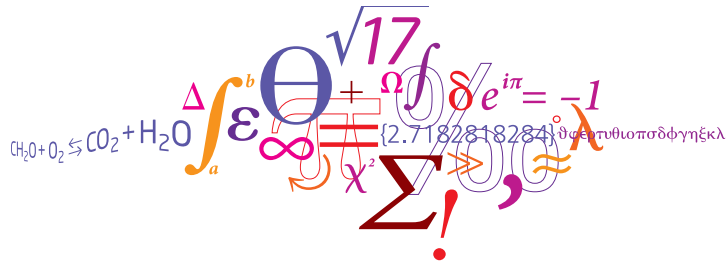
Distributional Counterfactual Explanations With Optimal Transport

Lei You, Ph.D.

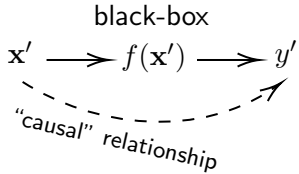
Assistant Professor in Applied Mathematics

Department of Engineering Technology

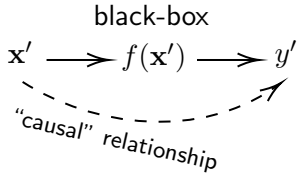
Technical University of Denmark (DTU)



Counterfactual Explanations



Counterfactual Explanations



Factual

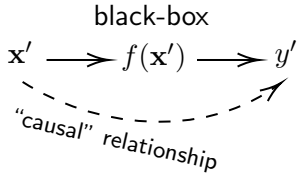
$\mathbf{x}'$ Reality $\longrightarrow y'$

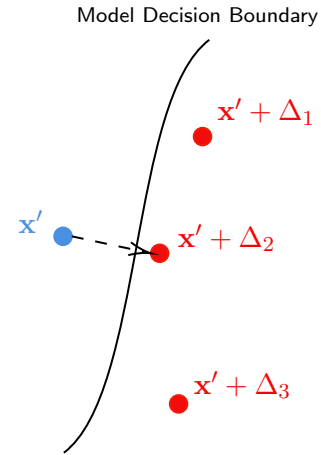
Counterfactual

$\mathbf{x}' + \Delta \longrightarrow f(\mathbf{x}' + \Delta)$

Hypothetical Reality

Counterfactual Explanations





Factual

$\mathbf{x}'$ ⟶ y'

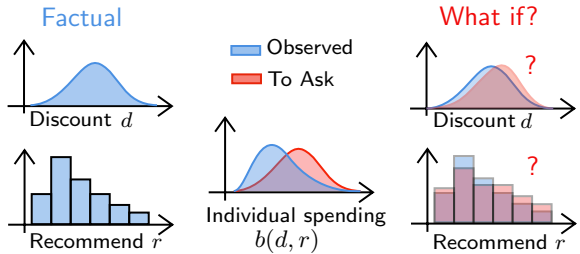
Reality

Counterfactual

$\mathbf{x}' + \Delta$ ⟶ $f(\mathbf{x}' + \Delta)$

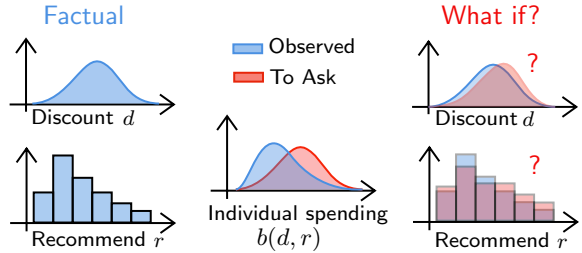
Hypothetical Reality

Distributional Counterfactual Explanations (DCE)



Consider a business that uses an ML model b to forecast revenue. Given any observed model outputs, the business wants to know how shifting the overall distributions of d and r would change predictions towards another pattern.

Distributional Counterfactual Explanations (DCE)



To answer this **"what-if"** question, the counterfactual d and r should **resemble** the observed factual d and r for practical and actionable changes.

Preliminaries of Optimal Transport (OT)

OT and the Wasserstein Distance. The one-dimensional (1D) squared 2-Wasserstein distance is defined as

$$\mathcal{W}^2(\gamma_1, \gamma_2) \triangleq \inf_{\pi \in \Pi} \int_{\mathbb{R}\times\mathbb{R}} \|a_1 - a_2\|^2 \, \mathrm{d}\pi(a_1, a_2),$$

which represents the optimal transport (OT) cost between γ_1 and γ_2 with respect to the squared Euclidean distance $\|a_1 - a_2\|^2$ under the optimized OT plan π.

Preliminaries of Optimal Transport (OT)

Sliced Wasserstein Distance. Let $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ stand for the d-dimensional unit sphere:

$$\mathcal{SW}^2(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \triangleq \int_{\mathbb{S}^{d-1}} \mathcal{W}^2(\boldsymbol{\theta}\sharp\boldsymbol{\gamma}_1, \boldsymbol{\theta}\sharp\boldsymbol{\gamma}_2)\, \mathrm{d}\sigma(\boldsymbol{\theta}),$$

where σ is the uniform distribution on $\mathbb{S}^{d-1}$ and $\sharp$ is the push-forward operator, projecting (high-dimensional) measures $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ onto 1D.

DCE Formulation

Denote b a black-box model. Given any data distribution $\boldsymbol{x}'$ as a factual (i.e. an observation), we aim to find a counterfactual $\boldsymbol{x}$ that achieves a desired target outcome y^*, which is obtained by solving the following optimization problem.

$$\max_{\boldsymbol{x}, P} P \tag{1a}$$

$$\text{s.t.} \quad P \leq \mathbb{P}\left[\underbrace{\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') < U_x}_{\text{Counterfactual proximity}}\right] \tag{1b}$$

$$P \leq \mathbb{P}\left[\underbrace{\mathcal{W}^2(b(\boldsymbol{x}), y^*) < U_y}_{\text{Counterfactual effect}}\right] \tag{1c}$$

$$\underbrace{P \geq 1 - \frac{\alpha}{2}}_{\text{We need a sufficiently large } P} \tag{1d}$$

Problem Solving Framework

1: **repeat**
2: Solve OT to obtain $\mathcal{W}^2(b(\boldsymbol{x}), y^*)$ and $\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}')$.
3: Stochastic gradient descent (SGD) for

$$\min_{\boldsymbol{x}} \ (1-\eta) \cdot \mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') + \eta \cdot \mathcal{W}^2(b(\boldsymbol{x}), y^*)$$

4: Estimate Upper Confidence Limit (UCL) $\overline{\mathcal{W}^2}(b(\boldsymbol{x}), y^*)$ and $\overline{\mathcal{SW}^2}(\boldsymbol{x}, \boldsymbol{x}')$ given α
5: Adjust η to maximize the smaller gap of $U_x - \overline{\mathcal{SW}^2}$ and $U_y - \overline{\mathcal{W}^2}$
6: **until** Counterfactual $\boldsymbol{x}$ converges

Problem Solving Framework

$$\max_{\boldsymbol{x}, P} P \tag{2a}$$

$$\text{s.t.} \quad P \leq \mathbb{P}\left[\underbrace{\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') < U_x}_{\text{Counterfactual proximity}}\right] \tag{2b}$$

$$P \leq \mathbb{P}\left[\underbrace{\mathcal{W}^2(b(\boldsymbol{x}), y^*) < U_y}_{\text{Counterfactual effect}}\right] \tag{2c}$$

$$\underbrace{P \geq 1 - \frac{\alpha}{2}}_{\text{We need a sufficiently large } P} \tag{2d}$$

Theoretical Foundation. The two chance constraints (2b) and (2c) holds with probability $1 - \alpha/2$ iff $\overline{\mathcal{SW}^2} \leq U_x$ and $\overline{\mathcal{W}^2} \leq U_y$, respectively.

Problem Solving Framework

1: **repeat**
2: Solve OT to obtain $\mathcal{W}^2(b(\boldsymbol{x}), y^*)$ and $\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}')$.
3: SGD for $\min_{\boldsymbol{x}} (1-\eta) \cdot \mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') + \eta \cdot \mathcal{W}^2(b(\boldsymbol{x}), y^*)$
4: **Estimate UCL $\overline{\mathcal{W}^2}(b(\boldsymbol{x}), y^*)$ and $\overline{\mathcal{SW}^2}(\boldsymbol{x}, \boldsymbol{x}')$ given α**
5: Adjust η to maximize the smaller gap of $U_x - \overline{\mathcal{SW}^2}$ and $U_y - \overline{\mathcal{W}^2}$
6: **until** Counterfactual $\boldsymbol{x}$ converges

Problem Solving Framework

1: **repeat**
2: Solve OT to obtain $\mathcal{W}^2(b(\boldsymbol{x}), y^*)$ and $\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}')$.
3: SGD for
$$\min_{\boldsymbol{x}} \ (1-\eta) \cdot \mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') + \eta \cdot \mathcal{W}^2(b(\boldsymbol{x}), y^*)$$
4: **Estimate UCL $\overline{\mathcal{W}^2}(b(\boldsymbol{x}), y^*)$ and $\overline{\mathcal{SW}^2}(\boldsymbol{x}, \boldsymbol{x}')$ given α**
5: Adjust η to maximize the smaller gap of $U_x - \overline{\mathcal{SW}^2}$ and $U_y - \overline{\mathcal{W}^2}$
6: **until** Counterfactual $\boldsymbol{x}$ converges

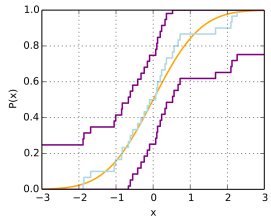


Figure: Dvoretzky–Kiefer–Wolfowitz inequality (DKW inequality)

Problem Solving Framework

1: **repeat**
2: Solve OT to obtain $\mathcal{W}^2(b(\boldsymbol{x}), y^*)$ and $\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}')$.
3: SGD for $\min_{\boldsymbol{x}} (1-\eta) \cdot \mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') + \eta \cdot \mathcal{W}^2(b(\boldsymbol{x}), y^*)$
4: **Estimate UCL $\overline{\mathcal{W}^2}(b(\boldsymbol{x}), y^*)$ and $\overline{\mathcal{SW}^2}(\boldsymbol{x}, \boldsymbol{x}')$ given α**
5: Adjust η to maximize the smaller gap of $U_x - \overline{\mathcal{SW}^2}$ and $U_y - \overline{\mathcal{W}^2}$
6: **until** Counterfactual $\boldsymbol{x}$ converges

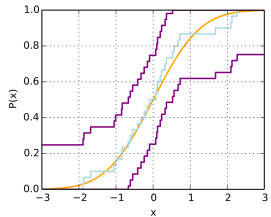


Figure: Dvoretzky–Kiefer–Wolfowitz inequality (DKW inequality)

1D Case: $\mathcal{W}^2(z, z) = \int_0^1 \underbrace{\left\| F_z^{-1}(q) - F_{z'}^{-1}(q) \right\|^2}_{\textit{Confidence bound} \text{ obtained by DKW}} \mathrm{d}q$

Problem Solving Framework

1: **repeat**
2: Solve OT to obtain $\mathcal{W}^2(b(\boldsymbol{x}), y^*)$ and $\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}')$.
3: SGD for $\min_{\boldsymbol{x}} (1-\eta) \cdot \mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') + \eta \cdot \mathcal{W}^2(b(\boldsymbol{x}), y^*)$
4: **Estimate UCL $\overline{\mathcal{W}^2}(b(\boldsymbol{x}), y^*)$ and $\overline{\mathcal{SW}^2}(\boldsymbol{x}, \boldsymbol{x}')$ given α**
5: Adjust η to maximize the smaller gap of $U_x - \overline{\mathcal{SW}^2}$ and $U_y - \overline{\mathcal{W}^2}$
6: **until** Counterfactual $\boldsymbol{x}$ converges

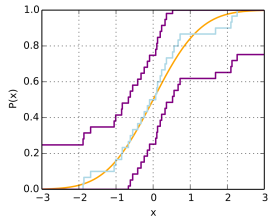


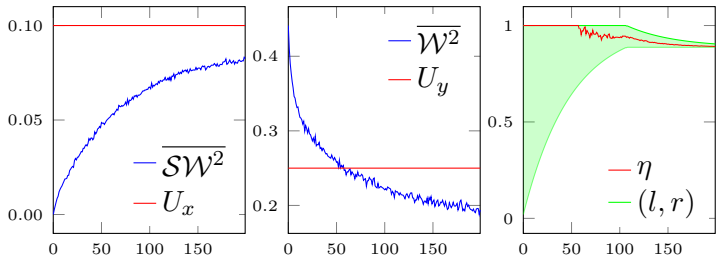
Figure: Dvoretzky–Kiefer–Wolfowitz inequality (DKW inequality)

1D Case: $\mathcal{W}^2(z, z) = \int_0^1 \underbrace{\left\| F_z^{-1}(q) - F_{z'}^{-1}(q) \right\|^2}_{\textit{Confidence bound}\text{ obtained by DKW}} \mathrm{d}q$

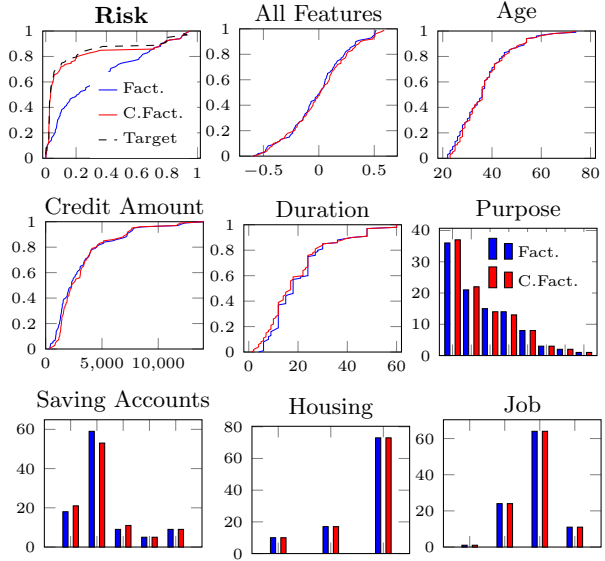
Hence: $\mathcal{W}^2(z, z) \leq \underbrace{\int_0^1 \textit{Confidence bound}\ \mathrm{d}q}_{\textbf{UCL of } \mathcal{W}^2(z, z')\textbf{, i.e. } \overline{\mathcal{W}^2}(z, z')}$

Convergence Analysis

1: **repeat**
2: Solve OT to obtain $\mathcal{W}^2(b(\boldsymbol{x}), y^*)$ and $\mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}')$.
3: SGD for $\min_{\boldsymbol{x}} (1-\eta) \cdot \mathcal{SW}^2(\boldsymbol{x}, \boldsymbol{x}') + \eta \cdot \mathcal{W}^2(b(\boldsymbol{x}), y^*)$
4: Estimate UCL $\overline{\mathcal{W}^2}(b(\boldsymbol{x}), y^*)$ and $\overline{\mathcal{SW}^2}(\boldsymbol{x}, \boldsymbol{x}')$ given α
5: **Adjust η to maximize the smaller gap of $U_x - \overline{\mathcal{SW}^2}$ and $U_y - \overline{\mathcal{W}^2}$**
6: **until** Counterfactual $\boldsymbol{x}$ converges



Counterfactual Proximity vs. Counterfactual Effect



Benchmarking

Model	Algo.	Cover.	Cost						Proximity		Time (s)
			AReS Cost	% Difference at Percentiles					OT	MMD	
				0-15	15-30	30-70	70-85	85-100			
DNN (74.9%)	AReS	0.019	0.038	0.000	0.187	43.83	0.321	0.000	0.000	0.000	12
	Globe	0.962	**3.346**	101.6	45.08	515.8	95.56	230.5	8.521	0.039	**5.3**
	DiCE	0.796	18.41	100.0	100.0	326.0	12.04	30.68	0.324	0.107	7235
	Discount	**0.981**	22.87	**7.801**	**8.183**	**265.0**	**9.422**	**6.325**	**0.202**	**0.036**	632
RBFNet (73.6%)	AReS	0.058	0.049	0.000	1.002	6.000	0.418	0.000	0.002	0.001	11
	Globe	0.038	3.302	41.93	28.28	87.70	57.29	117.8	2.899	0.039	4.9
	DiCE	0.174	10.54	100.0	100.0	95.28	4.257	32.39	0.330	0.148	2735
	Discount	**0.962**	**16.22**	**8.461**	**21.65**	**185.5**	**36.89**	**19.55**	**0.563**	**0.039**	**621**
SVM (75.0%)	AReS	0.038	0.577	0.000	0.000	0.794	0.935	0.000	0.001	0.000	14
	Globe	**1.000**	**3.600**	155.1	95.49	**152.2**	34.98	178.6	13.72	0.039	**4.8**
	DiCE	**1.000**	13.01	100.0	100.0	636.7	44.13	30.36	0.514	0.130	5767
	Discount	**1.000**	4.357	**5.591**	**14.07**	256.6	**16.63**	**9.811**	**0.342**	**0.036**	244

Open Problem: Information in the Optimal Transport Plan

It is worth exploring the connection between distribution shift research and the search for counterfactual explanations.