Towards Robust and Reliable Algorithmic Recourse

Sohini Upadhyay, Shalmali Joshi, Hima!

Motivation

Algorithmic Recourse

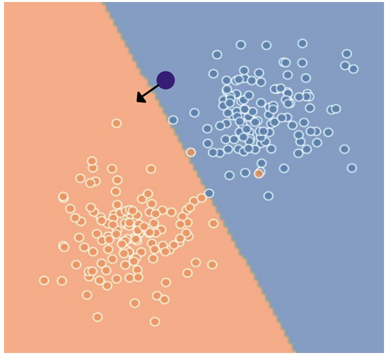
- ML models are deployed in high stakes scenarios
- If you receive an unfavorable outcome as a result of a prediction, how can you reverse it?
- Ex: A bank might tell you to increase your salary by $10,000

Model Updates

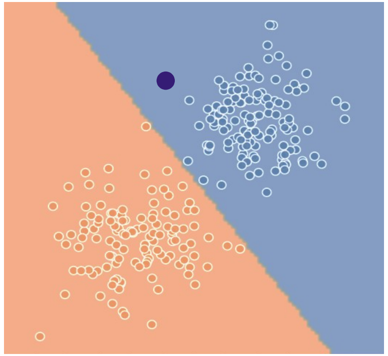
- In practice, data collectors are (hopefully) frequently updating their datasets
- Models are updated to reflect dataset changes
- Current algorithms to generate counterfactuals assume models are static

Motivation: An Example

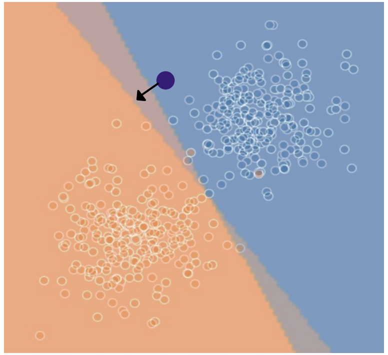
Model trained on $data_1$



Model trained on $data_2$



Overlaid



Original suggested recourse for the datapoint no longer crosses the decision boundary when the model is updated/retrained

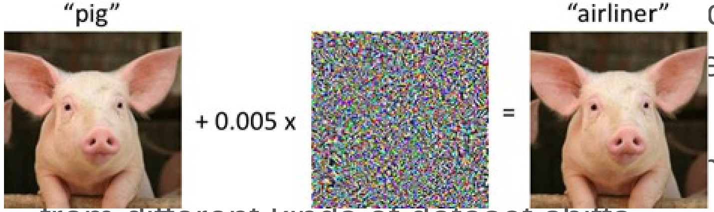
Summary of Contributions

- Outline model + data shifts that people should consider
  - Temporal shift
  - Geospatial shift
  - Data correction shift
- Propose a method for finding RObust Algorithmic Recourse (ROAR)
  - Introduces a novel minimax objective that can be used to construct robust actionable recourses while minimizing the recourse costs
- Theoretical analysis
  - How bad are regular counterfactuals under model shifts?
  - How much does the proposed method increase the cost of recourses when compared to normal CFs?
- Experimental analysis

Related work

Algorithmic recourse

- [Can i still trust you?: Understanding the



  from different kinds of dataset shifts

Adversarial Training

- Optimizes a **minimax objective** that captures the worst-case loss over a given set of perturbations to the input data
- At each gradient step, **computes the gradient at worst-case perturbation**
- Recent work explores the construction of other kinds of explanations (feature attribution and rule based explanations) that are robust to *dataset* shifts

Background: Recourse/Counterfactuals

Optimal CF:

$$x' = \arg\min_{x' \in \mathcal{A}} c(x, x') \quad \text{s.t} \quad \mathcal{M}(x') = 1$$

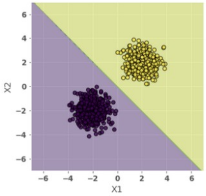
Unconstrained and differentiable relaxation:

$$x' = \arg\min_{x' \in \mathcal{A}} \ell(\mathcal{M}(x'), 1) + \lambda\, c(x, x')$$

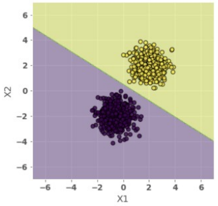
Notation Notes:

- *x*: specific data point
- *x'*: counterfactual
- *M*: model (or linear approximation of model around point x)
- *C*: cost function (how hard is it to achieve the counterfactual)
- *A*: actionable/possible counterfactuals

A Primer?



(a) Original data & $\mathcal{M}_1$



(b) Mean shift

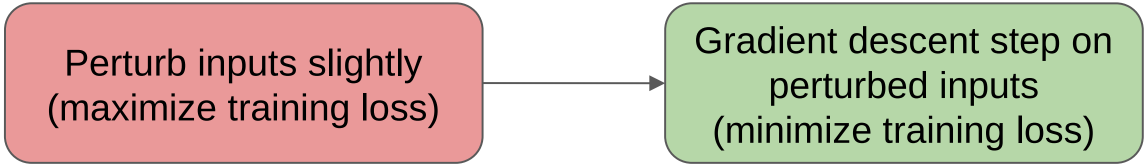
Model$_1$ parameters

Model$_2$ parameters

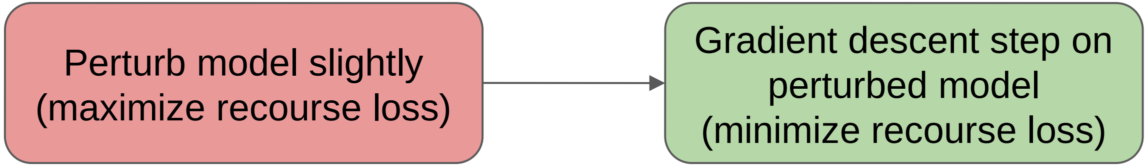
How to compare these models? (especially if they are nonlinear and high-dimensional)

Approach (Intuition)

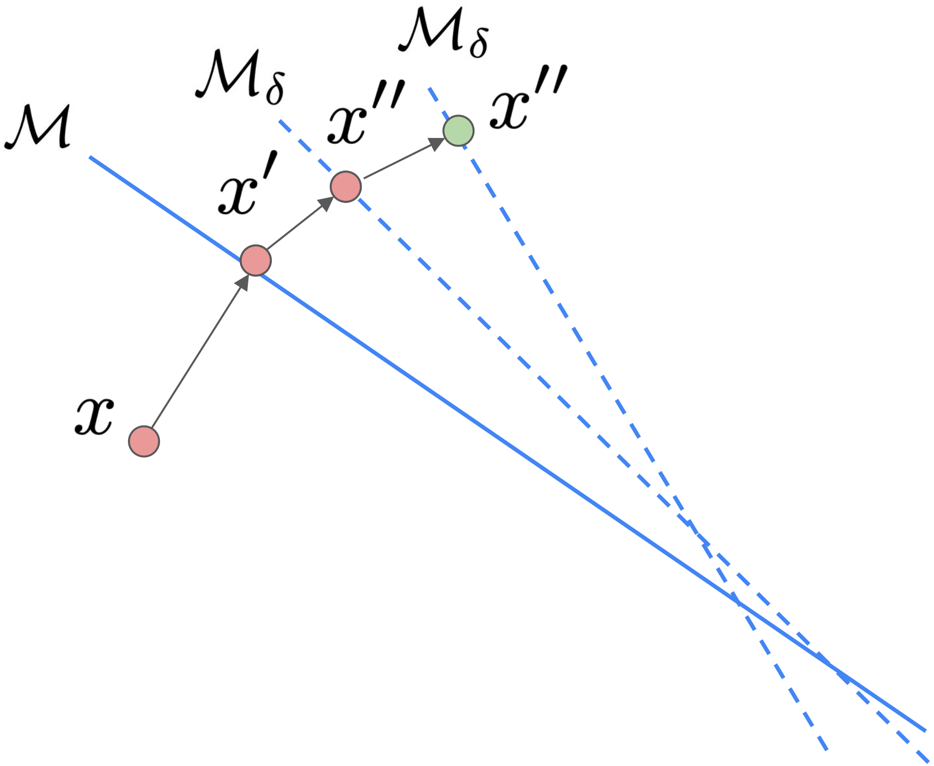
- Adversarial Training



- Robust Actional Recourse (ROAR)



Approach (Intuition)



$\delta \in \Delta$

Approach (Details)

$$x'' = \arg\min_{x'' \in \mathcal{A}} \max_{\delta \in \Delta} \ell(f_{w+\delta}(x''), 1) + \lambda c(x, x'')$$

Algorithm 1 Our Optimization Procedure

Input: x s.t. $f_w(x) = 0, f_w, \lambda > 0, \Delta,$ learning rate $\alpha > 0$.

Initialize $x'' = x, g = 0$

repeat

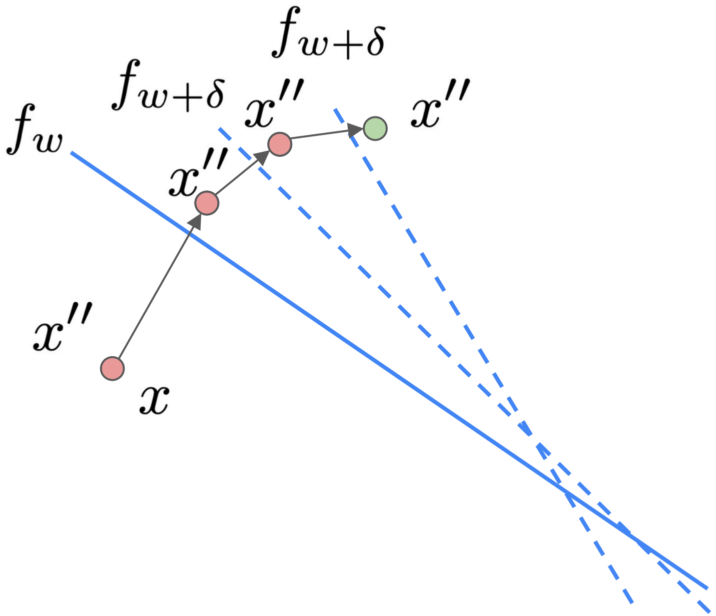
$\hat{\delta} = \arg\max_{\delta \in \Delta} \ell(f_{w+\delta}(x''), 1)$

$g = \nabla\left[\ell(f_{w+\hat{\delta}}(x''), 1) + \lambda c(x'', x)\right]$

$x'' -= \alpha g$

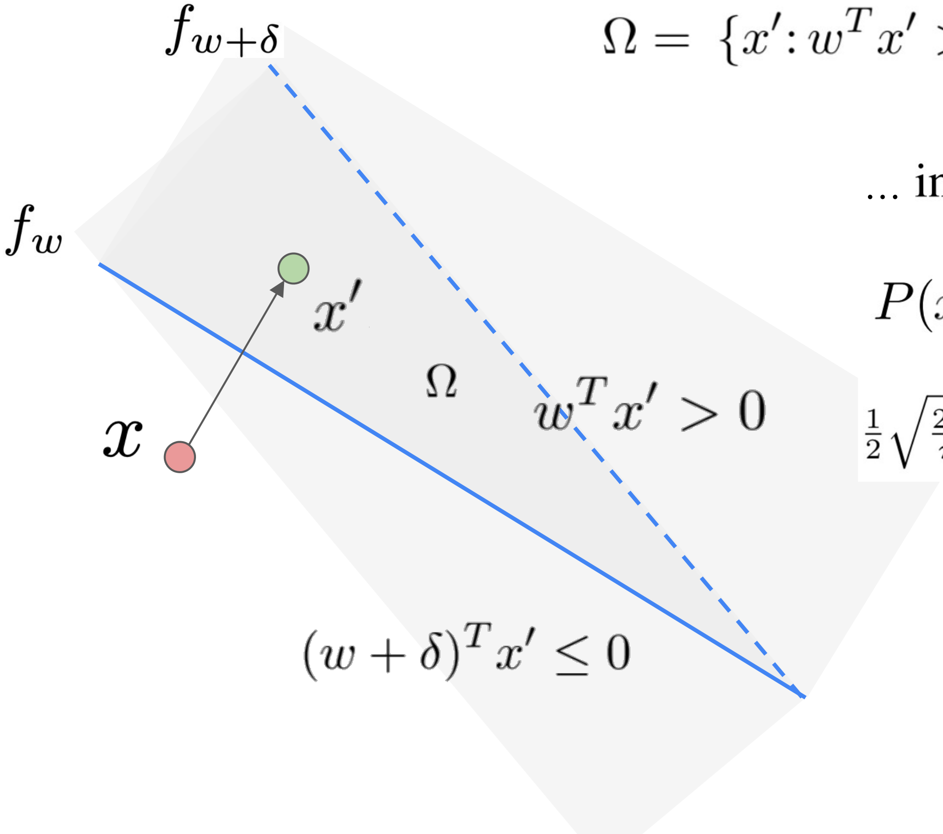
until convergence

Return x''



Proofs

Theorem 1



$$\Omega = \{x' \colon w^T x' > 0 \cap (w+\delta)^T x' \leq 0\}$$

... integrating over Ω

$$P(x' \text{is invalidated}) \geq$$

$$\frac{1}{2}\sqrt{\frac{2e}{\pi}}\frac{\sqrt{\beta-1}}{\beta}\exp^{-\beta\frac{(w^T\mu)^2}{2\|\sqrt{D}Uw\|^2}}$$

Assumptions

$x \sim \mathcal{N}(\mu, \Sigma)$

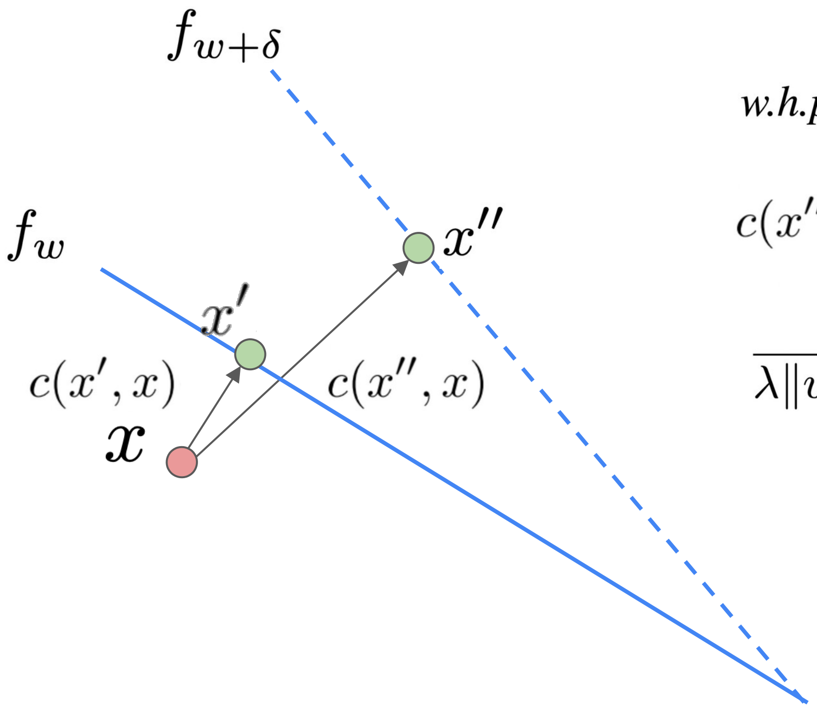
$x' \sim \mathcal{N}(\mu, \Sigma)$

$\Sigma = UDU^T$

$\beta \geq 1$

Proofs

Theorem 2



$$w.h.p.\ (1-\eta')$$

$$c(x'', x) \leq c(x', x) + \frac{1}{\lambda \|w+\delta\|}\alpha(w+\delta)^T\mu + \sqrt{\frac{D^2}{2}\log\left(\frac{1}{\eta'}\right)}$$

Assumptions

- Log-loss + simple model
- Optimal solution for x ''
- D is a bound for the "diameter" of dataset (l2)

Experimental results

- Real World Datasets
  - German Credit
    - Data Correction Shift
  - Small Business Administration
    - Temporal Shifts
  - Portuguese Student Performance
    - Geospatial Shift

- Synthetic Datasets
  - 2 Gaussians
    - Mean Shift
    - Variance Shift
    - Mean & Variance Shift

- Models
  - LR
  - SVM
  - 3 Layer Deep NN
- Cost (Distance) Functions
  - L1
  - Pairwise Feature Comparison (PFC)
    - Bradley Terry Comparison to parametrize p(i,j) = probability that feature i is less actionable than feature j
- Metrics:
  - Cost
    - How close is our counterfactual?
  - Validity
    - When undertaking recourse, (after a model shift) does it actually work?

Experimental Results: Logistic Regression

			Geospatial Shift		
Model	Cost	Recourse	Avg Cost	$\mathcal{M}_1$ Validity	$\mathcal{M}_2$ Validity
LR	L1	CFE	8.44 ± 0.30	1.00 ± 0.00	0.23 ± 0.04
		AR	5.24 ± 0.21	1.00 ± 0.00	0.35 ± 0.11
		ROAR	11.07 ± 0.48	1.00 ± 0.00	$\mathbf{0.68 \pm 0.05}$
		CR	NA	NA	NA
	PFC	CFE	0.34 ± 0.03	1.00 ± 0.00	0.18 ± 0.04
		AR	0.32 ± 0.02	1.00 ± 0.00	0.23 ± 0.04
		ROAR	1.13 ± 0.03	1.00 ± 0.00	$\mathbf{0.88 \pm 0.06}$
		CR	NA	NA	NA

$$\arg\min_{x'' \in \mathcal{A}} \max_{\delta \in \Delta} \ell(\mathcal{M}_\delta(x''), 1) + \lambda c(x, x'')$$

$$\arg\min_{x'} \max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

Experimental Results: Deep NN

			Avg Cost	$\mathcal{M}_1$ Validity	$\mathcal{M}_2$ Validity
NN	L1	CFE	9.50 ± 0.50	1.00 ± 0.00	0.46 ± 0.05
		AR-LIME	6.99 ± 0.29	0.60 ± 0.05	0.76 ± 0.08
		ROAR-LIME	18.65 ± 2.07	0.997 ± 0.003	$\mathbf{0.98} \pm \mathbf{0.01}$
		CR	NA	NA	NA
	PFC	CFE	0.44 ± 0.04	1.00 ± 0.00	0.36 ± 0.09
		AR-LIME	0.60 ± 0.09	0.57 ± 0.06	0.53 ± 0.05
		ROAR-LIME	1.59 ± 0.15	1.00 ± 0.00	$\mathbf{0.96} \pm \mathbf{0.01}$
		CR	NA	NA	NA

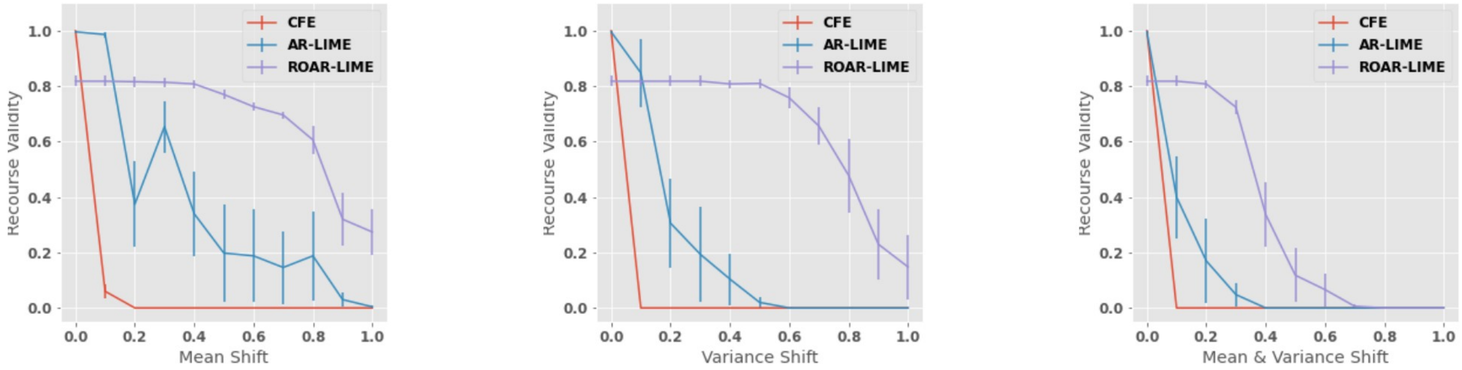
Takeaways

- Cost increases with robust recourse in general
  - The authors bound this!
- Linear approximations to complex models harms M1 validity
  - Robustness can actually helps!
- We can optimize for M1 validity by making our loss function more complex

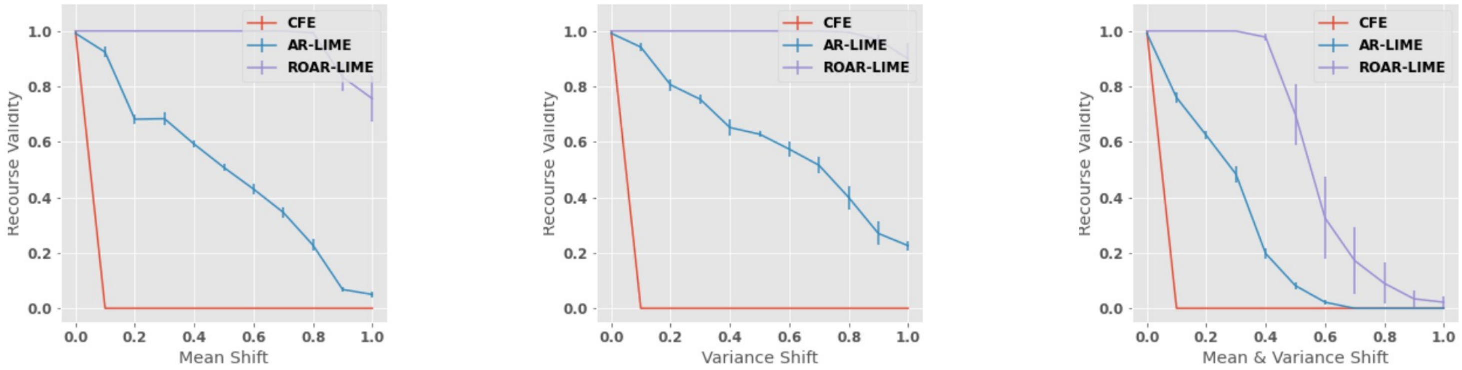
$$\arg\min_{x''}\max_{\delta}\max_{\lambda}\lambda\ell(M(x''), 1)+c(x, x'')$$

Experimental Results: Shift and Validity

L1:



PFC:



Conclusions

- Novel minimax objective and optimization strategy

- Bounds on error for non-robust counterfactuals under data shifts

- Bounds on cost of robust optimization

- Empirical results in both real world and synthetic scenarios validating method

Questions/Discussion

1. Do you think this problem should be worked on more from the CS/ML side or more from the policy side?
2. Do you think other types of explanations (that are not recourse motivated) need to be similarly robust?
3. What happens if model architecture/training dynamics change?
4. What incentive is there for companies/data controllers to implement this?