

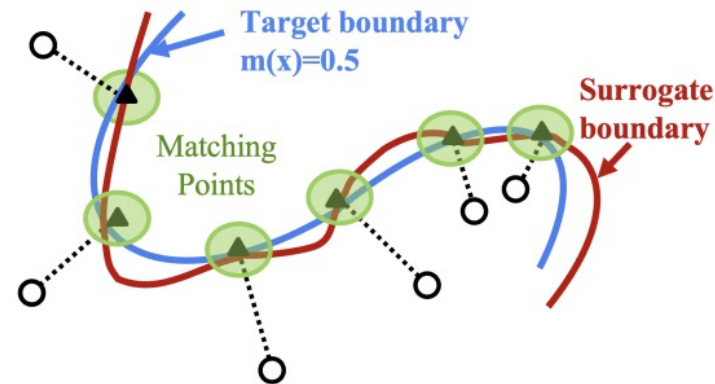


Model Reconstruction Using Counterfactual Explanations: A Perspective From Polytope Theory

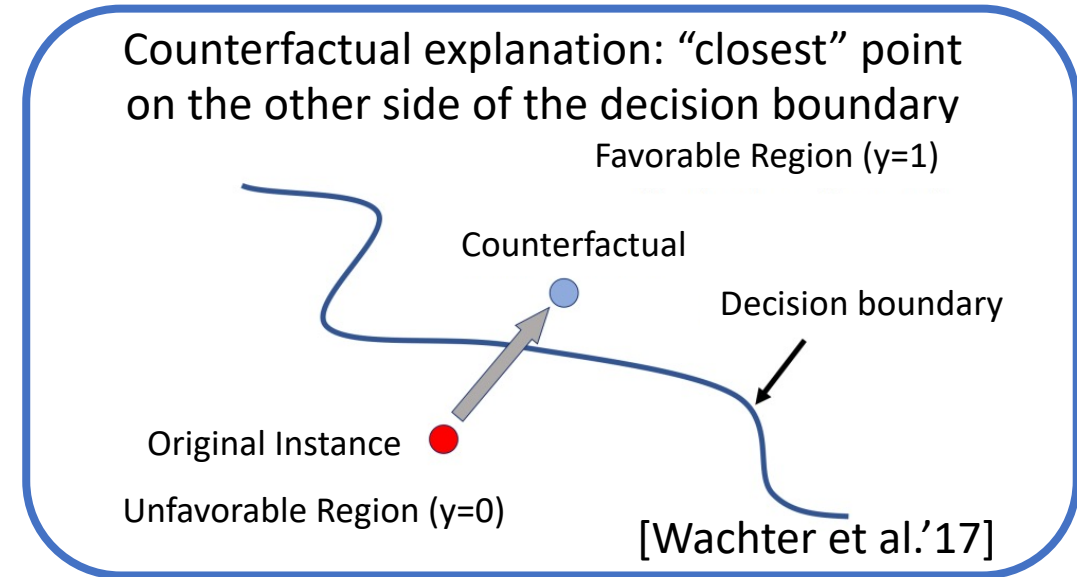
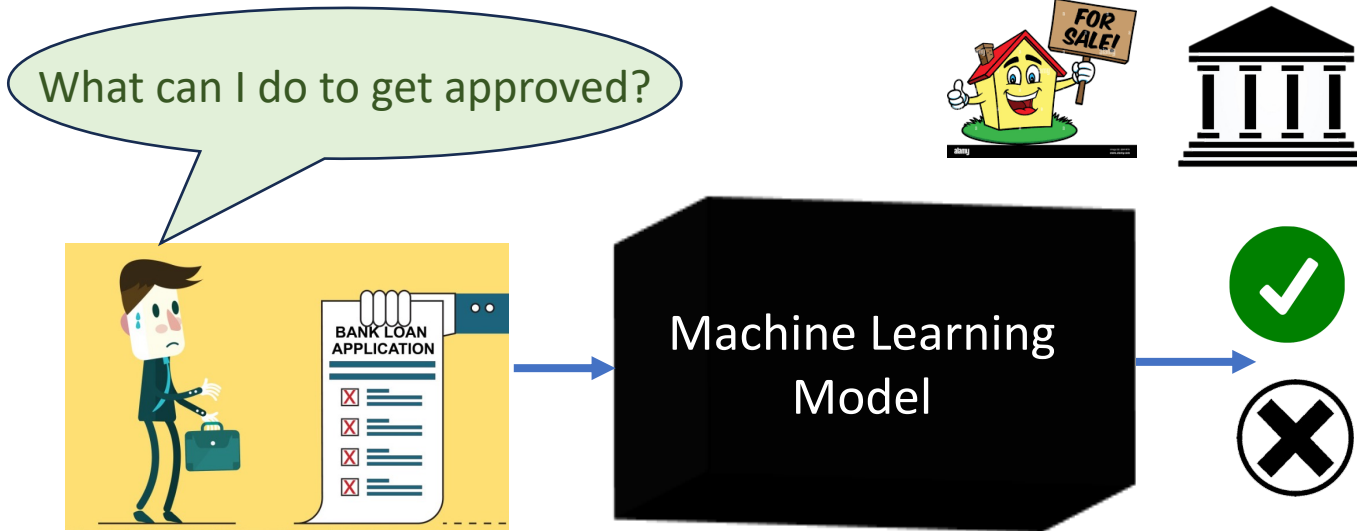
Pasan Dissanayake, Sanghamitra Dutta

University of Maryland College Park

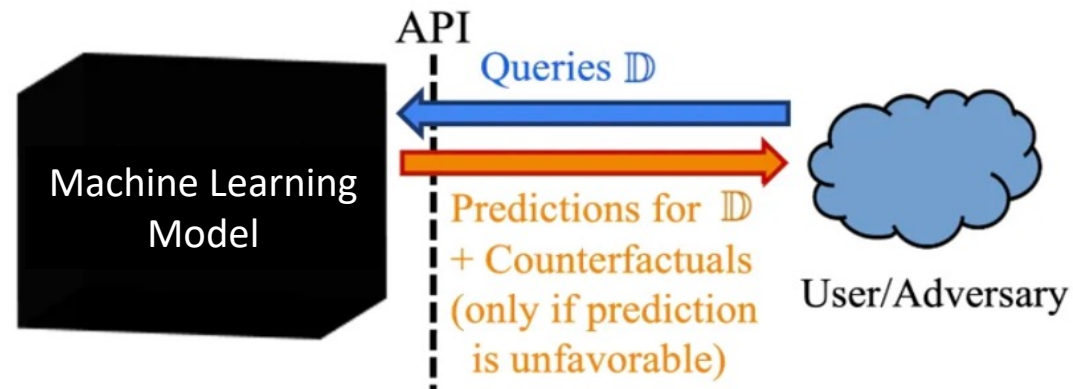
Poster: Wed 11 Dec 4:30 p.m. PST — 7:30 p.m. PST



What are Counterfactual Explanations?

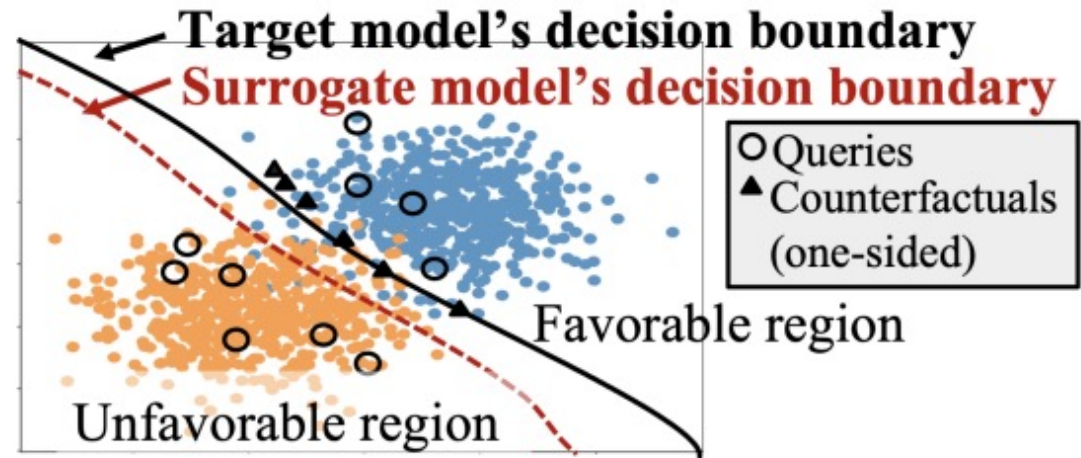


How faithfully can one reconstruct a model using counterfactual explanations?



Training a surrogate model using all the queried datapoints ($y=0/1$) and one-sided counterfactuals (for datapoints with $y=0$)

Counterfactuals as ordinary
labelled instances?
Decision boundary shift issue



Question: Can we improve model reconstruction specifically leveraging the fact that the counterfactuals are quite close to the boundary?

Main Contribution:

Novel Model Reconstruction Strategies Using Counterfactuals
With Theoretical Guarantees From Polytope Theory

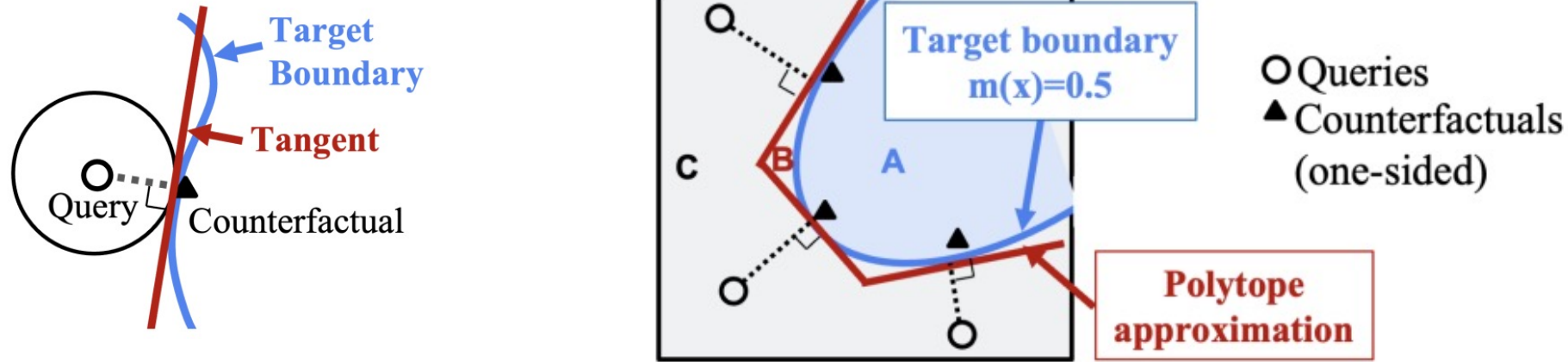
Related Works: [Aivodji et al.'20][Wang et al.'22]

Other Privacy + CF: [Pawelczyk et al.'23][Goethals et al.'23][Yadav et al.'23]

Model extraction in other settings: [Gong et al.'20][Milli et al.'19]

Main Results

1. Convex Decision Boundaries and Closest Counterfactuals

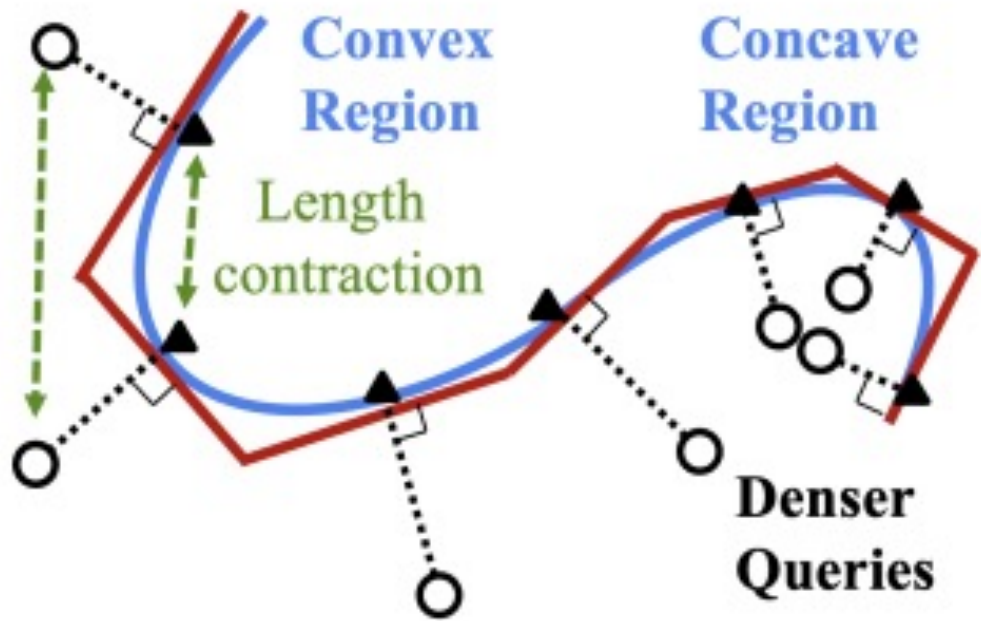


Theorem 3.2. *Let m be the target binary classifier whose decision boundary is convex (i.e., the set $\{\mathbf{x} \in [0, 1]^d : \lfloor m(\mathbf{x}) \rfloor = 1\}$ is convex) and has a continuous second derivative. Denote by \tilde{M}_n , the convex polytope approximation of m constructed with n supporting hyperplanes obtained through i.i.d. counterfactual queries. Assume that the fidelity is evaluated with respect to \mathbb{D}_{ref} which is uniformly distributed over $[0, 1]^d$. Then, when $n \rightarrow \infty$ the expected fidelity of \tilde{M}_n with respect to m is given by*

$$\mathbb{E} \left[\text{Fid}_{m, \mathbb{D}_{ref}}(\tilde{M}_n) \right] = 1 - \epsilon \quad (1)$$

where $\epsilon \sim \mathcal{O} \left(n^{-\frac{2}{d-1}} \right)$ and the expectation is over both \tilde{M}_n and \mathbb{D}_{ref} .

Theoretical guarantees on **exact volume approximation**
using **counterfactuals** leveraging polytope theory



Lipschitz
ReLU Networks

Main Results

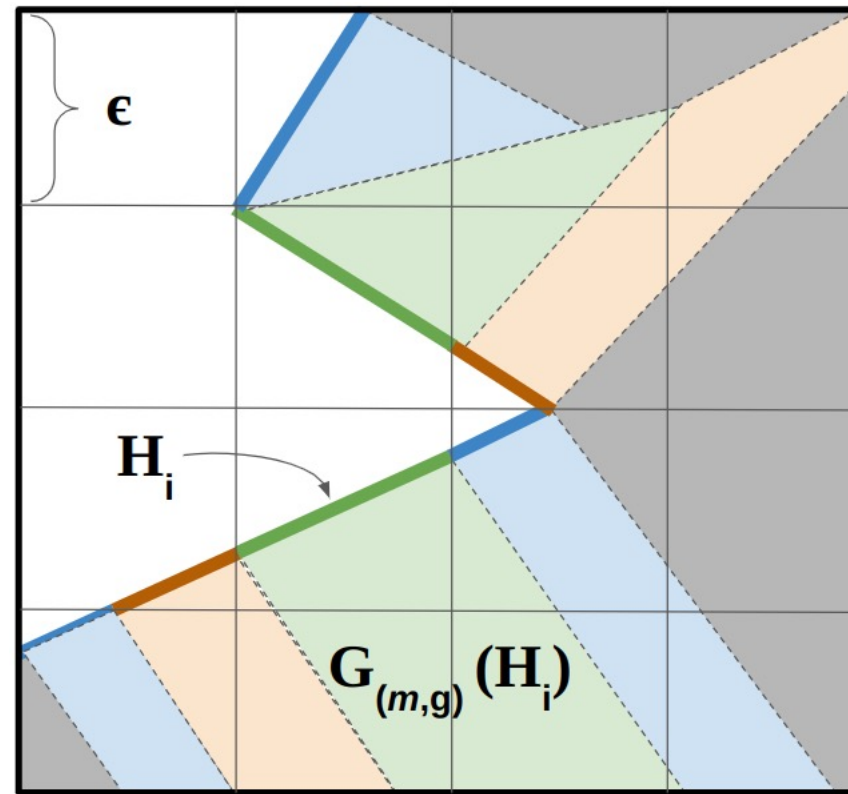
2. ReLU Networks and Closest Counterfactuals

Continuous Piece-Wise Linear (CPWL) Functions

Theorem 3.6. *Let m be a target binary classifier with ReLU activations. Let $k(\epsilon)$ be the number of cells through which the decision boundary passes. Define $\{\mathbb{H}_i\}_{i=1,\dots,k(\epsilon)}$ to be the set of affine pieces of the decision boundary within each decision boundary cell. Let $v_i(\epsilon) = V(\mathbb{G}_{m,g_m}(\mathbb{H}_i))$ where $V(\cdot)$ is the d -dimensional volume (i.e., the Lebesgue measure) and $\mathbb{G}_{m,g_m}(\cdot)$ is the inverse counterfactual region w.r.t. m and the closest counterfactual generator g_m . Then the probability of successful reconstruction with counterfactual queries distributed uniformly over $[0, 1]^d$ is lower-bounded as*

$$\mathbb{P}[\text{Reconstruction}] \geq 1 - k(\epsilon)(1 - v^*(\epsilon))^n \quad (2)$$

where $v^*(\epsilon) = \min_{i=1,\dots,k(\epsilon)} v_i(\epsilon)$ and n is the number of queries.

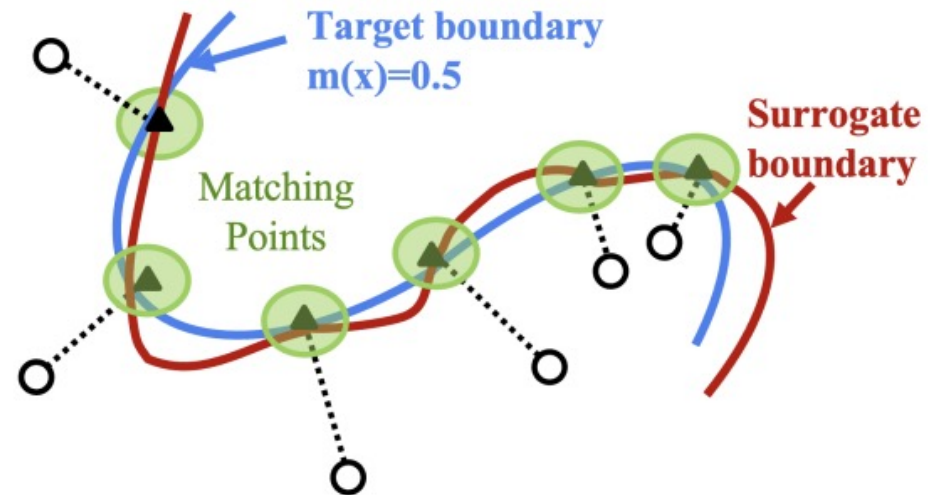


Main Results

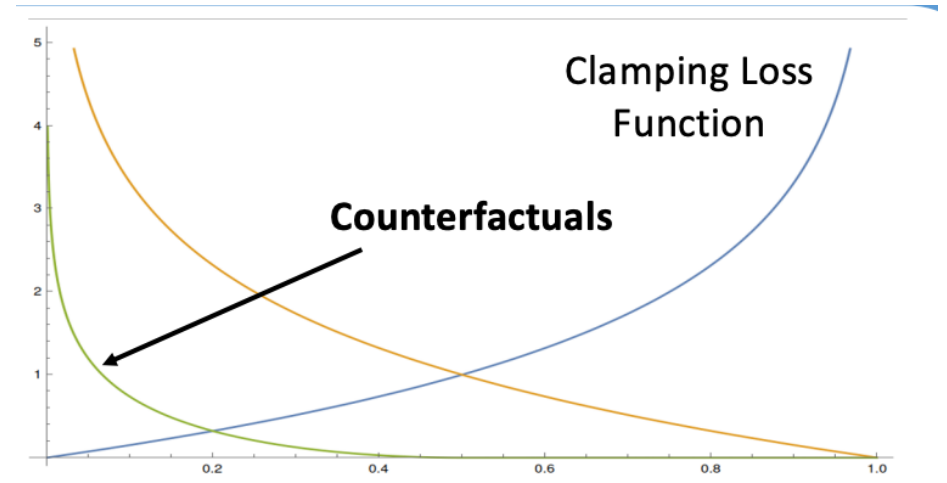
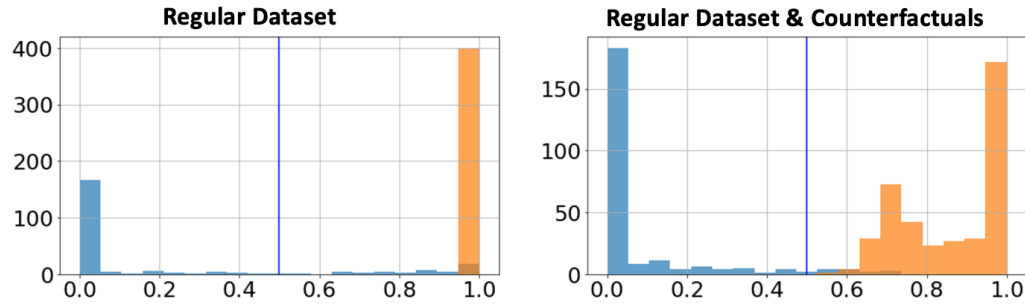
3. Beyond Closest Counterfactuals

Theorem 3.10. *Let the target m and surrogate \tilde{m} be ReLU classifiers such that $m(\mathbf{w}) = \tilde{m}(\mathbf{w})$ for every counterfactual \mathbf{w} . For any point \mathbf{x} that lies in a decision boundary cell, $|\tilde{m}(\mathbf{x}) - m(\mathbf{x})| \leq \sqrt{d}(\gamma_m + \gamma_{\tilde{m}})\epsilon$ holds with probability $p \geq 1 - k(\epsilon)(1 - v^*(\epsilon))^n$.*

Our Proposed Strategy:
Counterfactual Clamping Attack
(CCA)

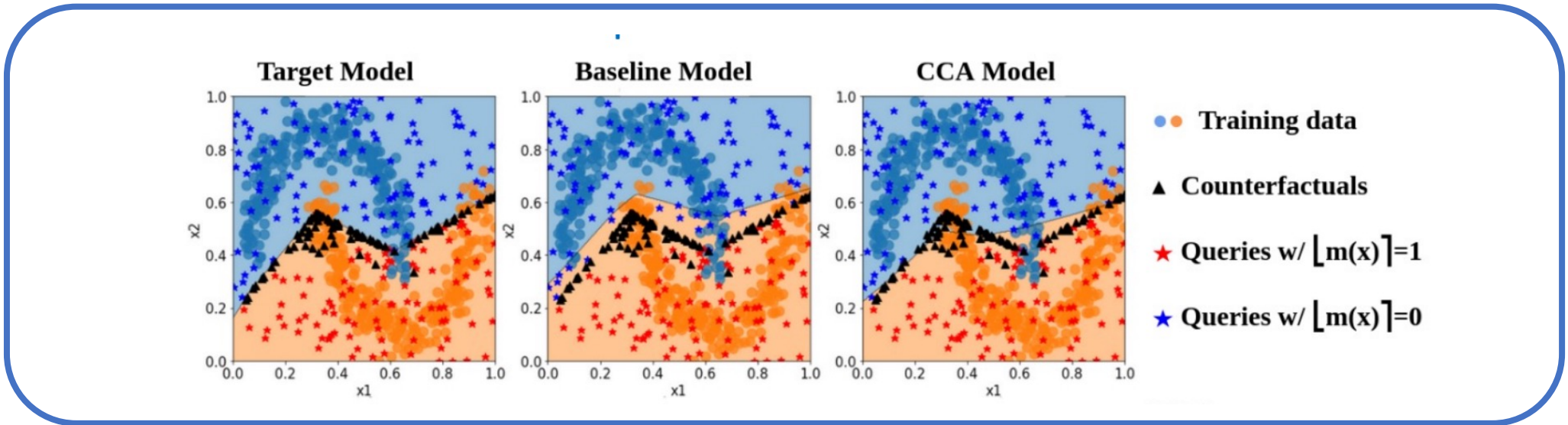


CCA Strategy: Unique Loss Function to Clamp Counterfactuals From One Side and Mitigate the Decision Boundary Shift Issue



$$L_k(\tilde{m}(\mathbf{x}), y_{\mathbf{x}}) = \mathbb{1}[y_{\mathbf{x}} = 0.5, \tilde{m}(\mathbf{x}) \leq k] \{L(\tilde{m}(\mathbf{x}), k) - h(k)\} + \mathbb{1}[y_{\mathbf{x}} \neq 0.5] L(\tilde{m}(\mathbf{x}), y_{\mathbf{x}})$$

neglect CFs that already have $g(w) > k$
for CFs
for normal examples



Experimental Validation: Fidelity Comparison Over Several Benchmark Datasets

Dataset	Architecture known (model 0)				Architecture unknown (model 1)			
	\mathbb{D}_{test}		\mathbb{D}_{uni}		\mathbb{D}_{test}		\mathbb{D}_{uni}	
	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
Adult In.	91±3.2	94±3.2	84±3.2	91±3.2	91±4.5	94±3.2	84±3.2	90±3.2
COMPAS	92±3.2	96±2.0	94±1.7	96±2.0	91±8.9	96±3.2	94±2.0	94±8.9
DCCC	89±8.9	99±0.9	95±2.2	96±1.4	90±7.7	97±4.5	95±2.2	95±11.8
HELOC	91±4.7	96±2.2	92±2.8	94±2.4	90±7.4	95±5.5	91±3.3	93±3.2

CCA provides high-fidelity model reconstruction

Comparison With Two-Sided Counterfactuals

Dataset		Architecture known (model 0)								Architecture unknown (model 1)							
		\mathbb{D}_{test}				\mathbb{D}_{uni}				\mathbb{D}_{test}				\mathbb{D}_{uni}			
		Base.	Dual.	CCA1	CCA2	Base.	Dual.	CCA1	CCA2	Base.	Dual.	CCA1	CCA2	Base.	Dual.	CCA1	CCA2
DCCC	n=100	0.95	0.99	0.94	0.99	0.90	0.95	0.92	0.97	0.92	0.98	0.93	0.98	0.88	0.92	0.89	0.93
	n=200	0.96	0.99	0.98	0.99	0.90	0.96	0.95	0.98	0.96	0.99	0.96	0.99	0.89	0.94	0.94	0.96
HELOC	n=100	0.94	0.97	0.90	0.98	0.91	0.98	0.84	0.98	0.92	0.91	0.90	0.96	0.88	0.92	0.84	0.96
	n=200	0.96	0.98	0.92	0.98	0.93	0.98	0.89	0.99	0.95	0.92	0.91	0.97	0.93	0.94	0.88	0.97

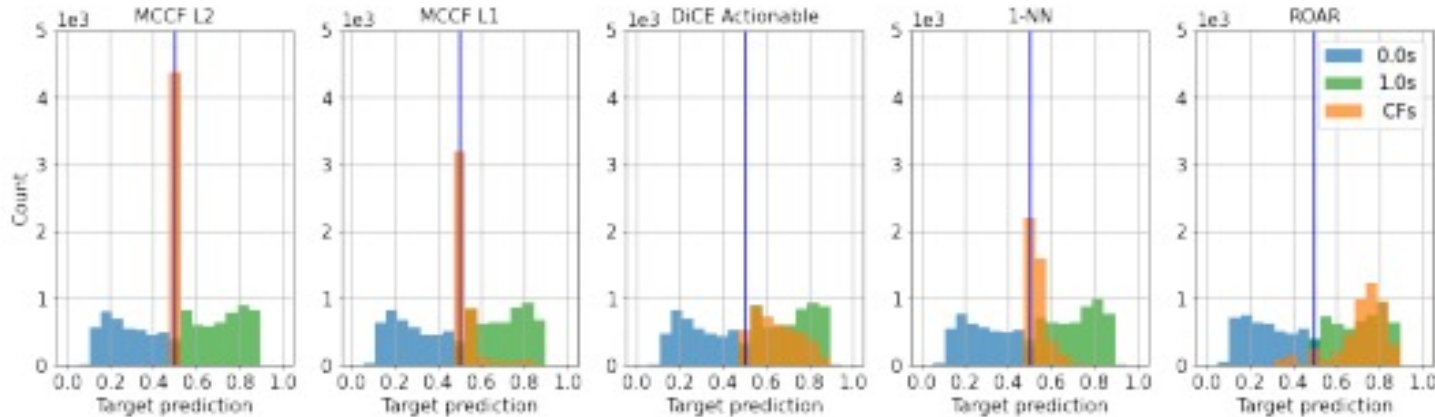
Baselines: [Aivodji et al.'20][Wang et al.'22]

Additional Experiments

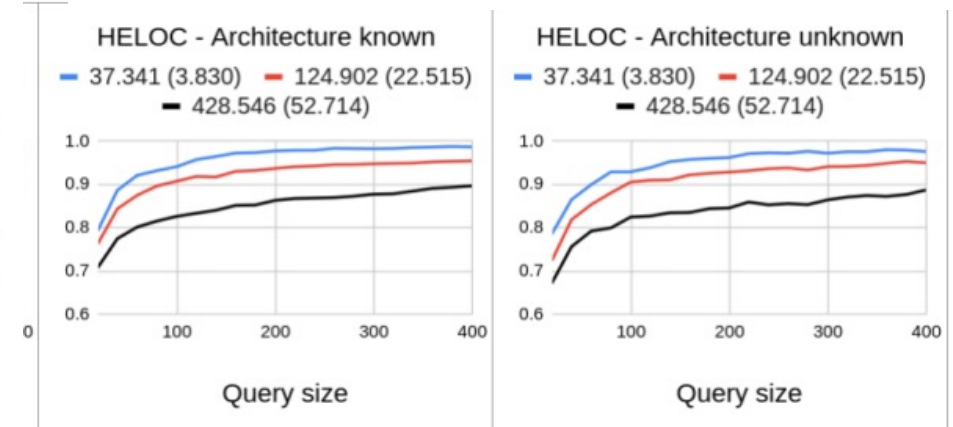
Other Counterfactual Generation Techniques

Table 2: Fidelity achieved with different counterfactual generating methods on HELOC dataset. Target model has hidden layers with neurons (20, 30, 10). Surrogate model architecture is (10, 20).

CF method	Fidelity over \mathbb{D}_{test}				Fidelity over \mathbb{D}_{uni}			
	n=100		n=200		n=100		n=200	
	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
MCCF L2-norm	91	95	93	96	91	93	93	95
MCCF L1-norm	93	95	94	96	89	92	91	95
DiCE Actionable	93	94	95	95	90	91	93	94
1-Nearest-Neighbor	93	95	94	96	93	93	94	95
ROAR [Upadhyay et al., 2021]	91	92	93	95	87	85	92	92
C-CHVAE [Pawelczyk et al., 2020]	77	80	78	82	90	89	85	78



Different Lipschitz Constants



Different Model Architectures

Dataset: HELOC - Fidelity over \mathbb{D}_{test}										
Target archi. →	(20,10)		(20,10,5)		(20,20,10,5)					
	n=100	n=200	n=100	n=200	n=100	n=200	n=100	n=200	n=100	n=200
Surrogate archi.	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
(20,10)	0.90	0.94	0.91	0.95	0.90	0.94	0.92	0.95	0.98	0.99
(20,10,5)	0.88	0.92	0.92	0.95	0.89	0.92	0.92	0.95	0.98	0.98
(20,20,10,5)	0.87	0.93	0.91	0.93	0.87	0.89	0.91	0.94	0.98	0.98

Dataset: HELOC - Fidelity over \mathbb{D}_{uni}										
Target archi. →	(20,10)		(20,10,5)		(20,20,10,5)					
	n=100	n=200	n=100	n=200	n=100	n=200	n=100	n=200	n=100	n=200
Surrogate archi.	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
(20,10)	0.92	0.92	0.94	0.95	0.91	0.91	0.94	0.95	0.98	0.98
(20,10,5)	0.91	0.90	0.94	0.93	0.91	0.89	0.93	0.94	0.97	0.97
(20,20,10,5)	0.91	0.91	0.93	0.94	0.91	0.87	0.93	0.92	0.97	0.97

CCA mostly outperforms baselines and gives high-fidelity model reconstruction!

Potential defenses: (i) Noisy Counterfactuals, or (ii) Robust Counterfactuals

Thank You!

Poster: Wed 11 Dec 4:30 p.m. PST — 7:30 p.m. PST



Model Reconstruction Using Counterfactual Explanations: A Perspective From Polytope Theory

Pasan Dissanayake, Sanghamitra Dutta

Department of Electrical and Computer Engineering, University of Maryland College Park



<https://arxiv.org/abs/2405.05369>



NeurIPS 2024

MOTIVATION

What can I do to get approved?

Machine Learning Model

Unfavourable region ($y=0$)

Decision boundary

Original instance

Favourable region ($y=1$)

Counterfactual

Definition: A counterfactual explanation for a given instance x is a point x_c such that $m(x_c) \neq m(x)$ (i.e. lies on the opposite side of the decision boundary), selected based on some criteria.

The **closest counterfactual** is the counterfactual which is closest to x , under some distance metric.

API: Queries D → Predictions for D → Counterfactuals (only if prediction is unfavourable)

Machine Learning Model ($m(x)$)

User/Adversary

How faithfully can one reconstruct a model using counterfactual explanations?

Create attack set \mathcal{D} → Query "m" with \mathcal{D} for labels+CFs → Train "m" on \mathcal{D}

Counterfactuals treated as ordinary labelled instances? → Boundary shift issue

Target model's decision boundary

Surrogate model's decision boundary

Unfavourable region

Favourable region

Question: Can we improve model reconstruction using counterfactuals specifically leveraging the fact that the counterfactuals are quite close to the boundary?

Main Contribution:
Novel Model Reconstruction Strategies & Fundamental Limits

Related Works: [Aivodji et al.'20][Wang et al.'22][Yadav et al.'23] Other Privacy + CF: [Pawelczyk et al.'23][Goethals et al.'23] Model extraction: [Gong et al.'20][Milli et al.'19]

MAIN RESULTS

1. Convex Decision Boundaries and Closest Counterfactuals

Theoretical guarantees on **volume approximation** using counterfactuals leveraging polytope theory

Theorem 3.2. Let m be the target binary classifier whose decision boundary is convex (i.e. the set $\{x \in [0, 1]^d : m(x) = 1\}$ is convex) and has a continuous second derivative. Denote by \tilde{M}_n the convex polytope approximation of m constructed with n supporting hyperplanes obtained through i.i.d. counterfactual queries. Assume that the fidelity is evaluated with respect to D_{vol} which is uniformly distributed over $[0, 1]^d$. Then, when $n \rightarrow \infty$ the expected fidelity of \tilde{M}_n with respect to m is given by

$$\mathbb{E}[F_{D_{vol}, \tilde{M}_n}(m)] \approx 1 - \epsilon \quad (1)$$

where $\epsilon \sim O(n^{-\frac{2}{d}})$ and the expectation is over both \tilde{M}_n and D_{vol} .

Convex Regions

Concave Regions

Length constraint

Queries

Convex Non-Convex Lipschitz ReLU Networks

2. ReLU Networks and Closest Counterfactuals

Continuous Piece-Wise Linear (CPWL) Functions

Theorem 3.6. Let m be a target binary classifier with ReLU activations. Let $k(\epsilon)$ be the number of cells through which the decision boundary passes. Define $\{H_1, \dots, H_{k(\epsilon)}\}$ to be the set of affine pieces of the decision boundary within each decision boundary cell. Let $v_i(\epsilon) = V(\mathcal{G}_{m, \epsilon}(H_i))$ where $V(\cdot)$ is the d -dimensional volume (i.e. the Lebesgue measure) and $\mathcal{G}_{m, \epsilon}(\cdot)$ is the inverse counterfactual region w.r.t. m and the closest counterfactual generator $\mathcal{G}_{m, \epsilon}$. Then the probability of successful reconstruction with counterfactual queries distributed uniformly over $[0, 1]^d$ is lower-bounded as

$$\mathbb{P}[\text{Reconstruction}] \geq 1 - k(\epsilon)(1 - v^*(\epsilon))^n \quad (2)$$

where $v^*(\epsilon) = \min_{i=1, \dots, k(\epsilon)} v_i(\epsilon)$ and n is the number of queries.

3. Beyond Closest Counterfactuals

Theorem 3.10. Let the target m and surrogate \tilde{m} be ReLU classifiers such that $m(w) = \tilde{m}(w)$ for every counterfactual w . For any point x that lies in a decision boundary cell, $\|m(x) - \tilde{m}(x)\| \leq \sqrt{d}(\tau_m + \tau_{\tilde{m}})$ holds with probability $p \geq 1 - k(\epsilon)(1 - v^*(\epsilon))^n$.

Our Proposed Algorithm: Counterfactual Clamping Attack (CCA)

Author Contacts: pasand@umd.edu, sanghamd@umd.edu

ALGORITHM & EXPERIMENTS

$L_1(\tilde{m}(x), m(x)) = 1$ ($y_x = 0, \tilde{m}(x) \leq 0$) $L_1(\tilde{m}(x), m(x)) = 1$ ($y_x \neq 0, \tilde{m}(x) > 0$)

neglect CFA that already have $|g(x)| > \epsilon$ for CFA for normal examples

Clamping Loss Function

Counterfactuals

Training data

Counterfactuals

Queries w/ $|m(x)| = 1$

Queries w/ $|m(x)| = 0$

Fidelity Comparisons

Dataset	Architecture known (model 0)				Architecture unknown (model 1)			
	Base	CCA	Base	CCA	Base	CCA	Base	CCA
Adult In.	91±3.2	94±3.2	84±3.2	91±3.2	91±4.5	94±3.2	84±3.2	90±3.2
COMPAS	92±3.2	96±2.0	94±1.7	96±2.0	91±2.9	96±3.2	94±2.0	94±8.9
DCCC	89±8.9	99±0.9	95±2.2	96±1.4	90±7.7	97±4.5	95±2.2	95±11.8
HELLOC	91±4.7	96±2.2	92±2.8	94±2.4	90±7.4	95±5.5	91±3.3	93±3.2

Table 6. Comparison with DualCFX. Legend: Base—Baseline model based on [Aivodji et al., 2020]. Dual-DualCFX, CCA1—CCA with one-sided counterfactuals, CCA2—CCA with counterfactuals from both sides.

Architecture known (model 0)

Dataset	Dual				Dual					
	Base	CCA1	CCA2	Base	CCA1	CCA2	Base	CCA1	CCA2	
DCCC	0.96	0.99	0.98	0.99	0.96	0.95	0.98	0.96	0.99	0.94
HELLOC	0.96	0.98	0.97	0.98	0.96	0.94	0.96	0.97	0.96	0.94

Architecture unknown (model 1)

Dataset	Dual				Dual					
	Base	CCA1	CCA2	Base	CCA1	CCA2	Base	CCA1	CCA2	
DCCC	0.96	0.98	0.97	0.98	0.96	0.94	0.96	0.97	0.96	0.94
HELLOC	0.96	0.98	0.97	0.98	0.96	0.94	0.96	0.97	0.96	0.94

CCA provides high-fidelity model reconstruction

Different Lipschitz Constants

Different Model Architectures

Other Counterfactual Generation Strategies

Table 2. Fidelity achieved with different counterfactual generation methods on HELLOC dataset. Target model has hidden layers with neurons {20, 30, 10}. Surrogate model architectures in {10, 20, 30}.

CF method	n=100			n=200			n=100			n=200		
	Base	CCA	CCA	Base	CCA	CCA	Base	CCA	CCA	Base	CCA	CCA
MCFL2.0-learn	91	91	91	91	91	91	91	91	91	91	91	91
MCFL2.0-learn	91	91	91	91	91	91	91	91	91	91	91	91
DCCC-AutoML	91	91	91	91	91	91	91	91	91	91	91	91
4-NearestNeighbor	91	91	91	91	91	91	91	91	91	91	91	91
Model Extraction (Gong et al., 2020)	91	91	91	91	91	91	91	91	91	91	91	91
CFM (Pawelczyk et al., 2023)	91	91	91	91	91	91	91	91	91	91	91	91

CCA outperforms baselines and gives high-fidelity model reconstruction!

Potential defenses:

(i) Noisy Counterfactuals

(ii) Robust Counterfactuals

References:

[1] U. Aivodji, A. Bolei, and S. Gamba. Model extraction from counterfactual explanations. arXiv:2009.01884, 2020.

[2] Y. Wang, L. Qian, and C. Mao. DualCF: Efficient model extraction attack from counterfactual explanations. In ICLR, 2022.

[3] C. Wiles, M. Washburn, and K. Chaturvedi. Sauris: A theoretical look at auditing with explanations. arXiv:2206.06740, 2023.

[4] M. Pawelczyk, M. Laskaraju, and S. Axel. On the privacy risks of algorithmic recourse. In International Conference on Artificial Intelligence and Statistics, 2023.

[5] F. Hammar, E. Noorani, S. Mishra, D. Magazzini, and S. Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In ICML, 2023.



<https://arxiv.org/abs/2405.05369>

<https://github.com/pasandissanayake/model-reconstruction-using-counterfactuals>

Author Contacts:

pasand@umd.edu, sanghamd@umd.edu

Broader Implications on the Interplay Between Explainability & Privacy