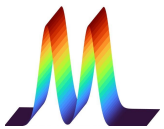
Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation

Yogesh Balaji, Rama Chellappa, Soheil Feizi

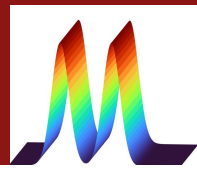
Reading group session#2
Presenter: Piyushi



ML Research Lab
Supervised by Dr J. Saketha Nath

July 18, 2021

Overview



Introduction

Prior Work for Robust OT

Proposed Formulation

Experimental Results

Optimal Transport Kantorovich Formulation

$\min_{\pi \in \Pi(\mathbb{P}_x, \mathbb{P}_y)} \int \int c(x,y)\pi(x,y)dxdy$ gives the minimum cost for transporting the density $\mathbb{P}_X$ to $\mathbb{P}_Y$

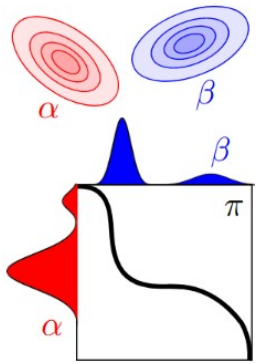


Figure 2: OT with continuous distributions

Optimal Transport Kantorovich Formulation

$\min_{\pi \in \Pi(\mathbb{P}_x, \mathbb{P}_y)} \int \int c(x,y)\pi(x,y)dxdy$ gives the minimum cost for transporting the density $\mathbb{P}_X$ to $\mathbb{P}_Y$

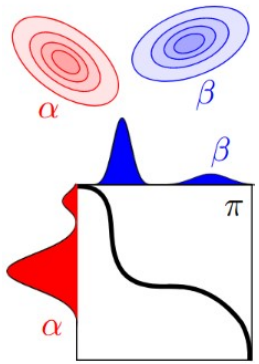


Figure 2: OT with continuous distributions

The optimal value of this problem gives Wasserstein distance.

Optimal Transport Kantorovich Formulation

$\min_{\pi \in \Pi(\mathbb{P}_x, \mathbb{P}_y)} \int \int c(x,y)\pi(x,y)dxdy$ gives the minimum cost for transporting the density $\mathbb{P}_X$ to $\mathbb{P}_Y$

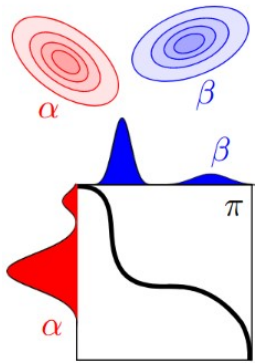


Figure 2: OT with continuous distributions

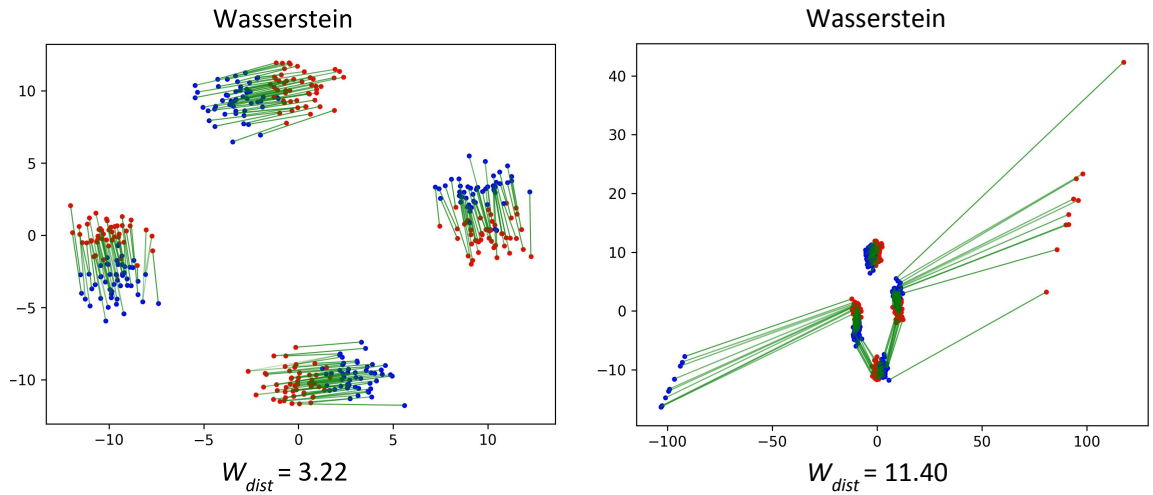
The optimal value of this problem gives Wasserstein distance.

Corresponding Dual

$\max_{\phi \in Lip-1} \int \phi(x)d\mathbb{P}_x - \int \phi(x)d\mathbb{P}_Y$

The notation $\phi \in Lip - L$ means $|\phi(x) - \phi(x')| \leq L||x - x'||$.

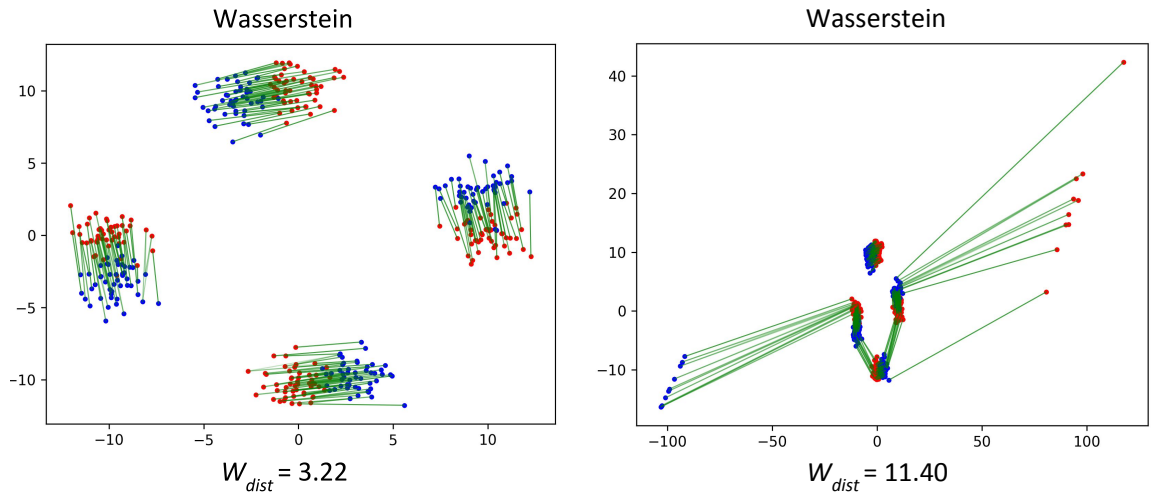
Need for robustness



(a) Clean dataset (b) Dataset with outliers

Figure 3: (a)Wasserstein couplings on clean dataset (b)Wasserstein couplings on dataset with 5% outliers (c)Proposed Robust Wasserstein couplings

Need for robustness

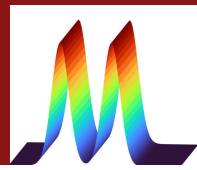


(a) Clean dataset (b) Dataset with outliers

Figure 3: (a)Wasserstein couplings on clean dataset (b)Wasserstein couplings on dataset with 5% outliers (c)Proposed Robust Wasserstein couplings

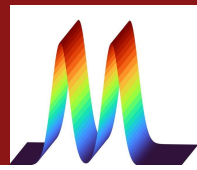
Problem: We can't ignore a point(even if it's an outlier) in the optimal coupling because we need to exactly match the marginals.

Prior Work for Robust OT



- **Relax the marginal constraint with Unbalanced OT formulation**

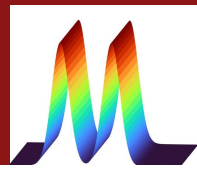
Prior Work for Robust OT



- **Relax the marginal constraint with Unbalanced OT formulation**

$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) =$
$\min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy + \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) + \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y)$

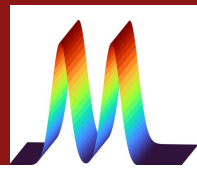
Prior Work for Robust OT



- **Relax the marginal constraint with Unbalanced OT formulation**

$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) =$
$\min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy + \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) + \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y)$
where $\mathcal{D}_f$ is the f-divergence defined as $\mathcal{D}_f(\mathbb{P}||\mathbb{Q}) = \int f(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)})\mathbb{Q}(x)dx$

Prior Work for Robust OT



• **Relax the marginal constraint with Unbalanced OT formulation**

$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) =$
$\min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy + \mathcal{D}_f(\mathbb{P}_{\tilde{X}}||\mathbb{P}_X) + \mathcal{D}_f(\mathbb{P}_{\tilde{Y}}||\mathbb{P}_Y)$
where $\mathcal{D}_f$ is the f-divergence defined as $\mathcal{D}_f(\mathbb{P}||\mathbb{Q}) = \int f(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)})\mathbb{Q}(x)dx$

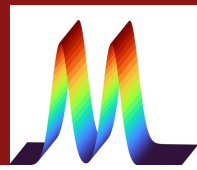
Corresponding Dual:

Define $r^*(x) := \sup_{s>0} \frac{x - f(s)}{s}$ where $f'_\infty := \lim_{s \to \infty} \frac{f(s)}{s}$. Then,

$$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) = \max_{\phi, \psi} \int \phi(x) d\mathbb{P}_X + \int \psi(y) d\mathbb{P}_Y$$
$$\text{s.t. } r^*(\phi(x)) + r^*(\psi(y)) \leq c(x,y)$$

Problem with this: Convergence issues while trying to scale to Deep Learning applications

Proposed Formulation

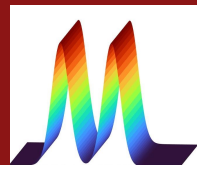


Primal problem for usual Unbalanced Wasserstein:

$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) =$

$\min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy + \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) + \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y)$

Proposed Formulation



Primal problem for usual Unbalanced Wasserstein:

$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) =$

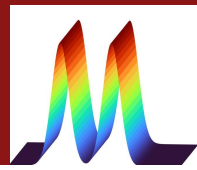
$\min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy + \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) + \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y)$

Reformulating to get Robust-Wasserstein(proposed)

$\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) := \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy$

s.t. $\mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1$, $\mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$

Proposed Formulation



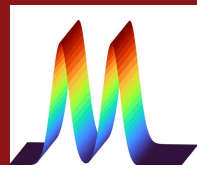
Primal problem for usual Unbalanced Wasserstein:
$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) =$
$\min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy + \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) + \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y)$

Reformulating to get Robust-Wasserstein(proposed)
$\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) := \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy$

s.t. $\mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1$, $\mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$

Key advantage Dual of the above primal Robust-Wasserstein has only 1 dual function with just a Lipschitz constraint.

Proposed Formulation



Primal problem for usual Unbalanced Wasserstein:
$\mathcal{W}^{ub}(\mathbb{P}_X, \mathbb{P}_Y) =$
$\min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy + \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) + \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y)$

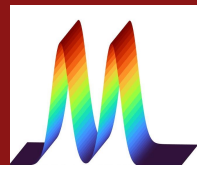
Reformulating to get Robust-Wasserstein(proposed)
$\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) := \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \min_{\pi \in \Pi(\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}})} \int \int c(x,y)\pi(x,y)dxdy$

s.t. $\mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1$, $\mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$

Key advantage Dual of the above primal Robust-Wasserstein has only 1 dual function with just a Lipschitz constraint.

Properties of Robust-Wasserstein: Symmetric when $\rho_1 = \rho_2$, Non-negative, Identity, Doesn't satisfy triangle inequality.

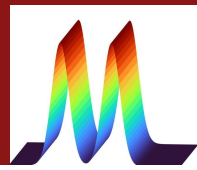
Dual of Robust-Wasserstein



1. Start with the Primal definition: $\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) := \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \min_{\pi} \int\int c(x,y)\pi(x,y)dxdy$

s.t. $\mathcal{D}_f(\mathbb{P}_{\tilde{X}}||\mathbb{P}_X) \leq \rho_1$, $\mathcal{D}_f(\mathbb{P}_{\tilde{Y}}||P_Y) \leq \rho_2$ and $\int \pi(x,y)dy = \mathbb{P}_{\tilde{X}}$, $\int \pi(x,y)dx = \mathbb{P}_{\tilde{Y}}$

Dual of Robust-Wasserstein



1. **Start with the Primal definition:** $\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) := \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \min_{\pi} \int\int c(x,y)\pi(x,y)dxdy$

s.t. $\mathcal{D}_f(\mathbb{P}_{\tilde{X}}||\mathbb{P}_X) \leq \rho_1$, $\mathcal{D}_f(\mathbb{P}_{\tilde{Y}}||P_Y) \leq \rho_2$ and $\int \pi(x,y)dy = \mathbb{P}_{\tilde{X}}$, $\int \pi(x,y)dx = \mathbb{P}_{\tilde{Y}}$

2. **Write Lagrangian Dual problem**

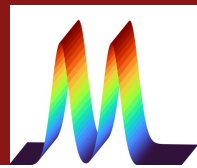
For a convex minimization problem with inequality constraints,

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

the Lagrangian dual problem is

$$\begin{aligned} \underset{u}{\text{maximize}} \quad & \inf_x \left(f(x) + \sum_{j=1}^{m} u_j g_j(x) \right) \\ \text{subject to} \quad & u_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

Dual of Robust-Wasserstein



1. Start with the Primal definition: $\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) := \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \min_{\pi} \int\int c(x,y)\pi(x,y)dxdy$

s.t. $\mathcal{D}_f(\mathbb{P}_{\tilde{X}}||\mathbb{P}_X) \leq \rho_1$, $\mathcal{D}_f(\mathbb{P}_{\tilde{Y}}||P_Y) \leq \rho_2$ and $\int \pi(x,y)dy = \mathbb{P}_{\tilde{X}}$, $\int \pi(x,y)dx = \mathbb{P}_{\tilde{Y}}$

2. Write Lagrangian Dual problem

For a convex minimization problem with inequality constraints,

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

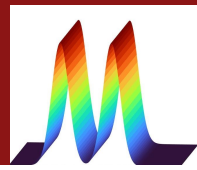
the Lagrangian dual problem is

$$\begin{aligned} \underset{u}{\text{maximize}} \quad & \inf_x \left(f(x) + \sum_{j=1}^{m} u_j g_j(x) \right) \\ \text{subject to} \quad & u_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

Lagrangian wrt marginal constraint for robust Wasserstein:

$$L = \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \min_{\pi>0} \max_{\phi(x),\psi(y)} \int\int c(x,y)\pi(x,y)dxdy + \int \phi(x)\left(\mathbb{P}_{\tilde{X}} - \int \pi(x,y)dy\right)dx + \int \psi(y)\left(\int \pi(x,y)dy - \mathbb{P}_{\tilde{Y}}\right)dy$$
$$\text{s.t.} \mathcal{D}_f(\mathbb{P}_{\tilde{X}}||\mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}}||\mathbb{P}_Y) \leq \rho_2$$

Dual of Robust-Wasserstein

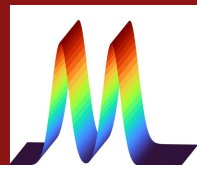


Re-arranging terms

$$= \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \max_{\phi(x), \psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} + \min_{\pi > 0} \int \int [c(x, y) - \phi(x) + \psi(y)] \pi(x, y) dx dy$$

$$\text{s.t.} \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$$

Dual of Robust-Wasserstein



Re-arranging terms

$$= \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \max_{\phi(x), \psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} + \min_{\pi > 0} \int \int [c(x, y) - \phi(x) + \psi(y)] \pi(x, y) dx dy$$

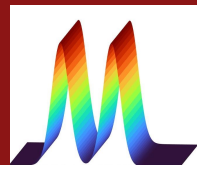
$$\text{s.t.} \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$$

Simplifying

$$= \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \max_{\phi(x), \psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} + \left\{ \begin{array}{cc} 0 & \text{if } c(x, y) - \phi(x) + \psi(y) \geq 0 \ \forall \ x, y \in \mathcal{X}, \\ -\infty & \text{otherwise} \end{array} \right\}$$

$$\text{s.t.} \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$$

Dual of Robust-Wasserstein



Re-arranging terms

$$= \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \max_{\phi(x), \psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} + \min_{\pi > 0} \int \int [c(x, y) - \phi(x) + \psi(y)] \pi(x, y) dx dy$$

$$\text{s.t.} \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$$

Simplifying

$$= \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \max_{\phi(x), \psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} + \left\{ \begin{array}{ll} 0 & \text{if } c(x, y) - \phi(x) + \psi(y) \geq 0 \ \forall \ x, y \in \mathcal{X}, \\ -\infty & \text{otherwise} \end{array} \right\}$$

$$\text{s.t.} \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$$

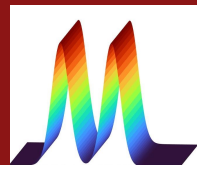
$$\mathcal{W}^{rob}_{\rho_1, \rho_2}(\mathbb{P}_X, \mathbb{P}_Y) = \min_{\mathbb{P}_{\tilde{X}}, \mathbb{P}_{\tilde{Y}}} \max_{\phi(x), \psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}}$$

$$\text{s.t.} \phi(x) - \psi(y) \leq c(x, y)$$

$$\mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2$$

This expression for Dual can be further simplified when the distributions lie in a metric space.

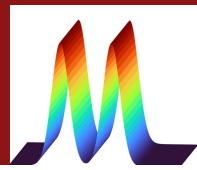
Further Simplifying Dual



We know that the dual problem satisfies

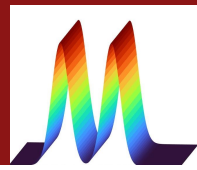
$$\phi(x) \leq \psi(y) + c(x,y) \implies \phi(x) \leq \inf_y \psi(y) + c(x,y).$$

Further Simplifying Dual



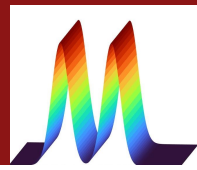
We know that the dual problem satisfies
$\phi(x) \leq \psi(y) + c(x,y) \implies \phi(x) \leq \inf_y \psi(y) + c(x,y)$.
Define $k(x) := \inf_y c(x,y) + \psi(y)$

Further Simplifying Dual



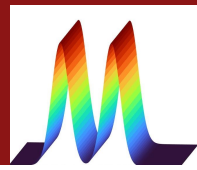
We know that the dual problem satisfies
$\phi(x) \leq \psi(y) + c(x,y) \implies \phi(x) \leq \inf_y \psi(y) + c(x,y)$.
Define $k(x) := \inf_y c(x,y) + \psi(y)$ then $\phi(x) \leq k(x)$

Further Simplifying Dual



We know that the dual problem satisfies
$\phi(x) \leq \psi(y) + c(x,y) \implies \phi(x) \leq \inf_y \psi(y) + c(x,y)$.
Define $k(x) := \inf_y c(x,y) + \psi(y)$ then $\phi(x) \leq k(x)$
Also, by definition, $k(y) := \inf_y c(y,y) + \psi(y) \implies k(y) \leq \psi(y)$.

Further Simplifying Dual



We know that the dual problem satisfies
$\phi(x) \leq \psi(y) + c(x,y) \implies \phi(x) \leq \inf_y \psi(y) + c(x,y)$.
Define $k(x) := \inf_y c(x,y) + \psi(y)$ then $\phi(x) \leq k(x)$
Also, by definition, $k(y) := \inf_y c(y,y) + \psi(y) \implies k(y) \leq \psi(y)$.
Hence, $\phi(x) \leq k(x) \leq \psi(x)$

Further Simplifying Dual

We know that the dual problem satisfies
$\phi(x) \leq \psi(y) + c(x,y) \implies \phi(x) \leq \inf_y \psi(y) + c(x,y)$.
Define $k(x) := \inf_y c(x,y) + \psi(y)$ then $\phi(x) \leq k(x)$
Also, by definition, $k(y) := \inf_y c(y,y) + \psi(y) \implies k(y) \leq \psi(y)$.
Hence, $\phi(x) \leq k(x) \leq \psi(x)$

The function k(x) also satisfies the 1-Lipschitz criteria:

$$
\begin{aligned}
|k(x) - k(x')| &= |\inf_y [c(x,y) + \psi(y)] - \inf_y [c(x',y) + \psi(y)]| \\
&\leq |[c(x,y) + \psi(y)] - \inf_y [c(x',y) + \psi(y)]| \\
&= |\inf_y [c(x',y) + \psi(y)] - [c(x,y) + \psi(y)]| \\
&\leq |c(x',y) + \psi(y) - [c(x,y) + \psi(y)]|
\end{aligned}
$$

Hence, $|k(x) - k(x')| \leq |c(x,y) - c(x',y)|$

Further Simplifying Dual

Using $\phi(x) \leq k(x) \leq \psi(x)$ in the following

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) = \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{\phi(x),\psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} \\ \text{s.t.} \phi(x) - \psi(y) \leq c(x,y) \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{1}$$

Further Simplifying Dual

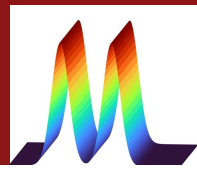
Using $\phi(x) \leq k(x) \leq \psi(x)$ in the following

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) = \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{\phi(x),\psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} \\ \text{s.t.} \phi(x) - \psi(y) \leq c(x,y) \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{1}$$

it becomes,

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{k \in Lip-1} \int k(x) d\mathbb{P}_{\tilde{X}} - \int k(y) d\mathbb{P}_{\tilde{Y}} \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{2}$$

Further Simplifying Dual



Using $\phi(x) \leq k(x) \leq \psi(x)$ in the following

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) = \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{\phi(x),\psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} \\ \text{s.t.} \phi(x) - \psi(y) \leq c(x,y) \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{1}$$

it becomes,

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{k \in Lip-1} \int k(x) d\mathbb{P}_{\tilde{X}} - \int k(y) d\mathbb{P}_{\tilde{Y}} \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{2}$$

It can also be seen that $\phi(x) = k(x), \psi(y) = k(y)$ is a feasible solution of eq(1) so it is also true that

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) \geq \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{k \in Lip-1} \int k(x) d\mathbb{P}_{\tilde{X}} - \int k(y) d\mathbb{P}_{\tilde{Y}} \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{3}$$

Further Simplifying Dual

Using $\phi(x) \leq k(x) \leq \psi(x)$ in the following

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) = \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{\phi(x),\psi(y)} \int \phi(x) d\mathbb{P}_{\tilde{X}} - \int \psi(y) d\mathbb{P}_{\tilde{Y}} \\ \text{s.t.} \phi(x) - \psi(y) \leq c(x,y) \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{1}$$

it becomes,

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{k \in Lip-1} \int k(x) d\mathbb{P}_{\tilde{X}} - \int k(y) d\mathbb{P}_{\tilde{Y}} \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{2}$$

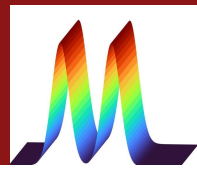
It can also be seen that $\phi(x) = k(x), \psi(y) = k(y)$ is a feasible solution of eq(1) so it is also true that

$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) \geq \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{k \in Lip-1} \int k(x) d\mathbb{P}_{\tilde{X}} - \int k(y) d\mathbb{P}_{\tilde{Y}} \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned} \tag{3}$$

From 2 & 3, the **dual robust Wasserstein becomes**

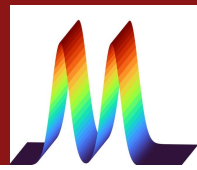
$$\begin{aligned} \mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X, \mathbb{P}_Y) = \min_{\mathbb{P}_{\tilde{X}},\mathbb{P}_{\tilde{Y}}} \max_{k \in Lip-1} \int k(x) d\mathbb{P}_{\tilde{X}} - \int k(y) d\mathbb{P}_{\tilde{Y}} \\ \mathcal{D}_f(\mathbb{P}_{\tilde{X}} || \mathbb{P}_X) \leq \rho_1, \mathcal{D}_f(\mathbb{P}_{\tilde{Y}} || \mathbb{P}_Y) \leq \rho_2 \end{aligned}$$

Continuous Stochastic Relaxation



Continuous Stochastic Relaxation: Let the supports of $\mathbb{P}_{\tilde{X}}$ & $\mathbb{P}_X$ match.
Then $\mathbb{P}_{\tilde{X}}(x)$ can be reparameterized as $W_x(x)\mathbb{P}_X(x)$ where $W_x(.)$ is a weight function which can be implemented using Neural Network.

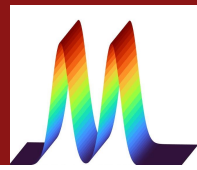
Continuous Stochastic Relaxation



Continuous Stochastic Relaxation: Let the supports of $\mathbb{P}_{\tilde{X}}$ & $\mathbb{P}_X$ match.
Then $\mathbb{P}_{\tilde{X}}(x)$ can be reparameterized as $W_x(x)\mathbb{P}_X(x)$ where $W_x(.)$ is a weight function which can be implemented using Neural Network.

- For $\mathbb{P}_{\tilde{X}}$ to be valid PDF, $\int W_x(x)\mathbb{P}_x(x)dx = 1 \implies \mathbb{E}[W_x(x)] = 1$
- $\mathcal{D}_f(\mathbb{P}_{\tilde{X}}||\mathbb{P}_X) \leq \rho$ becomes $\int f(\frac{W_x(x)\mathbb{P}_X(x)}{\mathbb{P}_X(x)})\mathbb{P}_X(x)dx \leq \rho$
  $\implies \mathbb{E}_{x\sim\mathbb{P}_X}[f(W_x(x))] \leq \rho$. For $\mathcal{X}^2$-div, $\mathbb{E}_{x\sim\mathbb{P}_X}[(W_x(x)-1)^2] \leq 2\rho$
- In the obj, $\int k(x)\mathbb{P}_{\tilde{X}}dx$ becomes $k(x)W_x(x)\mathbb{P}_X(x) = \mathbb{E}_{x\sim\mathbb{P}_X}[W_x(x)k(x)]$

Continuous Stochastic Relaxation



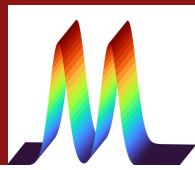
Continuous Stochastic Relaxation: Let the supports of $\mathbb{P}_{\tilde{X}}$ & $\mathbb{P}_X$ match.
Then $\mathbb{P}_{\tilde{X}}(x)$ can be reparameterized as $W_x(x)\mathbb{P}_X(x)$ where $W_x(.)$ is a weight function which can be implemented using Neural Network.

- For $\mathbb{P}_{\tilde{X}}$ to be valid PDF, $\int W_x(x)\mathbb{P}_x(x)dx = 1 \implies \mathbb{E}[W_x(x)] = 1$
- $\mathcal{D}_f(\mathbb{P}_{\tilde{X}}||\mathbb{P}_X) \leq \rho$ becomes $\int f(\frac{W_x(x)\mathbb{P}_X(x)}{\mathbb{P}_X(x)})\mathbb{P}_X(x)dx \leq \rho$
  $\implies \mathbb{E}_{x\sim\mathbb{P}_X}[f(W_x(x))] \leq \rho$. For $\mathcal{X}^2$-div, $\mathbb{E}_{x\sim\mathbb{P}_X}[(W_x(x)-1)^2] \leq 2\rho$
- In the obj, $\int k(x)\mathbb{P}_{\tilde{X}}dx$ becomes $k(x)W_x(x)\mathbb{P}_X(x) = \mathbb{E}_{x\sim\mathbb{P}_X}[W_x(x)k(x)]$

$$\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X,\mathbb{P}_Y) = \min_{W_x,W_y}\max_{k\in Lip-1} \mathbb{E}_{x\sim\mathbb{P}_X}[Wx(x)k(x)] - \mathbb{E}_{y\sim\mathbb{P}_Y}[Wy(y)k(y)]$$

s.t. $\mathbb{E}[f(W_x(x))] \leq \rho_1$, $\mathbb{E}[f(W_x(y))] \leq \rho_2$
$\mathbb{E}[W_x(x)] = 1$, $\mathbb{E}[W_y(y)] = 1$, $W_x(x) \geq 0, W_y(y) \geq 0$

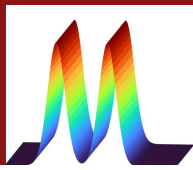
Provable Robustness



Theorem

Let $\mathbb{P}_X$ *&* $\mathbb{P}_Y$ *be two distributions such that* $\mathbb{P}_X = (1-\gamma)\mathbb{P}_X^c + \gamma\mathbb{P}_X^a$, *where* $\mathbb{P}_X^c$ *is the clean distribution &* $\mathbb{P}_X^a$ *is the anomaly distribution. Let* $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$, *with* $k \geq 1$. *Then* $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_y) \leq \max(1, 1 + k\gamma - k\sqrt{2\rho\gamma(1-\gamma)})\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

Provable Robustness



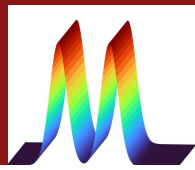
Theorem

Let $\mathbb{P}_X$ *&* $\mathbb{P}_Y$ *be two distributions such that* $\mathbb{P}_X = (1-\gamma)\mathbb{P}_X^c + \gamma\mathbb{P}_X^a$, *where* $\mathbb{P}_X^c$ *is the clean distribution &* $\mathbb{P}_X^a$ *is the anomaly distribution. Let* $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$, *with* $k \geq 1$. *Then* $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_y) \leq \max(1, 1 + k\gamma - k\sqrt{2\rho\gamma(1-\gamma)})\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

Proof-Sketch

1. Upper bound the obj of $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y)$ by (weighted)sum of obj of $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

Provable Robustness



Theorem

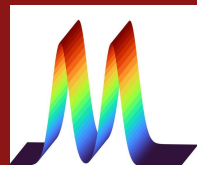
Let $\mathbb{P}_X$ & $\mathbb{P}_Y$ be two distributions such that $\mathbb{P}_X = (1-\gamma)\mathbb{P}_X^c + \gamma\mathbb{P}_X^a$, where $\mathbb{P}_X^c$ is the clean distribution & $\mathbb{P}_X^a$ is the anomaly distribution. Let $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$, with $k \geq 1$. Then $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_y) \leq \max(1, 1 + k\gamma - k\sqrt{2\rho\gamma(1-\gamma)})\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

Proof-Sketch

1. Upper bound the obj of $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y)$ by (weighted)sum of obj of $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

   1.1 This can be done by taking a transport map which is a convex combination of optimal maps of $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$.
   Under certain constraints, this transport map will be a feasible for $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y)$

Provable Robustness



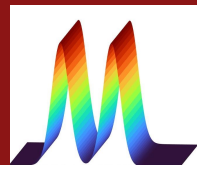
Theorem

Let $\mathbb{P}_X$ & $\mathbb{P}_Y$ be two distributions such that $\mathbb{P}_X = (1-\gamma)\mathbb{P}_X^c + \gamma\mathbb{P}_X^a$, where $\mathbb{P}_X^c$ is the clean distribution & $\mathbb{P}_X^a$ is the anomaly distribution. Let $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$, with $k \geq 1$. Then $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_y) \leq \max(1, 1 + k\gamma - k\sqrt{2\rho\gamma(1-\gamma)})\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

Proof-Sketch

1. Upper bound the obj of $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y)$ by (weighted)sum of obj of $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
   - 1.1 This can be done by taking a transport map which is a convex combination of optimal maps of $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$. Under certain constraints, this transport map will be a feasible for $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y)$
2. Apply triangle inequality $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) \leq \mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_X^c) + \mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & use $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ to eliminate $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ from the obj

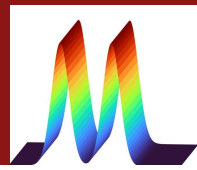
Provable Robustness



Proof.

- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.

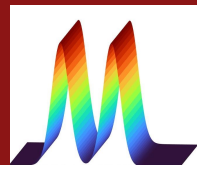
Provable Robustness



Proof.

- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.
- Consider a transport plan $\beta\pi^{a*} + (1-\beta)\pi^{c*}$ where π^{c*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & π^{a*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $0 \leq \beta \leq 1$

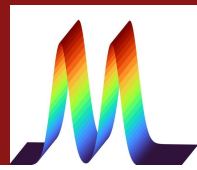
Provable Robustness



Proof.

- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.
- Consider a transport plan $\beta\pi^{a*} + (1-\beta)\pi^{c*}$ where π^{c*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & π^{a*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $0 \leq \beta \leq 1$
- Feasibility of this plan:
  1. The marginals are $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dx = \beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c$ and $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dy = \mathbb{P}_Y$

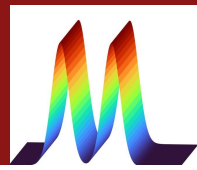
Provable Robustness



Proof.

- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.
- Consider a transport plan $\beta\pi^{a*} + (1-\beta)\pi^{c*}$ where π^{c*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & π^{a*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $0 \leq \beta \leq 1$
- Feasibility of this plan:
  1. The marginals are $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dx = \beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c$ and $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dy = \mathbb{P}_Y$
  2. The constraint $\mathcal{D}_{\mathcal{X}^2}(\beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c || \gamma\mathbb{P}_X^a + (1-\gamma)\mathbb{P}_X^c) \leq \rho$ becomes $(\beta-\gamma)^2 \leq 2\rho\gamma(1-\gamma)$

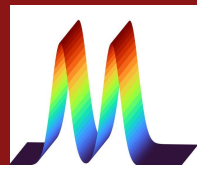
Provable Robustness



Proof.

- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.
- Consider a transport plan $\beta\pi^{a*} + (1-\beta)\pi^{c*}$ where π^{c*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & π^{a*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $0 \leq \beta \leq 1$
- Feasibility of this plan:
  1. The marginals are $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dx = \beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c$ and $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dy = \mathbb{P}_Y$
  2. The constraint $\mathcal{D}_{\mathcal{X}^2}(\beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c || \gamma\mathbb{P}_X^a + (1-\gamma)\mathbb{P}_X^c) \leq \rho$ becomes $(\beta-\gamma)^2 \leq 2\rho\gamma(1-\gamma)$
- The obj $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_\beta \sum_i \sum_j (\beta\pi_{i,j}^{a*} + (1-\beta)\pi_{i,j}^{c*})c_{i,j}$
$= \min_\beta \beta\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) + (1-\beta)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

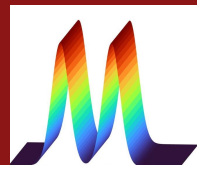
Provable Robustness



Proof.

- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.
- Consider a transport plan $\beta\pi^{a*} + (1-\beta)\pi^{c*}$ where π^{c*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & π^{a*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $0 \leq \beta \leq 1$
- Feasibility of this plan:
  1. The marginals are $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dx = \beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c$ and $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dy = \mathbb{P}_Y$
  2. The constraint $\mathcal{D}_{\mathcal{X}^2}(\beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c || \gamma\mathbb{P}_X^a + (1-\gamma)\mathbb{P}_X^c) \leq \rho$ becomes $(\beta-\gamma)^2 \leq 2\rho\gamma(1-\gamma)$
- The obj $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_\beta \sum_i \sum_j (\beta\pi_{i,j}^{a*} + (1-\beta)\pi_{i,j}^{c*})c_{i,j}$
  $= \min_\beta \beta\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) + (1-\beta)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
- Using triangle ineq, $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) \leq \mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_X^c) + \mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$.
  Using the assumption $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$,
  $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) \leq (k+1)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$

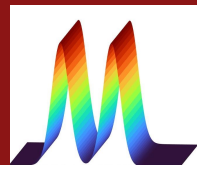
Provable Robustness



Proof.

- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.
- Consider a transport plan $\beta\pi^{a*} + (1-\beta)\pi^{c*}$ where π^{c*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & π^{a*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $0 \leq \beta \leq 1$
- Feasibility of this plan:
  1. The marginals are $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dx = \beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c$ and $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dy = \mathbb{P}_Y$
  2. The constraint $\mathcal{D}_{\mathcal{X}^2}(\beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c || \gamma\mathbb{P}_X^a + (1-\gamma)\mathbb{P}_X^c) \leq \rho$ becomes $(\beta-\gamma)^2 \leq 2\rho\gamma(1-\gamma)$
- The obj $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_\beta \sum_i \sum_j (\beta\pi_{i,j}^{a*} + (1-\beta)\pi_{i,j}^{c*})c_{i,j}$
  $= \min_\beta \beta\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) + (1-\beta)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
- Using triangle ineq, $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) \leq \mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_X^c) + \mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$.
  Using the assumption $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$,
  $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) \leq (k+1)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
- Hence, $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_\beta (1+\beta k)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
  s.t. $(\beta-\gamma)^2 \leq 2\rho\gamma(1-\gamma)$

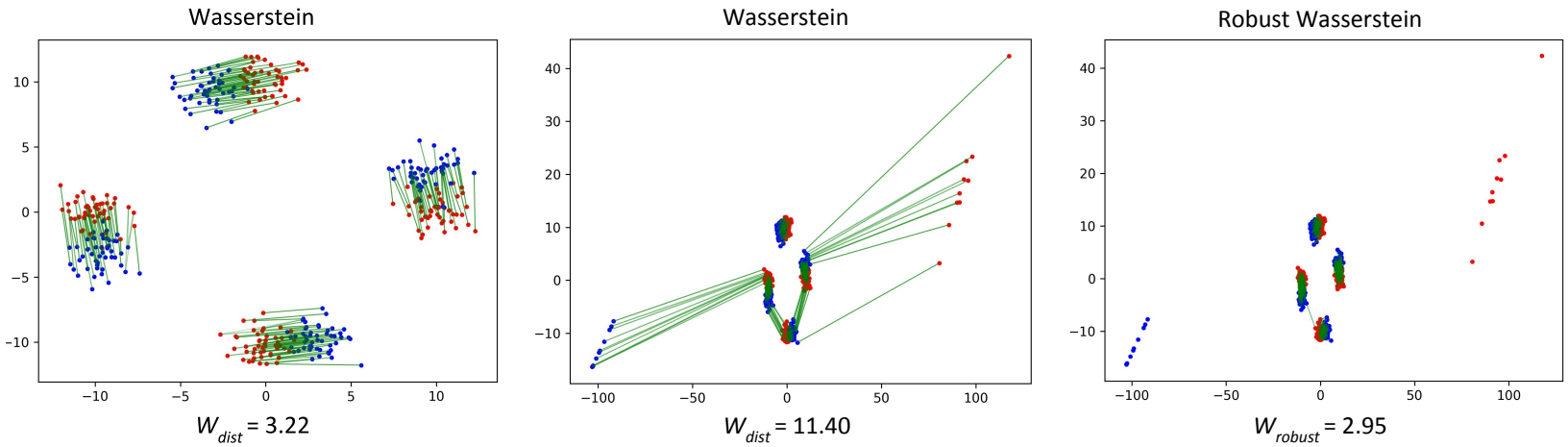
Provable Robustness



Proof.

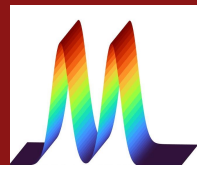
- Consider empirical distributions with $\{x_i^a\}_{i=1}^{n_a}$ samples from $\mathbb{P}_X^a$, $\{x_i^c\}_{i=1}^{n_c}$ samples from $\mathbb{P}_X^c$ & $\{y_i\}_{i=1}^{m}$ samples from $\mathbb{P}_Y$.
- Consider a transport plan $\beta\pi^{a*} + (1-\beta)\pi^{c*}$ where π^{c*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$ & π^{a*} is optimal plan for $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y)$ & $0 \leq \beta \leq 1$
- Feasibility of this plan:
  1. The marginals are $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dx = \beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c$ and $\int \beta\pi^{a*} + (1-\beta)\pi^{c*} dy = \mathbb{P}_Y$
  2. The constraint $\mathcal{D}_{\mathcal{X}^2}(\beta\mathbb{P}_X^a + (1-\beta)\mathbb{P}_X^c || \gamma\mathbb{P}_X^a + (1-\gamma)\mathbb{P}_X^c) \leq \rho$ becomes $(\beta-\gamma)^2 \leq 2\rho\gamma(1-\gamma)$
- The obj $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_\beta \sum_i \sum_j (\beta\pi_{i,j}^{a*} + (1-\beta)\pi_{i,j}^{c*})c_{i,j}$
  $= \min_\beta \beta\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) + (1-\beta)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
- Using triangle ineq, $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) \leq \mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_X^c) + \mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$.
  Using the assumption $\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_X^a) = k\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$,
  $\mathcal{W}(\mathbb{P}_X^a, \mathbb{P}_Y) \leq (k+1)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
- Hence, $\mathcal{W}_{\rho,0}^{rob}(\mathbb{P}_X, \mathbb{P}_Y) \leq \min_\beta (1+\beta k)\mathcal{W}(\mathbb{P}_X^c, \mathbb{P}_Y)$
  s.t. $(\beta-\gamma)^2 \leq 2\rho\gamma(1-\gamma)$
- $\beta_{min} = \gamma - \sqrt{2\rho\gamma(1-\gamma)}$

Experimental Results



(a) Clean dataset (b) Dataset with outliers (c) Dataset with outliers

Figure 4: Toy Example

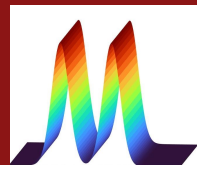


Continuous Stochastic Formulation:

$$\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X,\mathbb{P}_Y) = \min_{\mathbf{W_x},\mathbf{W_y}} \max_{\mathbf{D}\in Lip-1} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[\mathbf{W_x}(\mathbf{x})\mathbf{D}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim\mathbb{P}_Y}[\mathbf{W_y}(\mathbf{y})\mathbf{D}(\mathbf{y})]$$

s.t. $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[(\mathbf{W_x}(\mathbf{x})-1)^2] \leq 2\rho_1, \mathbb{E}_{\mathbf{y}\sim\mathbb{P}_Y}[(W_{\mathbf{y}}(\mathbf{y})-1)^2] \leq 2\rho_2$
$\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[\mathbf{W_x}(\mathbf{x})] = 1, \mathbb{E}_{\mathbf{y}\sim\mathbb{P}_Y}[\mathbf{W_y}(\mathbf{y})] = 1, \mathbf{W_x}(\mathbf{x}) \geq 0, \mathbf{W_y}(\mathbf{y}) \geq 0$

Generative Modeling



Continuous Stochastic Formulation:

$$\mathcal{W}^{rob}_{\rho_1,\rho_2}(\mathbb{P}_X,\mathbb{P}_Y) = \min_{\mathbf{W_x},\mathbf{W_y}} \max_{\mathbf{D}\in Lip-1} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[\mathbf{W_x}(\mathbf{x})\mathbf{D}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim\mathbb{P}_Y}[\mathbf{W_y}(\mathbf{y})\mathbf{D}(\mathbf{y})]$$

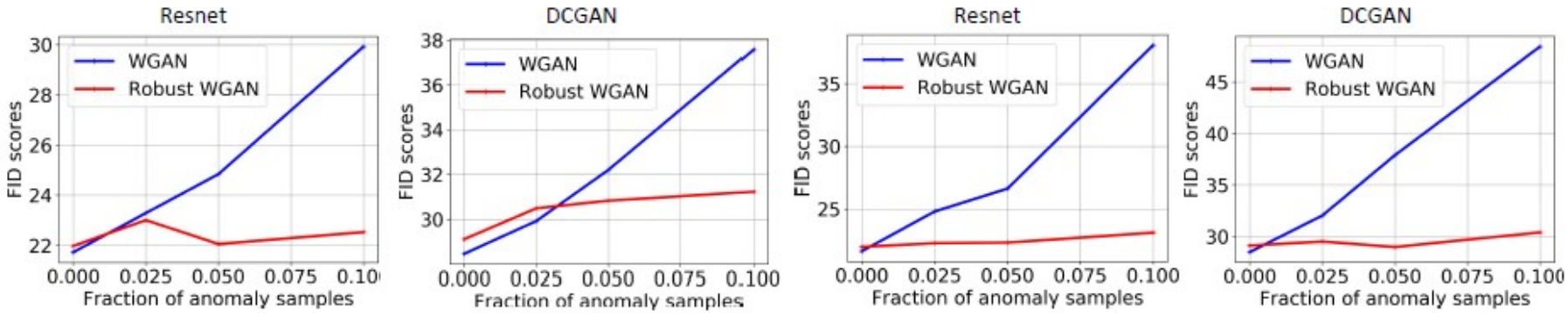
s.t. $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[(\mathbf{W_x}(\mathbf{x})-1)^2] \leq 2\rho_1, \mathbb{E}_{\mathbf{y}\sim\mathbb{P}_Y}[(W_{\mathbf{y}}(\mathbf{y})-1)^2] \leq 2\rho_2$
$\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[\mathbf{W_x}(\mathbf{x})] = 1, \mathbb{E}_{\mathbf{y}\sim\mathbb{P}_Y}[\mathbf{W_y}(\mathbf{y})] = 1, \mathbf{W_x}(\mathbf{x}) \geq 0, \mathbf{W_y}(\mathbf{y}) \geq 0$

Formulation for Generative Modeling:
Obj: To minimize robust-wasserstein measure between real & generated data distribution.

$$\min_{\mathbf{W},\mathbf{G}} \max_{\mathbf{D}\in Lip-1} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[\mathbf{W}(\mathbf{x})\mathbf{D}(\mathbf{x})] - \mathbb{E}_{\mathbf{z}}[\mathbf{D}(\mathbf{G}(\mathbf{z}))]$$

s.t. $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[(\mathbf{W}(\mathbf{x})-1)^2] \leq 2\rho, \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_X}[\mathbf{W}(\mathbf{x})] = 1, \mathbf{W}(\mathbf{x}) \geq 0$

Generative Modeling results



(a) CIFAR-10 + MNIST

(b) CIFAR-10 + Uniform noise

Figure 2: FID scores of GAN models trained on CIFAR-10 corrupted with outlier noise. In (a), samples from MNIST dataset are used as the outliers, while in (b), uniform noise is used. FID scores of WGAN increase with the increase in outlier fraction, while robust WGAN maintains FID scores.

Generative Modeling results (cont.)

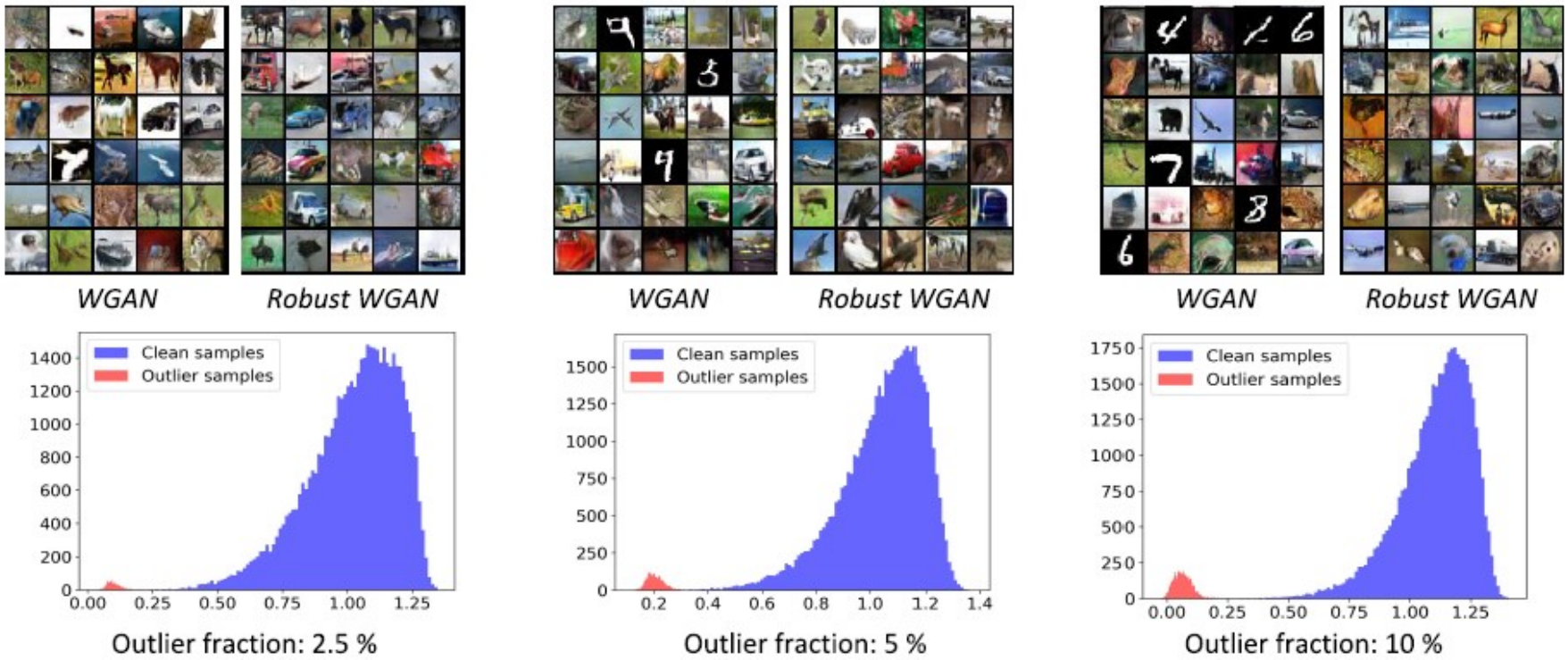
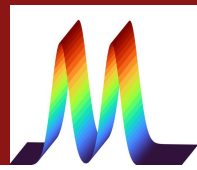


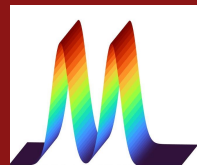
Figure 5: Top row: generated images; Bottom row: Output of $\mathbf{W}(.)$

Domain Adaptation



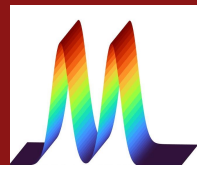
Unsupervised Domain Adaptation (UDA): Given a labeled source dataset $\mathbb{P}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ and an unlabeled target dataset $\mathbb{P}_t = \{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$, train classification model on source while minimizing a distributional distance between source and target distributions.

Domain Adaptation



Unsupervised Domain Adaptation (UDA): Given a labeled source dataset $\mathbb{P}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ and an unlabeled target dataset $\mathbb{P}_t = \{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$, train classification model on source while minimizing a distributional distance between source and target distributions.
The paper uses discrete formulation with the perturbed distribution $\mathbb{P}_{\tilde{t}}$ as weighted empirical distribution, i.e. $\mathbb{P}_t^{(n_t)}(\mathbf{x}_i) = \frac{1}{n_t}$ and $\mathbb{P}_{\tilde{t}}(\mathbf{x}_i) = w_i^x$.

Domain Adaptation



Unsupervised Domain Adaptation (UDA): Given a labeled source dataset $\mathbb{P}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ and an unlabeled target dataset $\mathbb{P}_t = \{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$, train classification model on source while minimizing a distributional distance between source and target distributions.
The paper uses discrete formulation with the perturbed distribution $\mathbb{P}_{\tilde{t}}$ as weighted empirical distribution, i.e. $\mathbb{P}_t^{(n_t)}(\mathbf{x}_i) = \frac{1}{n_t}$ and $\mathbb{P}_{\tilde{t}}(\mathbf{x}_i) = w_i^x$.
Formulation for UDA:

$$\min_{\mathbf{F},\mathbf{C}} \frac{1}{n_s} \sum_i \mathbf{C}(\mathbf{F}(x_i), y_i) + \lambda [\min_{\mathbf{w} \in \Delta^{n_t}} \max_{\mathbf{D} \in Lip-1} \frac{1}{n_s} \sum_i \mathbf{D}(\mathbf{F}(\mathbf{x}_i^s)) - \frac{1}{n_t} \sum_j w_j \mathbf{D}(\mathbf{F}(\mathbf{x}_j^t))]$$

s.t. $||n_t \mathbf{w} - 1||_2 \leq \sqrt{2\rho n_t}$
where F: feature network
C: classifier

UDA results

Table 2: Cross-domain recognition accuracy on VISDA-17 dataset using Resnet-18 model averaged over 3 runs.

Method	Accuracy (in %)
Source only	44.7
Adversarial (*no ent*)	55.4
Robust adversarial (*no ent*)	62.9
Adversarial (*with ent*)	59.5
Robust adversarial (*with ent*)	**63.9**

UDA results (cont.)

Table 3: Adaptation accuracy on VISDA-17 using Resnet-50 model averaged over 3 runs

	Method	Accuracy (in %)
	Source Only	50.7
	DAN [14]	53.0
	RTN [16]	53.6
	DANN [10]	55.0
	JAN-A [17]	61.6
	GTA [25]	69.5
	SimNet [22]	69.6
	CDAN-E [15]	70.0
Ours	Adversarial (*no ent*)	62.9
Ours	Robust adversarial (*no ent*)	68.6
Ours	Adversarial (*with ent*)	65.5
Ours	Robust adversarial (*with ent*)	**71.5**

UDA results (cont.)

Table 4: Adaptation accuracy on VISDA-17 using Resnet-101 model averaged over 3 runs

	Method	Accuracy (in %)
	Source only	55.3
	DAN [14]	61.1
	DANN [10]	57.4
	MCD [23]	71.9
Ours	Adversarial (*no ent*)	65.5
	Robust adversarial (*no ent*)	69.3
	Adversarial (*with ent*)	69.3
	Robust adversarial (*with ent*)	**72.7**