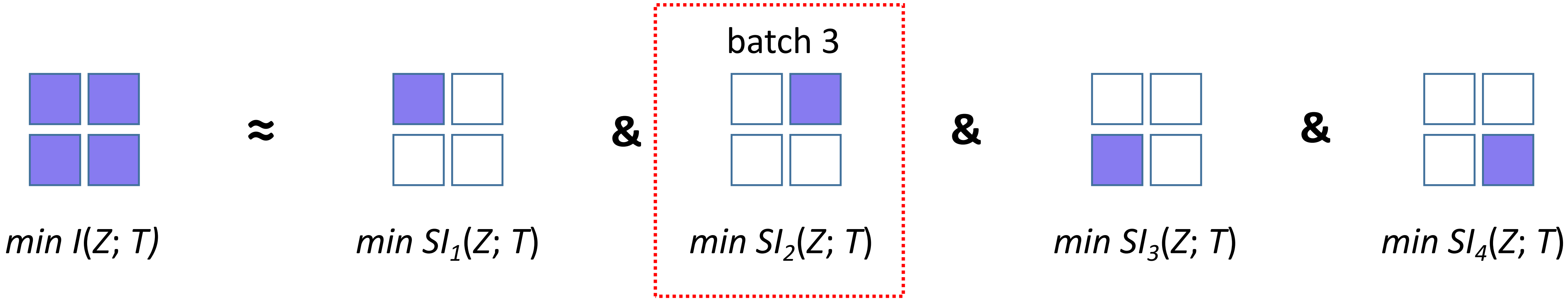
minimise randomly chosen *SI* in different mini-batches



Method

$$\min_f \mathcal{L}(f(X);Y) + \beta \cdot \underline{SI(f(X);T)},$$

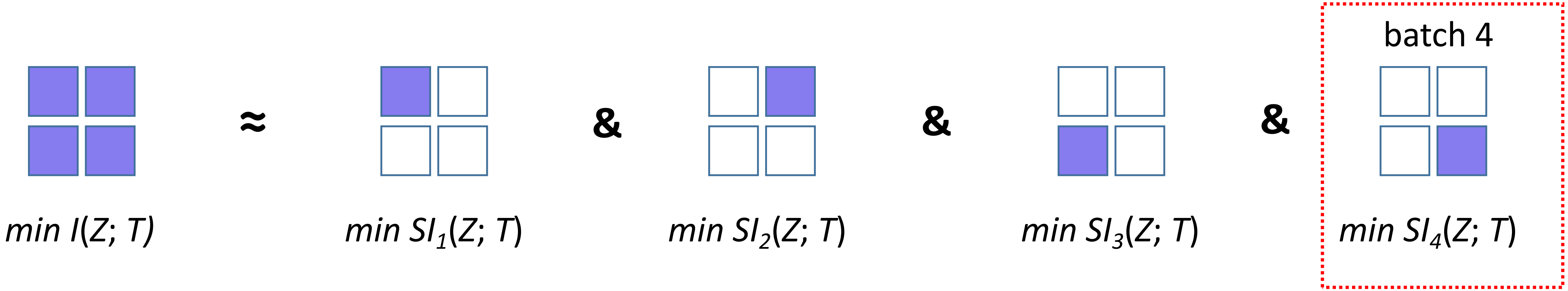
minimise randomly chosen *SI* in different mini-batches



Method

$$\min_f \mathcal{L}(f(X);Y) + \beta \cdot \underline{SI(f(X);T)},$$

minimise randomly chosen *SI* in different mini-batches



Method

Estimate of *SI*

1D polynomials

Pearson corr.

$$SI(Z;T) \approx \sup_{i,j} \sup_{h_i, g_j} \boxed{\rho}(h_i(\theta_i^\top Z), g_j(\phi_j^\top T)),$$

uniformly sampled from unit spheres

*can be solved analytically by eigendecomposition

Method

Properties of *SI*

1. $SI(Z;T) = 0 \quad \Leftrightarrow \quad I(Z;T) = 0$

2. $SI(Z;T) \in [0, 1]$

3. approximation has analytic solution

Method

Adversarial infomin learning

$$\min_f \max_t \mathcal{L}(f(X);Y) + \beta \cdot \hat{I}_t(f(X);T)$$

Estimate *I* by SGD

// Max-step
for l_2 in 1 to L_2 **do**
  $t \leftarrow t + \eta \nabla_t \hat{I}_t(f(X);T)$ with data in $\mathcal{D}'$
end for

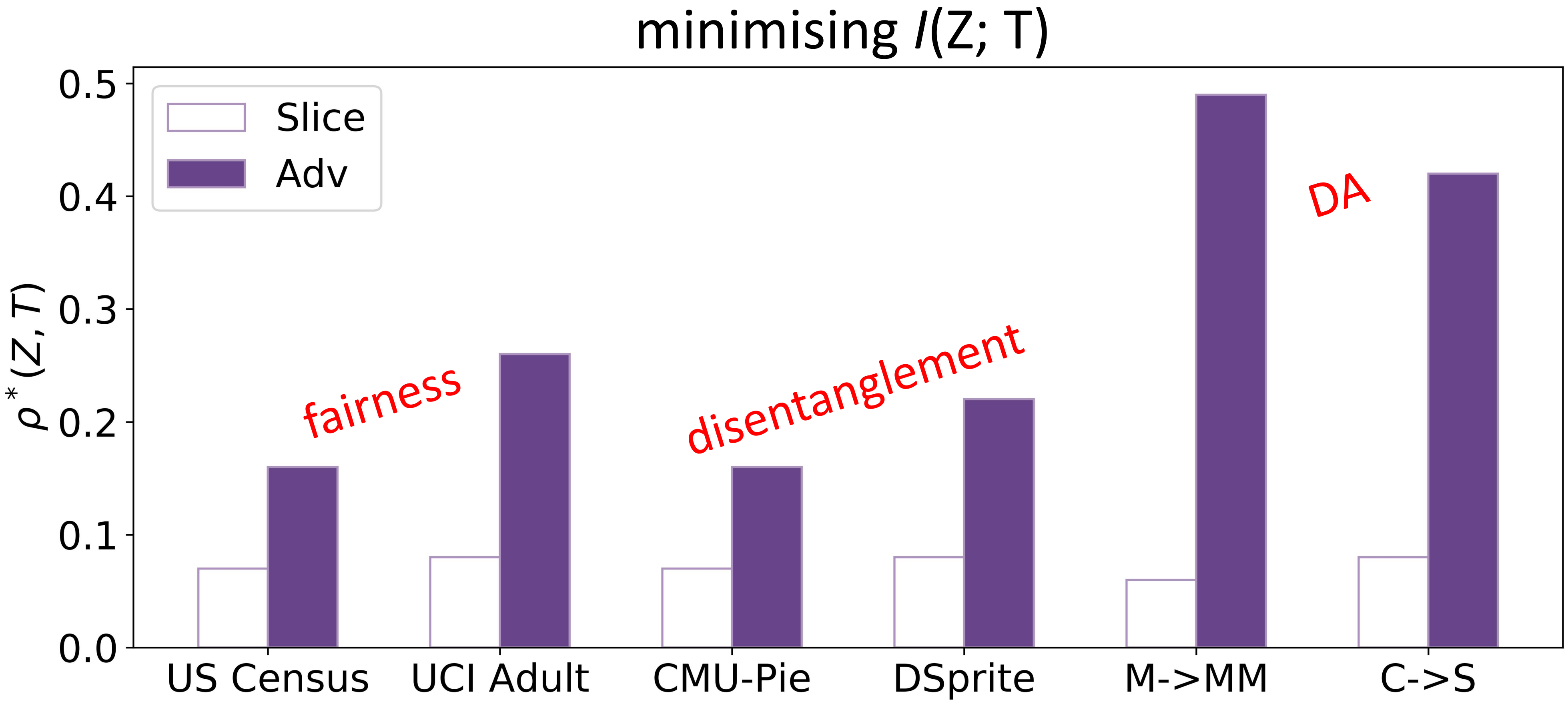
Slice infomin learning

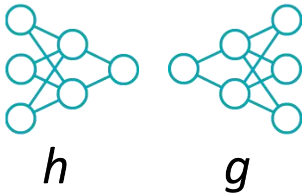
$$\min_f \mathcal{L}(f(X);Y) + \beta \cdot SI(f(X);T),$$

Estimate *SI* analytically

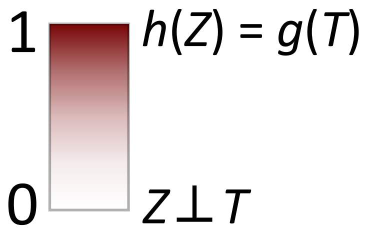
// Max-step
sample S slices $\Theta = \{\theta_i\}_{i=1}^S, \Phi = \{\phi_j\}_{j=1}^S$
solve the parameters w, v in $\hat{SI}$ analytically with $\Theta, \Phi, \mathcal{D}'$ by eigendecomposition

Results



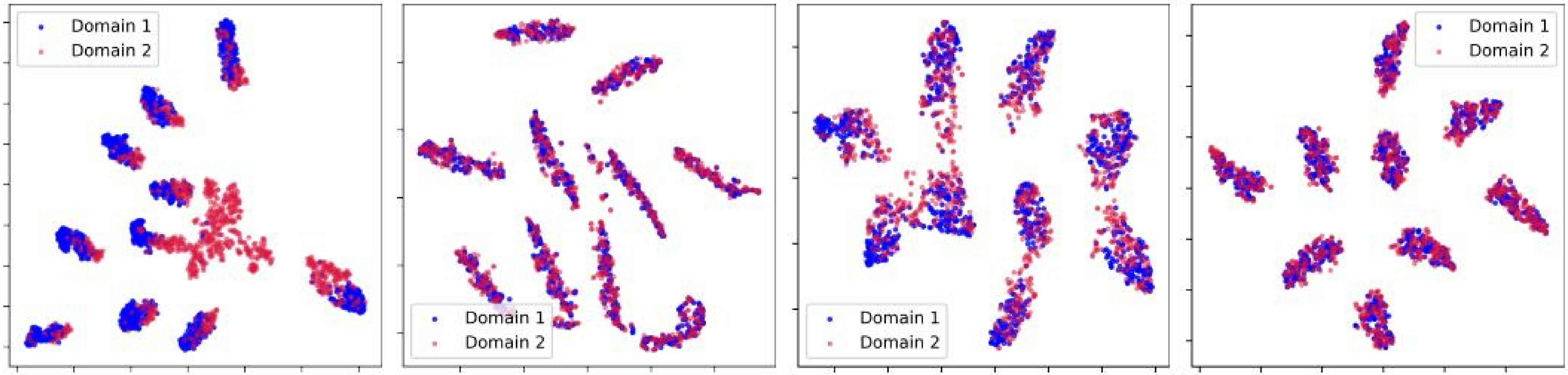


$$\rho^*(Z,T) = \sup_{h,g} \rho(h(Z), g(T))$$



Results

Domain-invariant representation learning

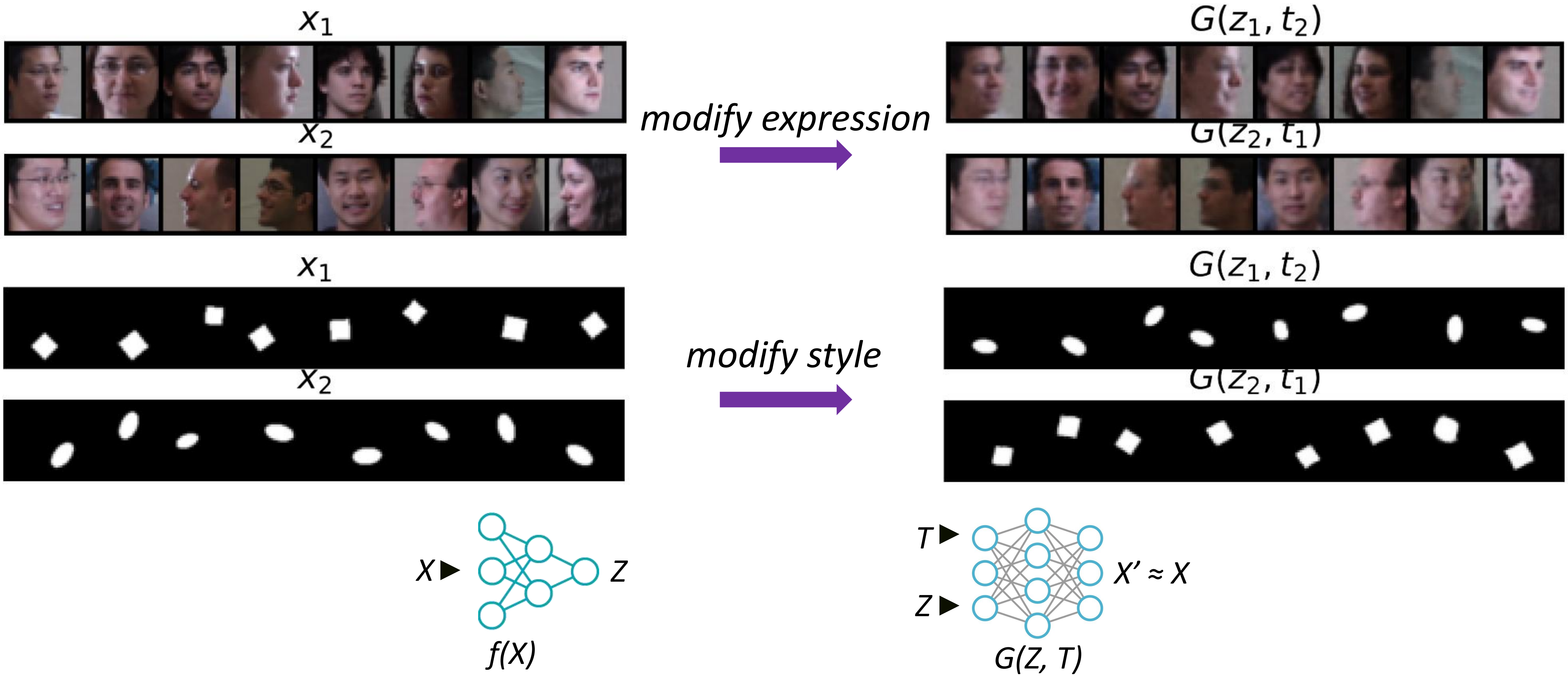


(a) M → MM, adversarial (b) M → MM, slice (c) C → S, adversarial (d) C → S, slice

*M: MNIST, MM: MNIST-M, C: CIFAR-10, S: STL-10

Results

Disentangled representation learning



Results

Fair representation learning

Table 4: Algorithmic fairness: US Census

Neural Rényi				Neural TC			
L_2	$\rho^*(Z,Y)$	$\rho^*(Z,T)$	time (sec./max step)	L_2	$\rho^*(Z,Y)$	$\rho^*(Z,T)$	time (sec./max step)
2	0.95 ± 0.02	0.23 ± 0.10	0.092	3	0.95 ± 0.02	0.27 ± 0.03	0.097
10	0.95 ± 0.02	0.19 ± 0.06	0.642	10	0.95 ± 0.02	0.21 ± 0.02	0.362
50	0.95 ± 0.01	0.06 ± 0.02	2.456	20	0.95 ± 0.01	0.08 ± 0.02	2.146
Slice	0.95 ± 0.01	0.07 ± 0.02	0.102	Slice	0.95 ± 0.00	0.07 ± 0.02	0.102

T = sensitive attribute

L_2 = num of gradient steps for training the adversary

Conclusions

- ***To minimise MI, we do not need to estimate it***

  only consider random facets of MI in each mini-batch

- ***A method widely applicable to many applications***

  fairness, disentanglement, domain adaptation, invariance, ...

Thanks!