Collective Counterfactual Explanations:

Balancing Individual Goals and Collective Dynamics

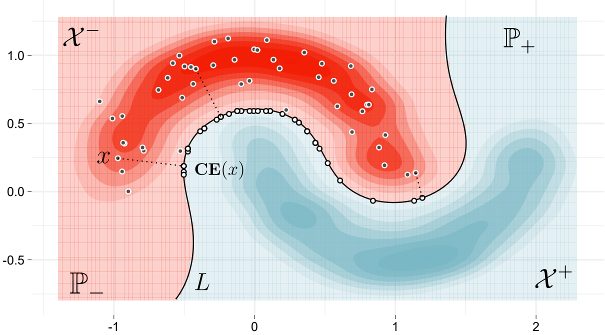
Ahmad-Reza Ehyaei[1], Ali Shirali [2] and Samira Samadi[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen AI Center
[2]University of California, Berkeley, USA

December, 2025

Why Individual Counterfactual Explanations (CE) Fails?

- **Externalities & Competition**: When many people follow similar recourse, they crowd into the same **desirable region**, creating competition and lowering everyone's utility (tragedy of the commons).
- **Ignores the Population Law**: CE that chases the closest decision boundary neglects the feature-space distribution $\mathbb{P}$, often funneling people into low-resource or unusual regions—recommendations users may perceive as impractical.



Collective Counterfactual Explanations (CCE): The Main Idea

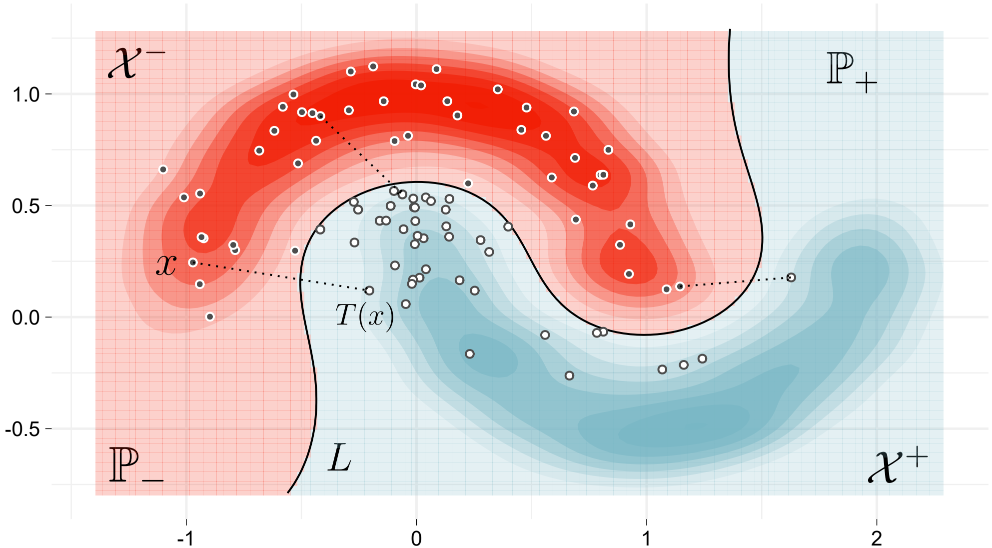
CCE gives collective recourse by balancing individual effort with the population impact after everyone (partially) acts. The key ingredients are:

- **Population dynamics lens**: Model competition via equilibrium: resources ($S(x)$) and density ($U(x,t)$) imply at equilibrium ($U^* \propto S$). Recourse should move the population to a new near-equilibrium, avoiding overcrowding.
- **Objective = effort + equilibrium penalty**: Learn a map ($T : X^- \to X^+$) that trades off individual cost and deviation from the positive-class distribution ($\mathbb{P}^+$):

$$\underset{T \in \mathcal{M}(\mathcal{X}^-, \mathcal{X}^+)}{\arg\min} \left\{ \left(\mathbb{E}_{x \sim \mathbb{P}^-}[c(x, T(x))^q] \right)^{1/q} + \lambda D_{\chi^2}\left(T_\# \mathbb{P}^- \| \mathbb{P}^+ \right) \right\}.$$

- **Outcomes:** less competition, data-manifold closeness, robustness, individual fairness, and amortized inference.

Collective Counterfactual Explanations



Relaxing CCE to Unbalanced Optimal Transport (UOT)

- **Step 1: Map $\to$ Plan (Monge $\to$ Kantorovich).** Existence can fail for map problems, so relax to a coupling/plan $\pi \in \mathcal{P}(X^- \times X^+)$ with fixed source marginal $\pi_1 = \mathbb{P}^-$ and keep the χ^2 penalty on the target marginal π_2:

  $$\min_{\pi} \left(\mathbb{E}_{(x,y)\sim\pi}[c(x,y)^q]\right)^{1/q} + \eta\lambda^2 D_{\chi^2}(\pi_2 \| \mathbb{P}^+) \quad \text{s.t. } \pi_1 = \mathbb{P}^-.$$

  Intuition: plans subsume all maps ($\pi \sim (X, T(X))$) and are convex/compact in weak topology.

- **Step 2: Hard constraint $\to$ Soft penalty (to get UOT).** Replace the hard marginal constraint $\pi_1 = \mathbb{P}^-$ by a φ-**divergence penalty** $\lambda_1 D_\psi(\pi_1 \| P^-)$. This gives the **unbalanced OT** objective:

  $$\min_{\pi} \left(\mathbb{E}_{(x,y)\sim\pi}[c(x,y)^q]\right)^{1/q} + \lambda_1 D_\psi(\pi_1 \| \mathbb{P}^-) + \lambda_2 D_{\chi^2}(\pi_2 \| \mathbb{P}^+).$$

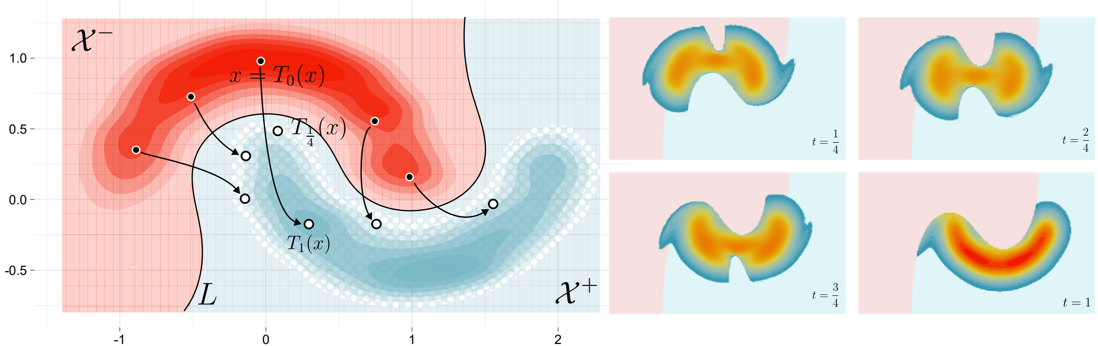
  Now, both marginals are softly matched—mass can be created/removed with a cost, which is exactly the UOT paradigm.

Compute Once, Explain Many!

- **Algorithmic payoff.** Picking $D_\psi = \mathrm{KL}$ (common in UOT) yields fast, projected-gradient solvers for CCE. This is why the relaxation is practical. The projected-gradient CCE solver runs in $O(Tmn)$ time for $m = |X^-|$, $n = |X^+|$ over T iterations.
- **Amortized Inference.** After this global solve, explanations are **amortized**—you read them directly from the learned plan π, i.e., **no additional per-person optimization** is needed. This removes the computational bottleneck of individual CE solves while staying faithful to collective dynamics.
- **Sinkhorn Algorithms.** Picking $D_{\chi^2} = \mathrm{KL}$ yields fast Sinkhorn-style solvers for CCE.

Beyond Point Recourse: Path-Guided CE

- *Path-Guided CE* gives a step-by-step route, not just a destination: introduce a time-indexed map $T_t(x)$, $t \in [0,1]$, that moves each individual from the current state x ($t{=}0$) to the target $T_1(x)$ ($t{=}1$) along a (near) constant-speed path.
- Use a back-and-forth scheme to approximate T_t and v_t; produce actionable intermediate states (mini-steps) that respect collective dynamics and avoid crowding while guiding users to feasible outcomes.



Numerical Experiments