

Group counterfactual explanations: State-of-the-art and future research

Explainable AI and causality

(And are they the same?)

Hello there!



Enrique Valero-Leal

- Ongoing ELLIS PhD in AI
 - Technical University of Madrid
 - Ludwig-Maximilians Universität of Munich

Research

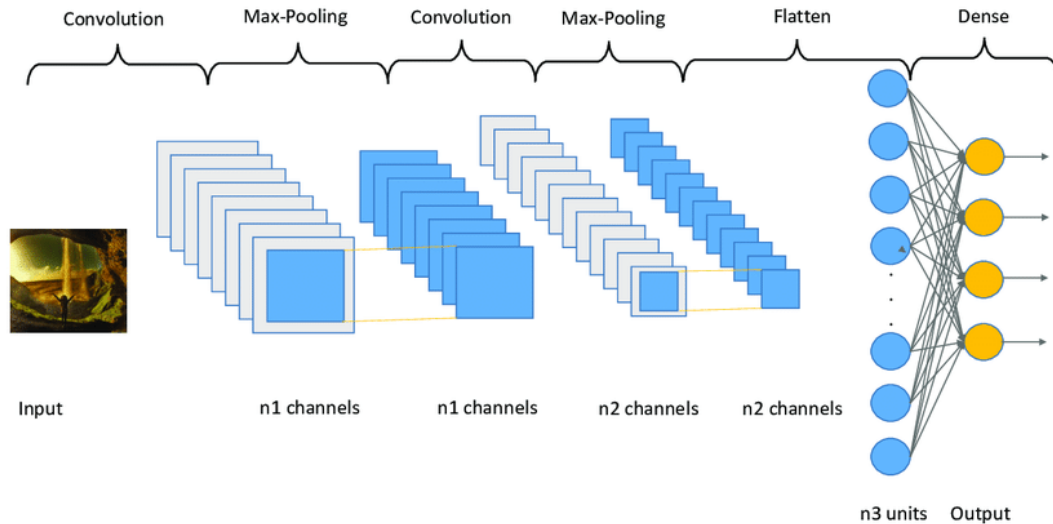
- I am interested in probabilistic models, explainability and causality

Teaching

- Bachelor and advanced courses in machine learning and explainable AI

Deep learning revolution

- Development of powerful machine learning models
- Deployment of those models in our society



A couple of initial claims

- Claim 1: Subjectivity exists
 - And that is ok!
- Claim 2: Artificial intelligence is not intelligent
 - Clever Hans effect. The world of correlations...



Clever Hans

Do maths



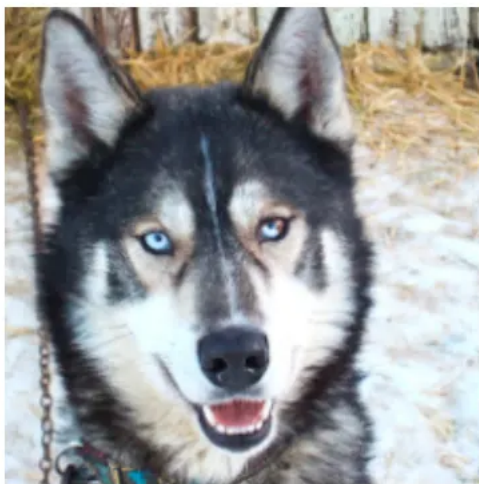
Clever Hans

Carefully analyze
the body language
of the public
to solve a math
problem

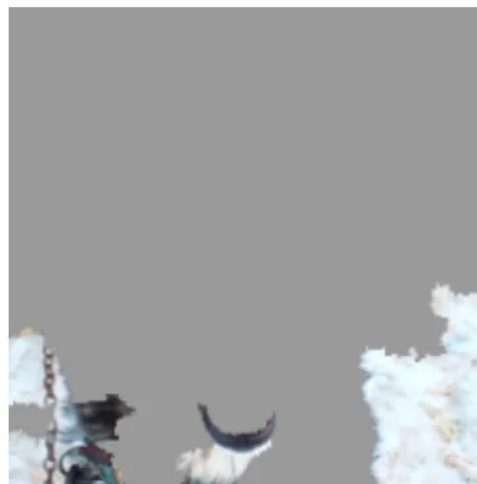
Example in image classification

Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty

Edited by Douglas Massey, Princeton University, Princeton, NJ; received March 6, 2022; accepted August 22, 2022



(a) Husky classified as wolf



(b) Explanation

Solutions?

There are multiple ways to address this problem

- Uncertainty quantification
- Calibration and stability
- ...

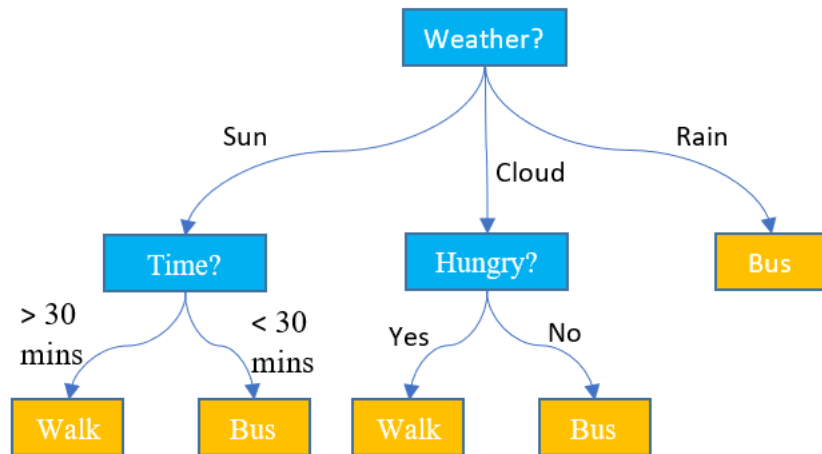
I see an opportunity for:

- Explainable AI: How to interpret a model?
- Causal models: What kind of reasoning can we do with a model?
- Are they somehow related?

Explainable AI

We can explain either a whole model or a specific prediction

- Example: How to get to my friend house?



How does the model generally works?

VS.

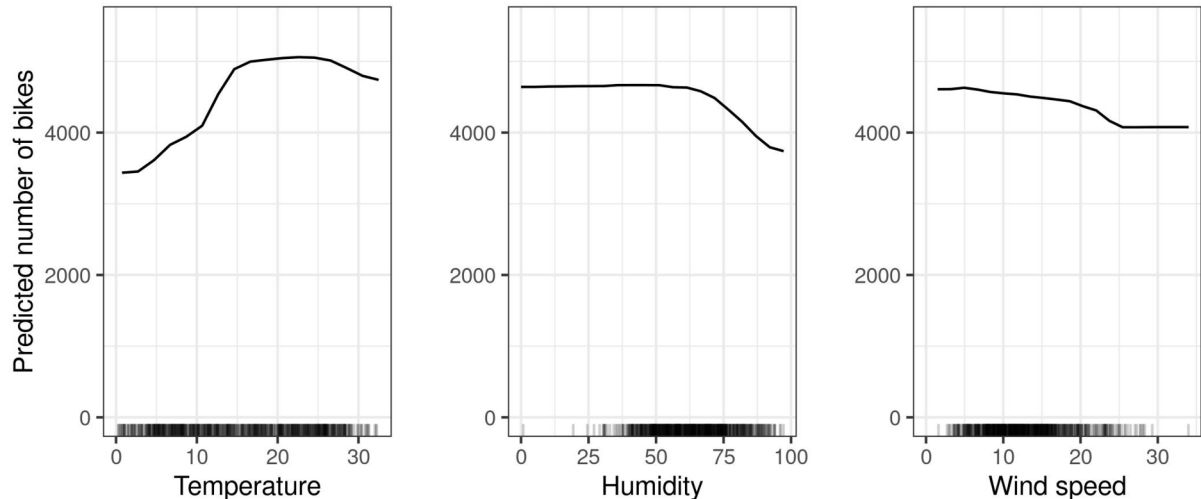
Why did the system recommended me to go walking?

Partial dependence plots

How does changing the variables modify the output?

- I have a bike renting company in Germany. I want to predict the number of bikes I should move from my warehouse to the field.
- Variables: Price, weather, rain, windspeed, ...

$$\hat{f}_S(\mathbf{x}_S) = E_{\mathbf{x}_C} [\hat{f}(\mathbf{x}_S, \mathbf{X}_C)]$$



Counterfactual explanations

Give me a counterexample for an AI prediction.

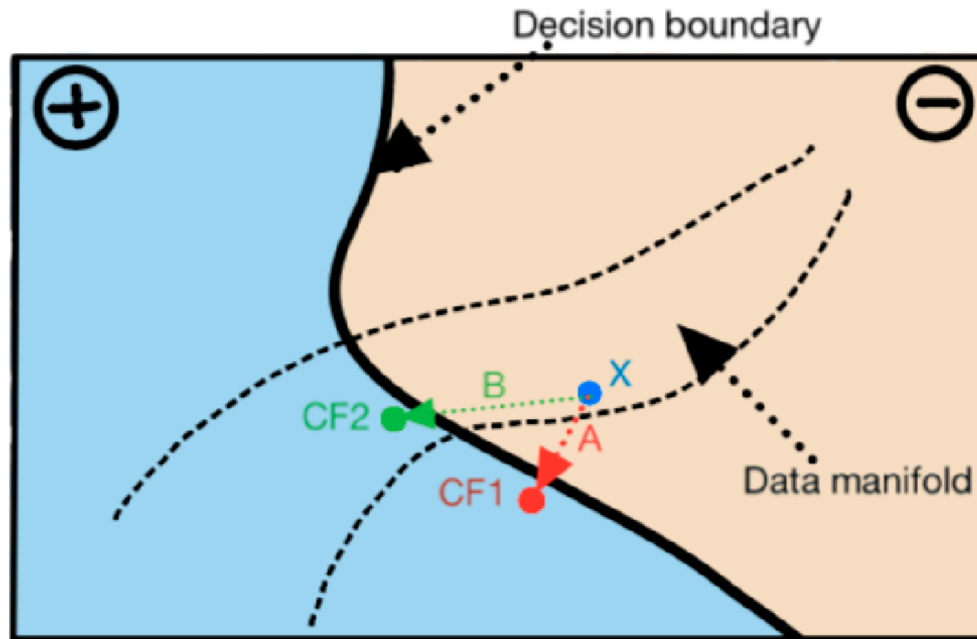
- Example: why should I take 4000 bikes instead of 5000 bikes?
- “Had the wind be milder, you would rent more bikes”.



Counterfactual explanations

It's just an optimisation problems

- Optimize similarity counterfactual s.t. having a different label



Causal models

Causal models explicitly represent causal relations between

- Which is kinda close to how we think!

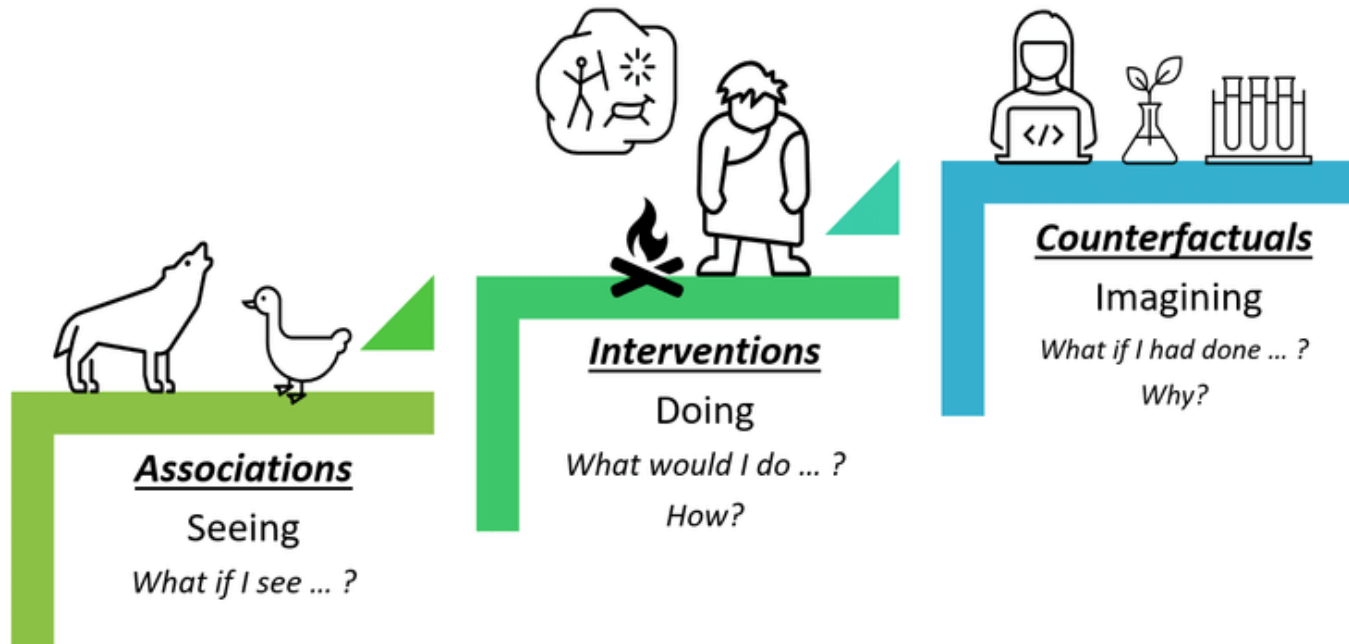


- However, these two models are indistinguishable for a machine

$$P(R, S, G) = P(R) \cdot P(S \mid R) \cdot P(G \mid R, S) = P(S) \cdot P(R \mid S) \cdot P(G \mid R, S)$$

Ladder of causality

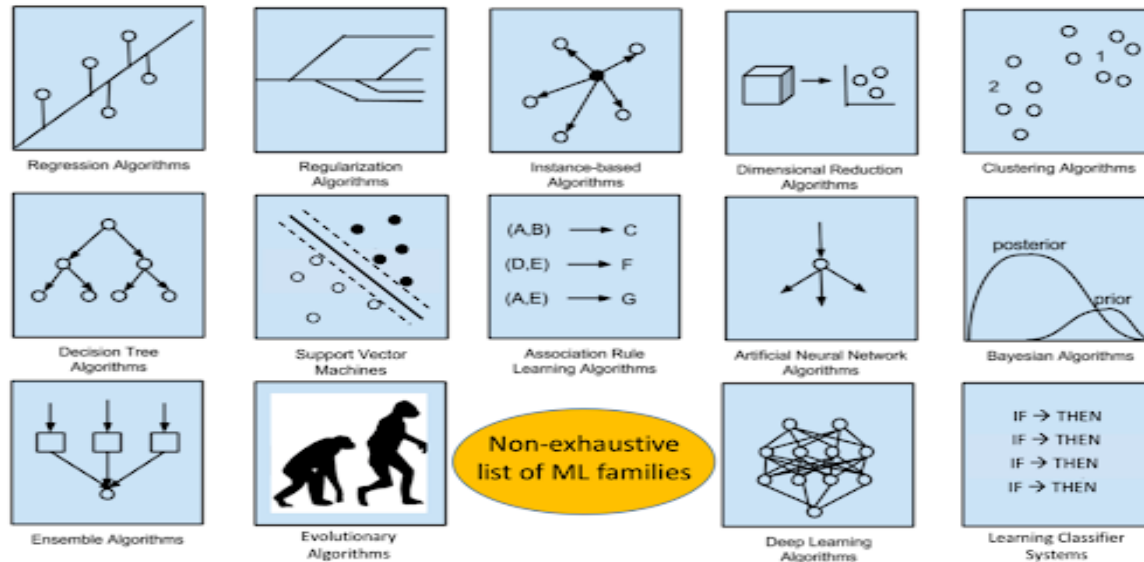
Classification of causal model attending how it can reason



Level 1: Associations

Models can answer “present” questions

- What if I see this?
- What is the number of bikes if I observe certain weather conditions?



Level 1: Associations (II)

 Microsoft +  OpenAI

 Google +  DeepMind

 TESLA +  xAI

 Meta +  Meta AI

*“Human-Level Intelligence
or Animal-Like Abilities?” -
Adnan Darwiche*

Level 2: Interventions

Models can answer “future” questions

- What will happen if I do this now?
- What will happen if I increase the price of the bikes by two?



Level 2: Interventions (II)

Observational



- What happens in other companies where the bike price is double?
- Correlation is not causation!

Interventional

- What will happen in MY company?
- Intuition of variable importance

Level 2: Interventions (III)

Causal discovery: How do we create this model?

- Repeated randomized controlled trials
- That yield an interventional distributions on top of the observational



The functional world

Is a coin toss deterministic or probabilistic?



Level 3: Counterfactuals

Models can answer “hypothetical present” questions

- What would have happened if something went different?
- Yesterday it was sunny and prepared 5000 bikes. How many should I have prepared if it was cloudy instead



Level 3: Counterfactuals (II)

Counterfactuals are a mathematical formalization of imagination

- How would it be like if in today's world there were dragons?

HOWEVER, the proper causal functions are needed

- Otherwise, counterfactuals are “non-falsable”

But now, what?

Recently, it was studied that some XAI methods are “practical implementations” of causality concepts

- Partial dependence = Interventions
- Counterfactual explanations = Counterfactuals in causal models

Both can help to understand the model and avoid model degeneration over time

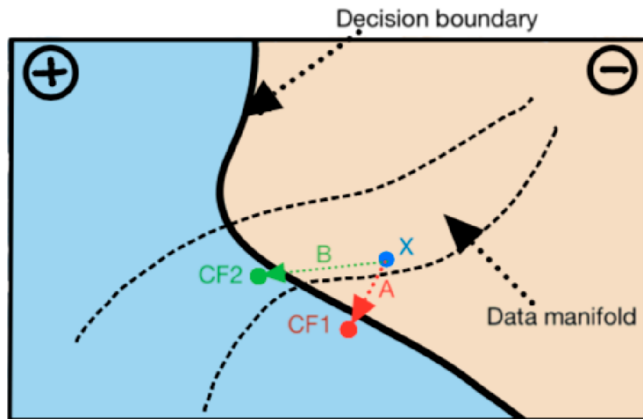
Maybe is the time to move from deep learning to deep understanding –
Judea Pearl

Optimal Transport Group Counterfactual Explanations

The need for group counterfactuals

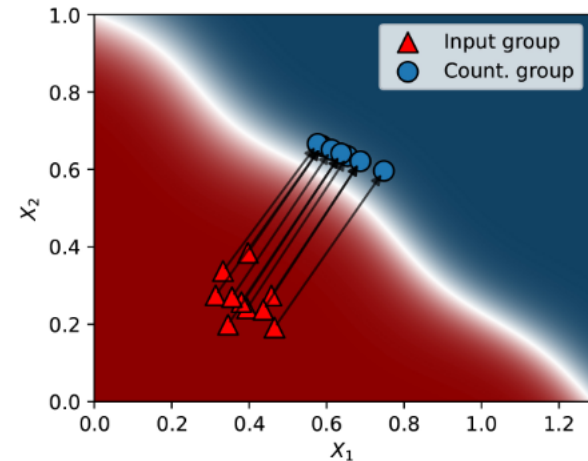
Individual explanations

- Standard counterfactuals explain how changes to input features alter a prediction for a single instance



Group explanations

- Group counterfactuals provide a set of instance to explain a group contrastively, potentially aiming for a consistent recourse



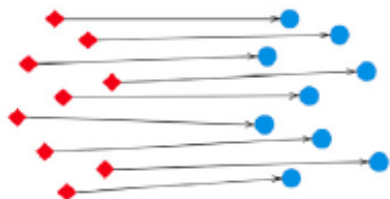
Interesting scenarios

- In business with a limited stock, we might need to redirect buyers to alternative products. If all are redirected to the same product, we create congestion
- In group trials, we might be interested in finding a common and cohesive compensation for the victims

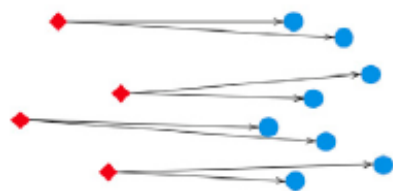
The proposal by Carrizosa et al. (2024)

- Different scenarios for group counterfactuals
 - Depending on **input size S** and **output size R**
 - We focus on the first scheme. For a group $\underline{x}$ find a counterfactual $\underline{x}'$

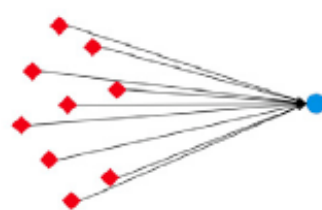
$$\min_{\underline{x}'} \sum_{i=1}^n \|\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}\|_2^2 \quad \text{Potentially s.t. } \|\mathbf{x}'^{(i)} - \mathbf{x}'^{(j)}\|_2 \leq K \cdot \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_2, \quad \forall i \neq j,$$



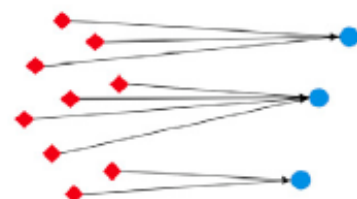
(a) One-for-one



(b) Many-for-one



(c) One-for-all



(d) One-for-many

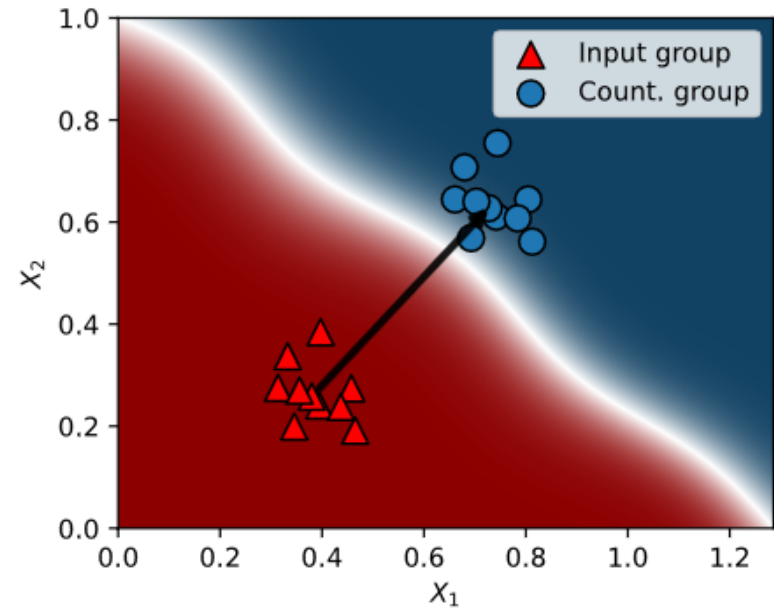
Limitations of current approaches

- Violation of Feasibility (Congestion): Current methods optimize instances individually, directing everyone to the single nearest target point, causing severe congestion in recourse.
- Violation of Fairness: Tailors compensation strictly to individual data noise, resulting in heterogeneous and potentially unfair decisions.
- Lack of Generalization: Existing algorithms must completely re-run expensive optimization processes for any new group member.
- Computational Intractability: The problem intractable for non-convex or non-linear classifiers.

Our ideal method

- Preserves geometry in the counterfactual group
- Generalizes to new instances
- The problem remains tractable and potentially convex

How do we do this?



Optimal transport

- You are in the Munich bothanical garden
- You see 3 beautiful butterflies



Optimal transport

- You look away and now they have moved!
- Which butterfly is each?
- Or... how did they move?



Optimal transport

- Optimal transport tries to answer this.
- What is the most likely moving pattern?

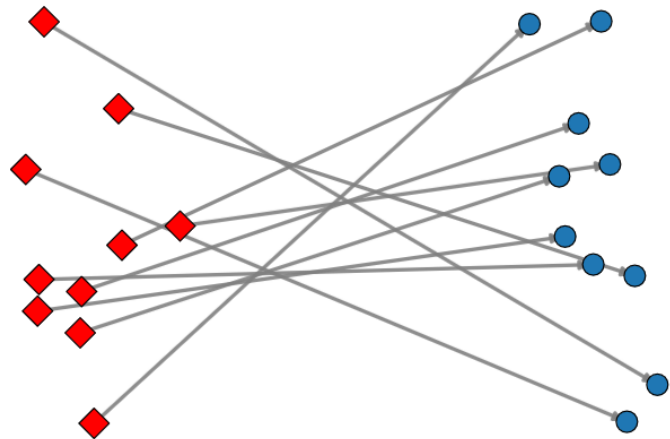


Theory time: Optimal transport

Optimal transport studies how to find the best mapping between two sets of two metric spaces $\mathcal{X}$, $\mathcal{Y}$ with probability measures as p and q .

- We have cost function c and a push-over function T .
- Monge formulation (deterministic):

$$\inf_{T: T_{\#}p=q} \int_{\mathcal{X}} c(\mathbf{x}, T(\mathbf{x})) dp(\mathbf{x})$$



Theory time: Optimal transport (probabilistic)

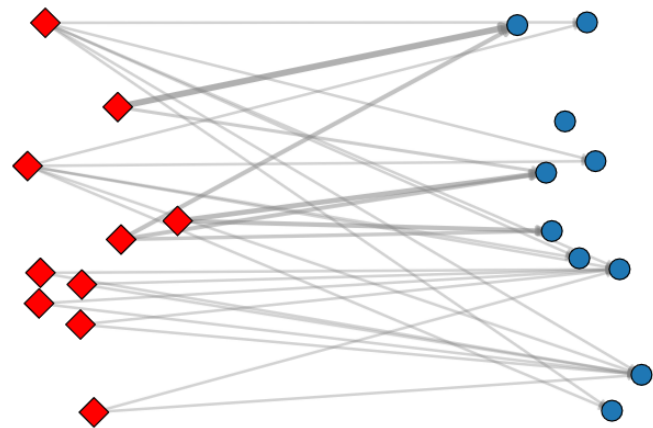
Optimal transport studies how to find the best mapping between two sets of two metric spaces $\mathcal{X}$, $\mathcal{Y}$ with probability measures as p and q .

- We can also aim to “move” the distribution by establishing a coupling π
- Kantorovich formulation:

$$\inf_{\pi \in \Pi(p, q)} \int_{\mathcal{X} \times \mathcal{Y}} c(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y})$$

This defines the Wasserstein distance:

$$W_2(p, q) = \inf_{\pi \in \Pi(p, q)} \sqrt{\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \pi} \|\mathbf{x} - \mathbf{y}\|_2^2}$$



Proposal: OT Group Counterfactuals

Let's find a function that maps every point in the input space to a counterfactual space.

$$\underline{\mathbf{x}}' = \hat{g}(\mathbf{x}) \text{ instead of finding } x'^{(i)} \forall x^{(i)}.$$

Main idea:

- Much easier to control its properties and to interpret it.
- Since it is a function, we can generalize.
- The function should have some properties to preserve geometry.

How to define the OT function?

To find the function $\hat{g} : \mathcal{X} \rightarrow \mathcal{X}$, optimize:

$$\hat{g} = \arg \min_{g \in \mathcal{G}} \frac{1}{|\underline{\mathbf{x}}|} \sum_{\mathbf{x} \in \underline{\mathbf{x}}} \|\mathbf{x} - g(\mathbf{x})\|_2$$

We still lack a way to control the density preservation

- Bilipschitz constraint:
$$\frac{1}{K} \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \leq \|\hat{g}(\mathbf{x}_1) - \hat{g}(\mathbf{x}_2)\|_2^2 \leq K \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \underline{\mathbf{x}}$$
- We reduce parameters and added generalization, but still non-convex...

Affine transforms as OT maps

Sensible to assume that affine transforms are enough:

- We look for simplicity in the transport
- Groups are typically homogenous without outliers.

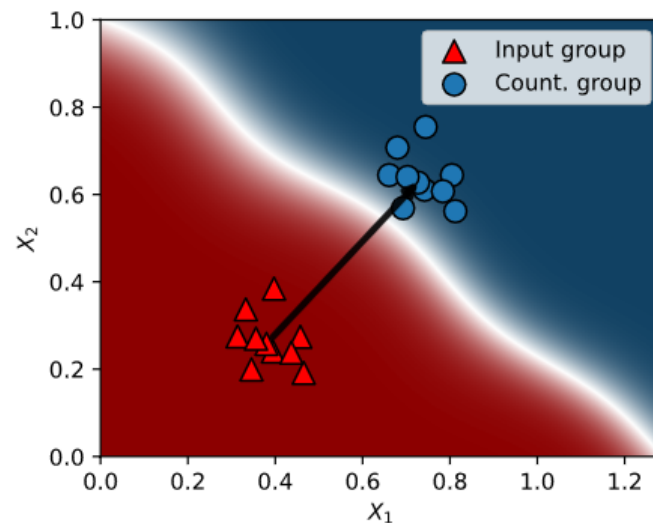
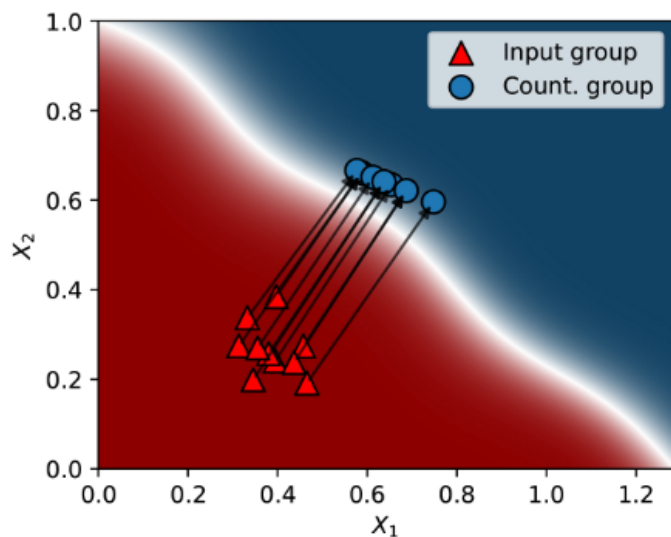
The result?

- Checking Bilipschitz continuity reduces to bound the Eigenvalues of the transform matrix A between 0 and K , which is a convex constraint
- We go from complexity n^2d to d

Affine transforms as OT maps

Example of geometry preservation control.

- To the left, big K (low enforcement)
- To the right, $K=1$ (maximum enforcement)



Another bottleneck

Let's look again to the objective

$$\hat{g} = \arg \min_{g \in \mathcal{G}} \frac{1}{|\underline{\mathbf{x}}|} \sum_{\mathbf{x} \in \underline{\mathbf{x}}} \|\mathbf{x} - g(\mathbf{x})\|_2$$

The constraint is does not scale with the group size, but the objective still does... Can we do better?

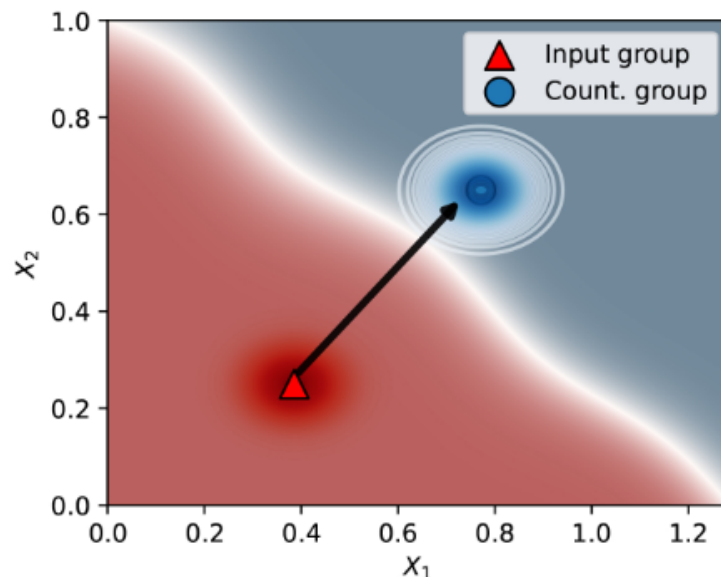
Gaussian OT maps

We can further reduce the complexity if we assume Gaussianity.

The Wasserstein distance between two Gaussian has a closed form.

$$W_2^2(p, q) = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\|_2^2 + \text{tr} \left(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q - 2 \left(\boldsymbol{\Sigma}_q^{\frac{1}{2}} \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_q^{\frac{1}{2}} \right)^{\frac{1}{2}} \right).$$

- i.e., we need to find $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the target distribution q



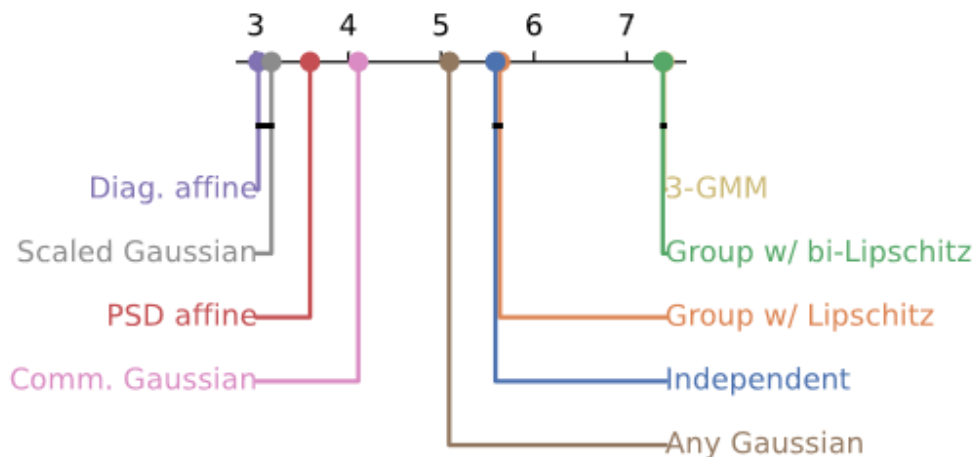
Theoretical summary

Method	Convex	Math opt.	Closed-form	Params
Group w/ Bi-Lipschitz		QCQP		nd
Affine transform	✓	SDP		$\frac{d(d+1)}{2} + d$
Diagonal affine transform	✓	QP		$2d$
Scaled Gaussian	✓	SDP	✓	$d(d+1) + d$
Gaussian commutative	✓	QCQP	✓	$2d$
Gaussian scaled	✓	QP	✓	$d+1$
k -GMM	✓	SDP	✓	$k(d(d+1) + d)$

Results (linear classifiers)

Efficiency (W2 w.r.t to the #params). Convex optimization.

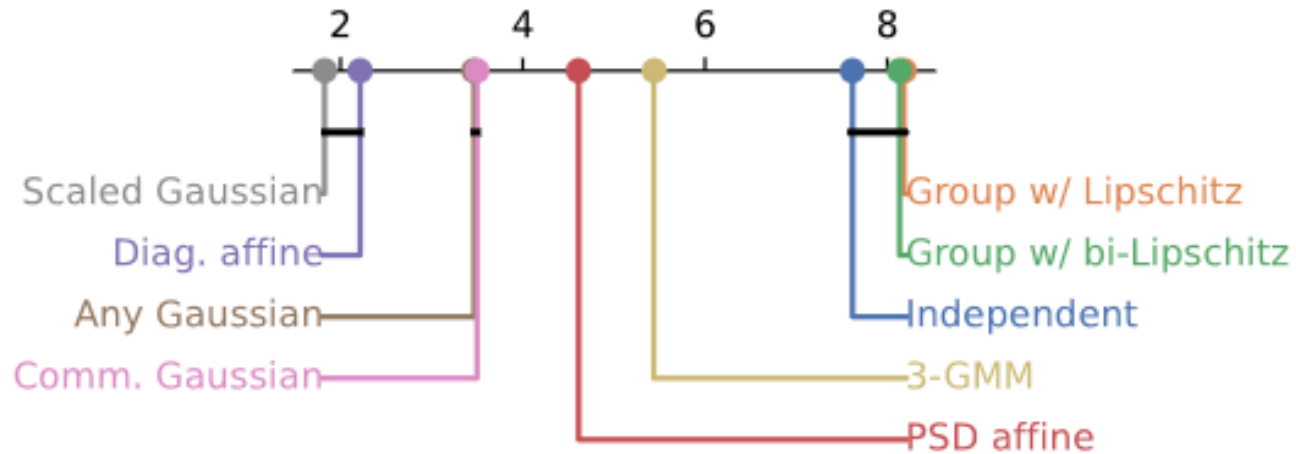
- Our sparse proposals perform best
- Speed up of up to x1000
- Proposals still competitive even if we ignore the reduction in #params



Results (non-linear classifiers)

Hypervolume (geometry preservation, W2). Heuristic MO optimization.

- Our proposals perform even better relatively
- The baselines do not converge due to the great number of #params



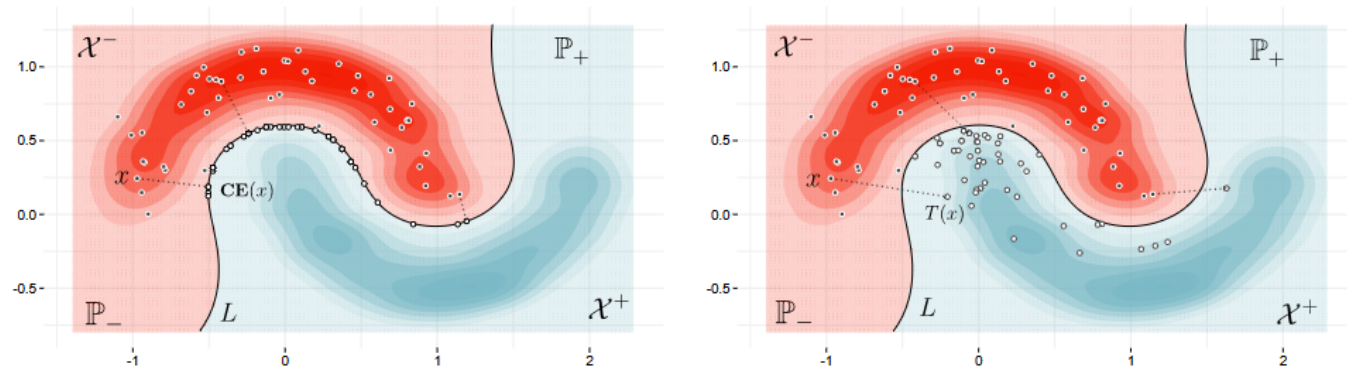
Problems that are still open

Taking into account data distributions

For simplicity, this and previous work ignore data distributions

Challenges

- Define a congestion metric that is guided by the density
- Perhaps quadratic data density estimators can be helpful (splines)



Extension to image data

Take a group of pictures that are similar, but wrongly classified

Challenges

- Defining a kernel function between images that make human sense.
 - Similarity = Similarity in classification scores
 - Relation with mechanistic interpretability



Mechanistic interpretability

Reverse engineer a neural network

Sparse dictionary learning

- The same neuron can represent different concepts
- Learn the concepts as linear combinations of neuron activations.

