Axiomatic Explainer Globalness via Optimal Transport

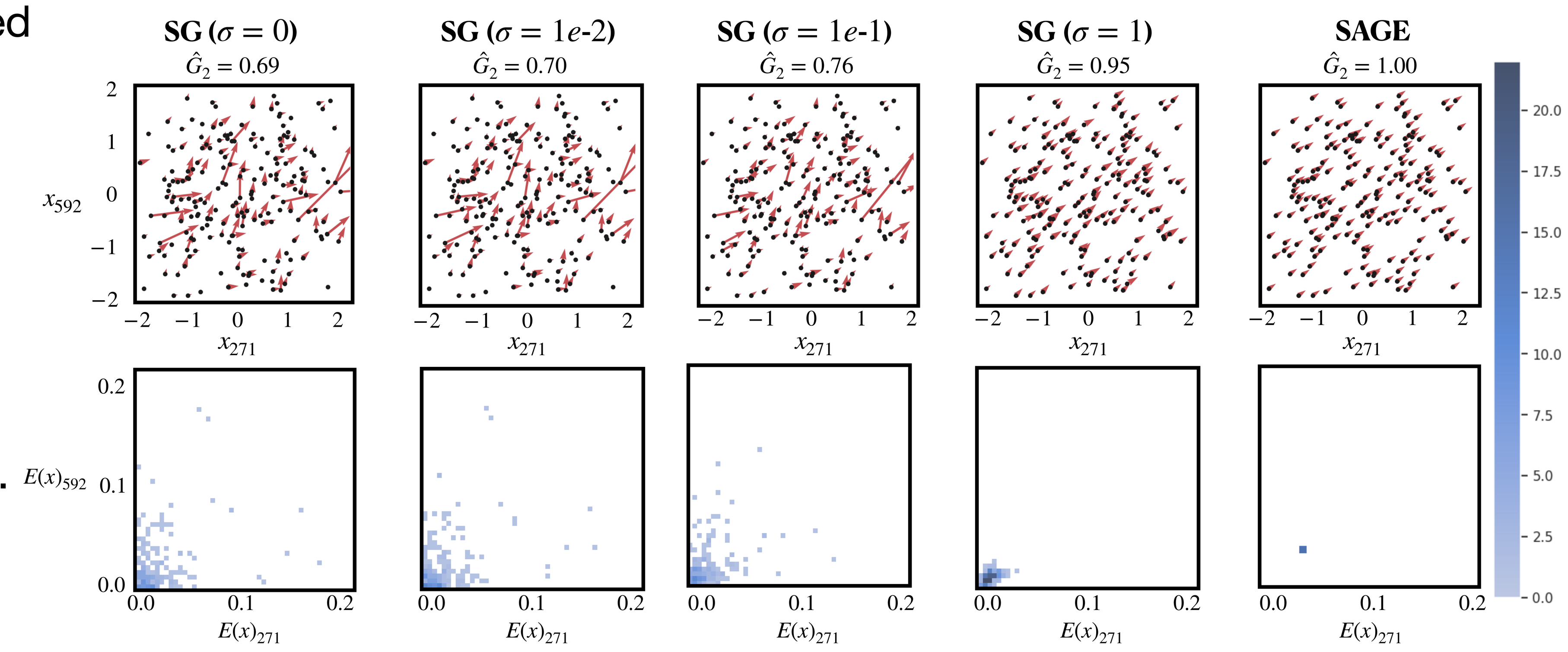
Davin Hill*, Josh Bone*, Aria Masoomi, Max Torop, Jennifer Dy

Department of Electrical and Computer Engineering, Northeastern University, Boston, MA

Northeastern University

Background and Motivation

- Many feature attribution methods have been proposed – explainers are typically compared using *faithfulness* metrics.
- When explainers show similar faithfulness scores, we need additional selection criteria – e.g. *complexity* —just as we prefer simpler models with equivalent accuracy.
- The term “local” is vague, and often used interchangeably with “instance-wise”.
- We propose **Wasserstein Globalness**, a continuous metric that measures explanation diversity, where higher globalness indicates more similar explanations across the dataset.
- Wasserstein Globalness can be applied to feature attribution or selection, and allows choice of explanation similarity metric.



Axiomatic Explainer Globalness

Desired Properties for a Globalness Metric

P1 (Non-negativity) *For any probability measure* $\mu \in \mathcal{P}(\mathcal{E})$, $G(\mu) \geq 0$.
P2 (Continuity) *Let* $\{\mu^{(n)}\}$ *be a sequence of probability measures which converges weakly to* μ. *Then* $G(\mu^{(n)}) \to G(\mu)$.
P3 (Convexity) *Let* $\nu = \lambda P + (1-\lambda)Q$ *for some* $\lambda \in [0,1]$ *and* $P, Q \in \mathcal{P}(\mathcal{E})$. *Then* $G(\nu) \leq \lambda G(P) + (1-\lambda)G(Q)$.

P4 (Fully-local measure). *Let* $U_\mathcal{E} \in \mathcal{P}(\mathcal{E})$ *be a uniform measure with density defined in Eq. 1. Then* $\eta_{\mathcal{E}\#}\mu = U_\mathcal{E} \iff G(\mu) = 0$.
P5 (Fully-global measure) *There exists an* $x_0 \in \mathcal{E}$ *for which* $G_p(\delta_{x_0}) \geq G_p(\mu)$ *for all* $\mu \in \mathcal{P}(\mathcal{E})$.

P6 (Selective Invariance to Transformations) *Let* $T_{\mathcal{E},d_\mathcal{E}}$ *be the group of distance-preserving measurable maps* ϕ.

$$T_{\mathcal{E},d} = \{\phi : \mathcal{E} \to \mathcal{E} \mid d_\mathcal{E}(x,y) = d_\mathcal{E}(\phi(x), \phi(y)) \forall x, y \in \mathcal{E}\}$$

Then G *is invariant with respect to the group* $T_{\mathcal{E},d_\mathcal{E}}$:

$$G(\mu) = G(\phi_{\#}\mu) \text{ for all } \phi \in T_{\mathcal{E},d_\mathcal{E}} \tag{1}$$

Additionally, let S *denote all measurable bijective maps* $\psi : \mathcal{E} \to \mathcal{E}$. *Then* G *is not invariant to* S:

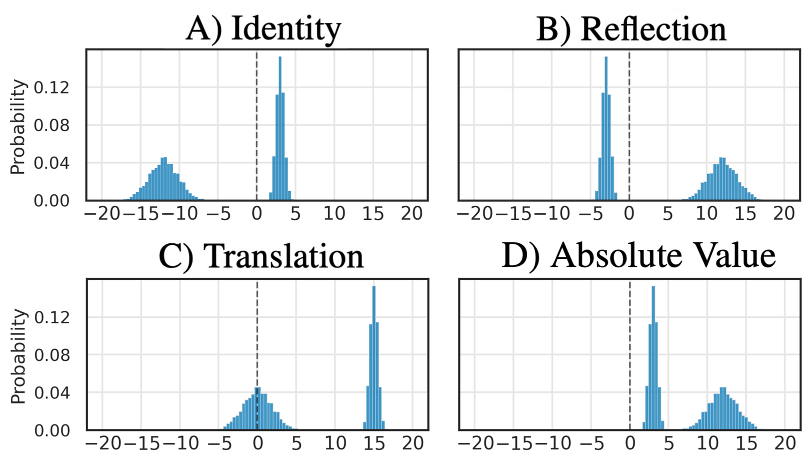
$$\exists \psi \in S \text{ for which } G(\mu) \neq G(\psi_{\#}\mu) \tag{2}$$

Wasserstein Globalness

Definition 1 (Wasserstein Globalness): The Wasserstein Distance between the **Centered Explanation Distribution** $\boldsymbol{\eta_{\epsilon\#}\mu}$ and the **Uniform Distribution** $\boldsymbol{U_\epsilon}$

$$G_p(\mu) := d_W^p(\eta_{\mathcal{E}\#}\mu, U_\mathcal{E})$$

$$= \inf_{\pi \in \Pi(\eta_{\mathcal{E}\#}\mu, U_\mathcal{E})} \left[\mathbb{E}_{(x,y)\sim\pi}[d_\mathcal{E}(x,y)^p]\right]^{1/p}$$



Transformation		$\hat{G}_2$ (Ours)	SE	KL	TV
Isometric					
A) Identity	$t(x) = x$	0.43	2.09	29.4	0.71
B) Reflection	$t(x) = -x$	0.43	2.09	10.0	0.71
C) Translation	$t(x) = x + 12$	0.43	2.09	29.5	0.71
Non-Isometric					
D) Absolute Value	$t(x) = \|x\|$	0.60	2.09	DNE	0.71

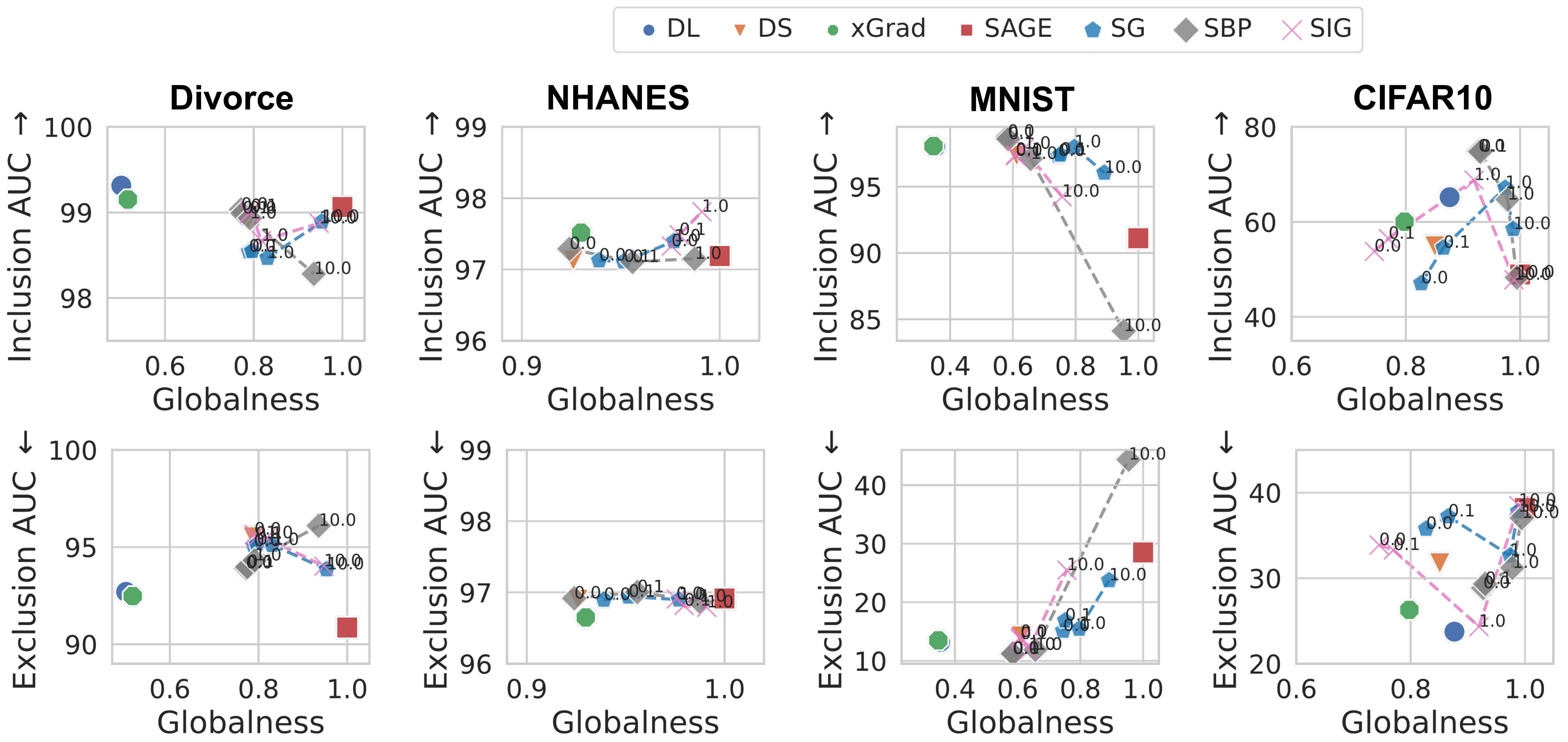
Theorem 1: Wasserstein Globalness satisfies Properties 1-6 for both continuous (feature attribution) and discrete (feature selection) domains.

Theorem 2: Sample Complexity bounds for approximating Wasserstein Globalness

$$\mathbb{E}\left[\left|\hat{G}_p(\mu^{(N)}) - G_p(\mu)\right|\right] \leq \kappa_{p,q} M_q(\eta_{\mathcal{E}\#}\mu)^{1/q} N^{-1/d} + C_{p,q,d} M_q^{p/q}(\eta_{\mathcal{E}\#}\mu)(N^{-p/d} + N^{-(q-p)/q})$$

Experiments

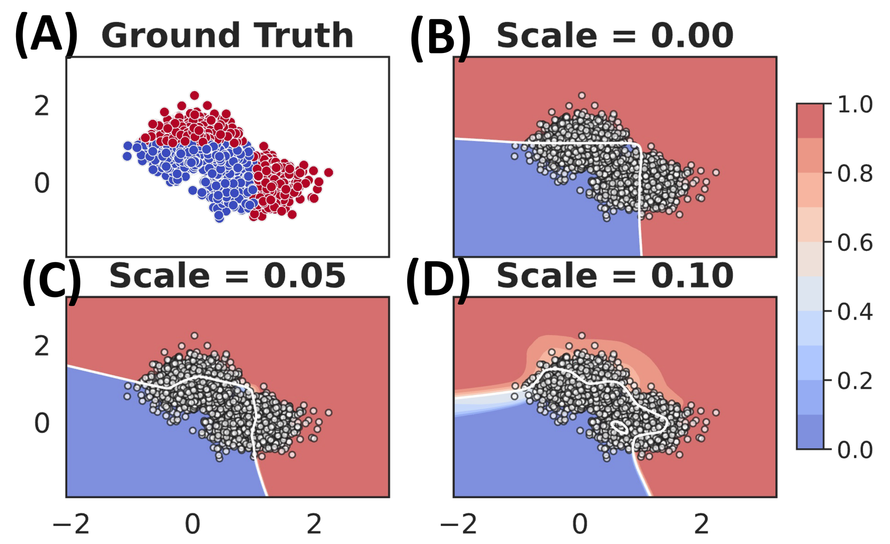
Comparing Explainer Faithfulness and Globalness

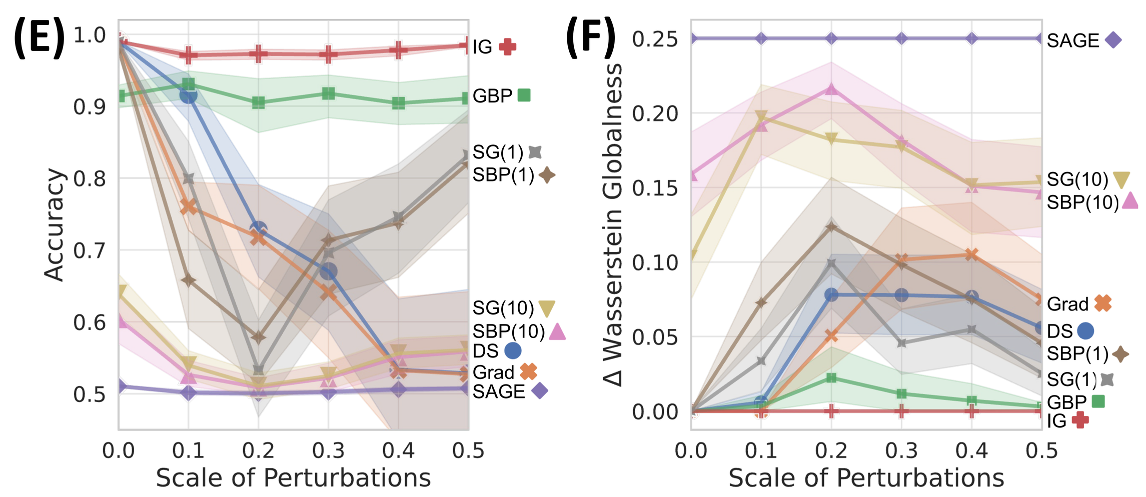


Left: A comparison of explanation *faithfulness* (IncAUC and ExAUC) and *globalness*}. We observe that explainers can often exhibit similar faithfulness with varying levels of globalness, especially for simpler datasets. In this situation, we would prefer high-faithfulness explainers with higher globalness (i.e. lower complexity)

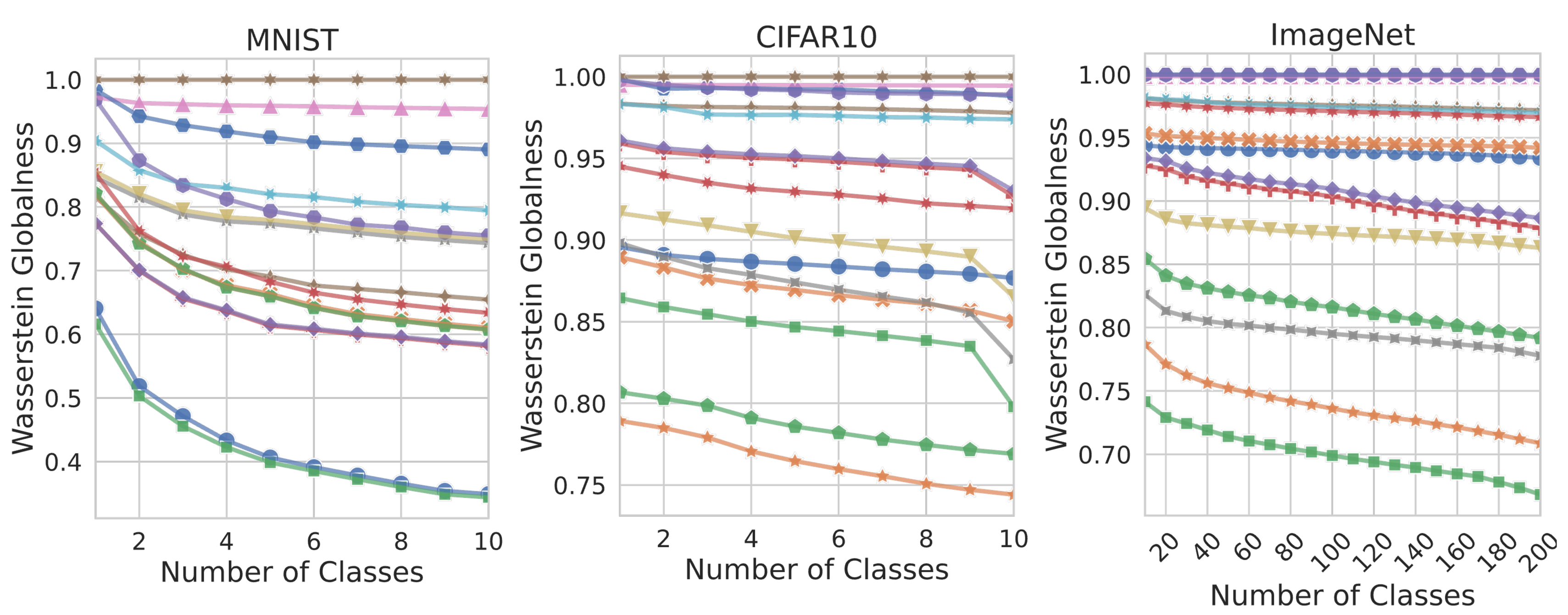
Evaluating “Ground Truth” Globalness on Synthetic Data

Right: (A-D) Synthetic 2-explanation dataset with increasing noise. **(E, F)** We compare the accuracy of each explainer in identifying the relevant features and compare to Globalness.





Evaluating Globalness on Increasing Sample Classes



Left: We evaluate whether WG can capture an increasing diversity of explanations. As we increase the number of label classes in a set of samples, the corresponding explanations become more diverse, resulting in a decreasing WG score

Acknowledgements

Paper!

Code!

This work was supported in part by grants NIH 2T32HL007427-41, U01HL089856, and 5R01HL167072 from the National Heart, Lung, and Blood Institute, and grants NIH R01CA240771 and U24CA264369 from the National Cancer Institute.