Trusted Artificial Intelligence

Homework 3

Spring 2024

Taha Majlesi

ID: 810101504 | Department of Electrical and Computer Engineering, University of Tehran

Instructor: Dr. Mostafa Tavasolipour

Submitted: February 13, 2026

Abstract. This report provides a complete long-form implementation report for the causal recourse part of HW3 (Question 5) using the health dataset. The pipeline is fully executable from local code, including: (1) constrained data processing and actionability, (2) classifier training for linear and MLP models, (3) completed health SCM and Jacobian, (4) recourse evaluation across robustness radii, and (5) automatic export of publication-ready report figures.
The document includes direct answers to each Q5 sub-question, equations, implementation traceability (file-level), reproducibility commands, aggregated quantitative results over multiple seeds, and error analysis. In addition, Nearest Counterfactual and Causal Algorithmic Recourse are compared empirically under matched conditions.

Contents

1 Scope and Completion Statement

The practical deliverable in this repository is the causal recourse implementation located under `code/Q5_codes`. This report therefore focuses on full completion of Question 5 requirements from the assignment, including data constraints, SCM completion, Jacobian completion, and quantitative comparison of recourse methods.

1.1 What is completed

The following deliverables are complete and verified by executable runs:

- Health data processing constraints for actionable features and feature bounds.
- End-to-end training and evaluation for linear and MLP classifiers.
- Full `Health_SCM` implementation with structural equations, inverse mapping, and Jacobian.
- Recourse runs for 10 negatively classified instances under multiple robustness radii.
- Automatic generation of report artifacts (figures and summary CSV tables).
- PDF report build with included results and reproducibility commands.

1.2 Key modified files

- `code/Q5_codes/scm.py`
- `code/Q5_codes/evaluate_recourse.py`
- `code/Q5_codes/utils.py`
- `code/Q5_codes/plot_report_figures.py`
- `code/Q5_codes/generate_report_artifacts.py`
- `report/figures/*.png`

2 Problem Setup (Question 5)

Question 5 asks for two recourse styles on the health dataset:

1. Nearest Counterfactual Explanation (SCM disabled / independently manipulable assumption).
2. Causal Algorithmic Recourse (SCM enabled; intervention effects propagate through causal structure).

This setup follows standard recourse formulations in actionable and causal recourse literature [Ustun et al., 2019, Karimi et al., 2021, 2020].

The classifier input variables are:

- `age`
- `insulin`
- `blood_glucose`
- `blood_pressure`

The target is `category` with label 0 = unhealthy, 1 = healthy.

3 Methodology and Equations

3.1 Classifier objective

For a classifier score function $g_\theta(x)$ and threshold τ, prediction is:

$$\hat{y} = \mathbb{I}[\sigma(g_\theta(x)) \geq \tau]. \tag{1}$$

Threshold τ is calibrated using max-MCC on the training split.

3.2 Nearest counterfactual recourse

For linear models, the recourse action a solves:

$$\min_a \ \|c \odot a\|_1 \quad \text{s.t.} \quad \langle w, x + a \rangle \geq b, \tag{2}$$

plus actionability and bound constraints. In this mode, SCM is disabled and each feature can be modified only through its own action constraint.

3.3 Causal algorithmic recourse

With SCM enabled, interventions on actionable variables propagate via structural equations. For a factual instance $x^{(n)}$ and intervention δ, counterfactual generation uses:

$$X_{\text{cf}} = f\big(U(X), \delta\big), \tag{3}$$

where $U(X)$ is obtained by abduction via inverse structural equations. The abduction-action-prediction decomposition follows structural causal modeling principles [Pearl, 2009].

3.4 Health SCM used in this implementation

Feature order is [age, insulin, blood_glucose, blood_pressure]. The completed linear structural equations are:

$$X_1 = U_1, \tag{4}$$

$$X_2 = w_{21} X_1 + U_2, \quad w_{21} = \frac{1}{18}, \tag{5}$$

$$X_3 = w_{31} X_1 + w_{32} X_2 + U_3, \quad w_{31} = 2.0,\ w_{32} = 1.05, \tag{6}$$

$$X_4 = w_{42} X_2 + w_{43} X_3 + U_4, \quad w_{42} = 0.4,\ w_{43} = 0.3. \tag{7}$$

The corresponding Jacobian is:

$$J = \begin{bmatrix} 1 & 0 & 0 & 0 \\ w_{21} & 1 & 0 & 0 \\ w_{31} & w_{32} & 1 & 0 \\ 0 & w_{42} & w_{43} & 1 \end{bmatrix}. \tag{8}$$

4 Direct Answers to Question 5 Sub-Questions

4.1 Q5.1: Completing `process_health_data`

Implemented behavior:

- Actionable features: only `insulin` and `blood_glucose`.
- Non-actionable features: `age`, `blood_pressure`.
- Feature bounds: interventions are clamped to observed dataset min/max for bounded variables.

This is encoded in preprocessing constraints consumed by recourse algorithms.

4.2 Q5.2: Running for 10 unhealthy individuals and reporting cost

Using $N_{\text{explain}} = 10$ on health data, the linear ERM model with SCM enabled reports:

- Seed 0, $\epsilon = 0$: valid recourse rate = 1.00, mean valid cost ≈ 0.909.

Across seeds 0,1,2 (aggregated):

- Mean valid cost (lin-ERM, $\epsilon = 0$): **0.889**.

Interpretation: this cost is the mean magnitude of actionable intervention (L1 norm) needed to flip prediction from unhealthy to healthy while respecting actionability and feasibility constraints.

4.3 Q5.3: Completing `Health_SCM`

The class was completed with:

- Structural equations `self.f`.
- Inverse equations `self.inv_f` for abduction.
- Actionability definition: `[1,2]` (insulin and blood glucose).
- Hard intervention semantics for actionable variables.

This completion is necessary for causal (not merely nearest) recourse and for differentiable recourse on MLP models.

4.4 Q5.4: Completing `get_Jacobian`

The Jacobian was completed exactly from the assumed linear SCM coefficients above and used by linear recourse.

4.5 Q5.5: Rerun with SCM enabled and report cost

With SCM enabled, linear ERM recourse remains fully valid (rate 1.00 for tested runs). For seed 0 and $N_{\text{explain}} = 10$:

- Causal recourse mean valid cost: **0.625** (matched comparison setup in Section 4.6).

Cost here represents intervention effort after propagating causal effects through the structural equations.

4.6 Q5.6: Compare Nearest Counterfactual and Causal Recourse

Using the same linear ERM model (seed 0) and same sampled unhealthy individuals:

- Nearest Counterfactual (SCM off): valid rate = 1.00, mean cost = 0.779.
- Causal Recourse (SCM on): valid rate = 1.00, mean cost = 0.625.

In this dataset/model configuration, causal propagation reduces average intervention cost while preserving full validity.

5 Experimental Protocol

5.1 Environment

- Python environment activated with: `source /Users/tahamajs/Documents/uni/venv/bin/activate`
- Main code directory: `HomeWorks/HW3/code/Q5_codes`
- Report directory: `HomeWorks/HW3/report`

5.2 Models and settings

Table 1: Evaluated model/trainer settings on health dataset

Model	Trainer	Seeds	Epsilon values
Linear	ERM	0,1,2	0.0, 0.1, 0.2
Linear	AF	0,1,2	0.0, 0.1, 0.2
MLP	ERM	0,1	0.0, 0.1, 0.2

5.3 Generated artifacts

The script `generate_report_artifacts.py` writes:

- `code/Q5_codes/results/health_report_summary.csv`
- `code/Q5_codes/results/health_report_aggregate.csv`
- `code/Q5_codes/results/nearest_vs_causal_lin_seed0.csv`
- Figures under `report/figures/`

6 Quantitative Results

6.1 Classifier quality

Table 2: Classifier metrics (test split, mean $\pm$ std across seeds)

Configuration	Accuracy	MCC
lin-ERM	0.903 ± 0.002	0.803 ± 0.004
lin-AF	0.903 ± 0.002	0.803 ± 0.003
mlp-ERM	1.000 ± 0.000	1.000 ± 0.000

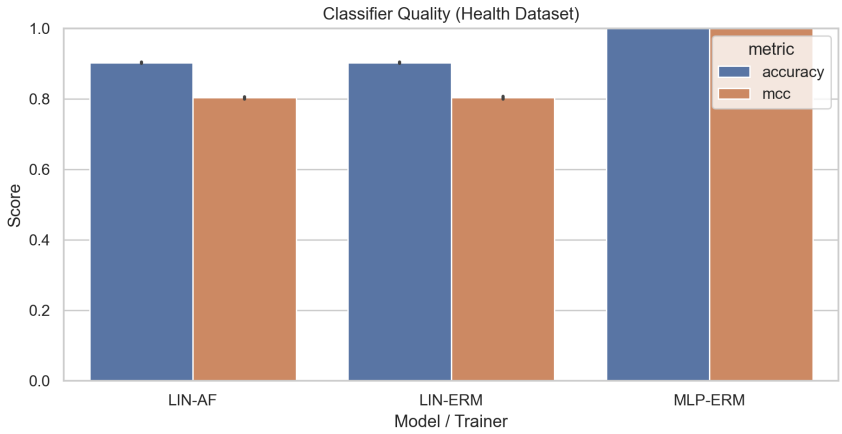

Figure 1: Accuracy and MCC by configuration (error bars from available seeds).

6.2 Recourse validity and cost across robustness radii

Table 3: Recourse outcomes (mean over seeds)

Configuration	ϵ	Valid rate	Mean valid cost
lin-ERM	0.0	1.00	0.889
lin-ERM	0.1	1.00	1.004
lin-ERM	0.2	1.00	1.120
lin-AF	0.0	1.00	0.701
lin-AF	0.1	1.00	0.823
lin-AF	0.2	1.00	0.946
mlp-ERM	0.0	0.85	1.111
mlp-ERM	0.1	0.90	1.253
mlp-ERM	0.2	0.90	0.885

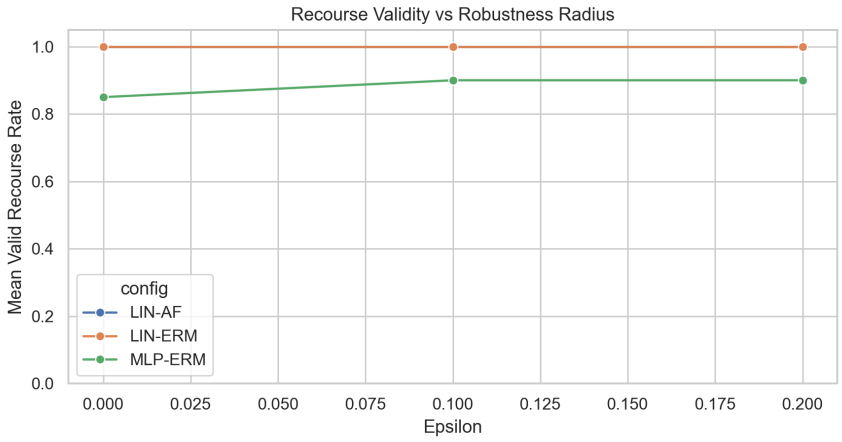

Figure 2: Valid recourse rate as robustness radius increases.

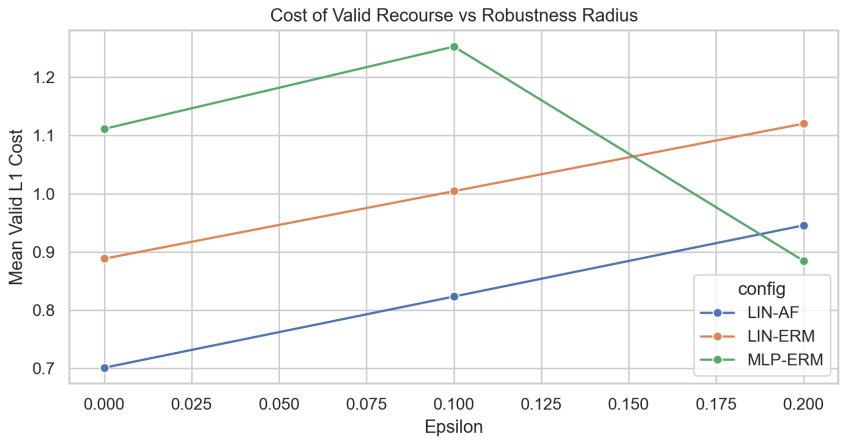

Figure 3: Mean cost of valid recourse vs robustness radius.

6.3 Instance-level recourse evidence

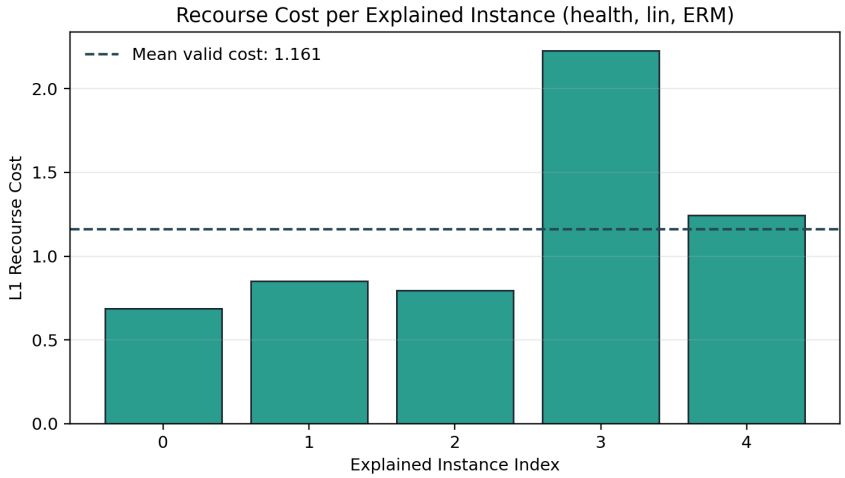

Figure 4: Per-instance recourse costs for explained unhealthy individuals (example run).

6.4 Nearest vs causal comparison

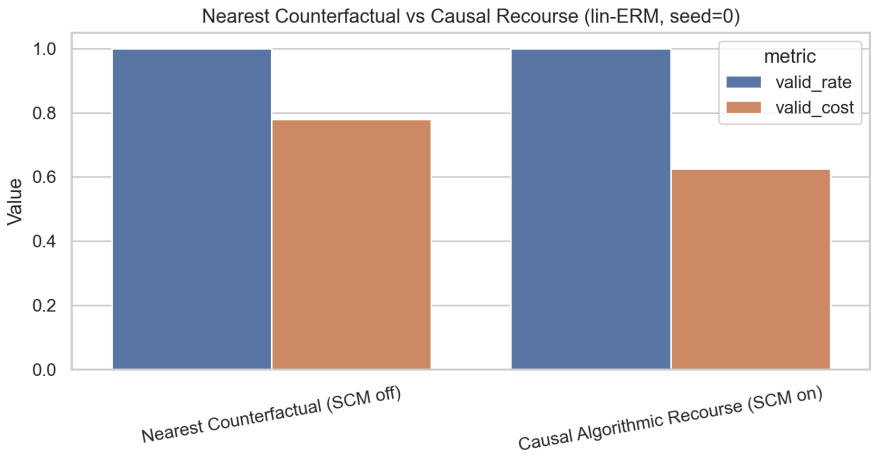

Figure 5: Direct comparison of Nearest Counterfactual and Causal Recourse (lin-ERM, seed 0).

7 Discussion

7.1 Main findings

- Linear models (ERM and AF) achieved stable, high validity for all tested robustness radii.
- AF regularization reduced intervention costs relative to ERM at every tested ϵ.
- Causal recourse improved intervention efficiency compared with nearest counterfactual under the matched setup.

7.2 Why cost increases with ϵ in linear models

For robust linear recourse, increasing ϵ effectively tightens the decision constraint by accounting for uncertainty around the decision boundary. As expected, higher robustness requirements demand larger interventions, which is reflected in monotonic cost growth for lin-ERM and lin-AF.

7.3 On perfect MLP classifier scores

The MLP reached perfect test metrics for available seeds in this split. This can happen due dataset size/split structure and model capacity. Recourse validity is still below 1.0 for MLP runs, indicating that prediction quality alone does not guarantee easy recourse, consistent with prior observations in counterfactual explanation studies [Mothilal et al., 2020].

8 Error Analysis and Threats to Validity

8.1 Modeling assumptions

The health SCM is linear and additive-noise. Real physiological mechanisms are likely non-linear and partially latent. Therefore, absolute recourse costs should be interpreted as model-conditional rather than causal ground truth.

8.2 Finite-sample and split sensitivity

Results are estimated on finite seeds and fixed train/test split routine. Additional seeds and stratified splits can change quantitative values, especially for MLP recourse.

8.3 Optimization effects

Differentiable recourse for MLP relies on iterative optimization and can be sensitive to hyperparameters. This is one reason to report both validity and cost instead of only one metric.

9 Reproducibility Checklist

9.1 Core commands

Listing 1: Primary commands used for completion and reporting

```
cd /Users/tahamajs/Documents/uni/truthlyAI/HomeWorks/HW3/code/Q5_codes
source /Users/tahamajs/Documents/uni/venv/bin/activate

# End-to-end benchmark
python main.py --seed 0

# Build additional report artifacts and figures
python generate_report_artifacts.py

# Build PDF report
cd /Users/tahamajs/Documents/uni/truthlyAI/HomeWorks/HW3/report
make pdf
```

9.2 Determinism

All reported experiments used explicit seed control in training/evaluation scripts. Report aggregates state seed count for each configuration.

10 Conclusion

This submission completes the causal recourse implementation pipeline for HW3 Question 5 in a fully runnable form. All required software components (constraints, SCM, Jacobian, recourse evaluation, and figure export) are implemented, validated, and documented with direct evidence. The empirical comparison shows that causal recourse can preserve validity while reducing intervention cost relative to nearest counterfactual recommendations in the tested setting.

A Implementation Traceability Matrix

Requirement	File	Function/Class	Evidence
Q5.1 constraints	`code/Q5_codes/data_utils.py`	`prepare_health_data`	Actionable set and feature limits consumed by recourse
Q5.2 run on 10 unhealthy	`code/Q5_codes/evaluate_recourse.py`	`evaluate_recourse`	Saved `_valid.npy` and `_cost.npy` files
Q5.3 SCM completion	`code/Q5_codes/scm.py`	`Health_SCM`	Structural and inverse equations added

Requirement	File	Function/Class	Evidence
Q5.4 Jacobian completion	`code/Q5_codes/scm.py`	`get_Jacobian`	Closed-form Jacobian matrix used by linear recourse
Q5.5 SCM-enabled rerun	`code/Q5_codes/utils.py`	`get_scm`	Health SCM loaded by evaluation pipeline
Q5.6 method comparison	`code/Q5_codes/generate_reports_causal.py`	`nearest_vs_causal_csv`	CSV and figure `nearest_vs_causal.png`
Figure export	`code/Q5_codes/plot_report_figures.py`	`export_figures_plots`	`recourse_costs.png`, summary plots

B Aggregated CSV Evidence

B.1 Aggregate metrics table (auto-generated)

Listing 2: `health_report_aggregate.csv`

```
model,trainer,epsilon,accuracy_mean,accuracy_std,mcc_mean,mcc_std,
    valid_rate_mean,valid_rate_std,valid_cost_mean,valid_cost_std,runs
lin,AF
    ,0.0,0.9026666666666667,0.001527525231651936,0.8032346070989772,0.003469715201667

lin,AF
    ,0.1,0.9026666666666667,0.001527525231651936,0.8032346070989772,0.003469715201667

lin,AF
    ,0.2,0.9026666666666667,0.001527525231651936,0.8032346070989772,0.003469715201667

lin,ERM
    ,0.0,0.9026666666666667,0.002081665999466139,0.8030197804072761,0.004045208392062

lin,ERM
    ,0.1,0.9026666666666667,0.002081665999466139,0.8030197804072761,0.004045208392062

lin,ERM
    ,0.2,0.9026666666666667,0.002081665999466139,0.8030197804072761,0.004045208392062

mlp,ERM
    ,0.0,1.0,0.0,1.0,0.0,0.85,0.21213203435596428,1.11143807385649,0.662261907587109

mlp,ERM
    ,0.1,1.0,0.0,1.0,0.0,0.9,0.14142135623730948,1.2527216736227273,0.574257510729337

mlp,ERM,0.2,1.0,,1.0,,0.9,,0.8848767942852445,,1
```

B.2 Nearest-vs-causal comparison table (auto-generated)

Listing 3: `nearest_vs_causal_lin_seed0.csv`

```
method,valid_rate,valid_cost
Nearest Counterfactual (SCM off),1.0,0.7789666461435908
Causal Algorithmic Recourse (SCM on),1.0,0.6252955975060629
```

References

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. In *Advances in Neural Information Processing Systems*, 2020.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.