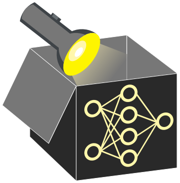
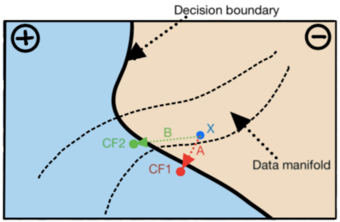
# Interpretable Machine Learning

## Counterfactual Explanations (CEs)
## Optimization Problem and Objectives






**Learning goals**

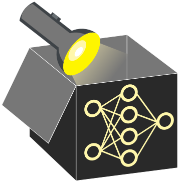
- Formulate CEs as optimization problem
- Identify key objectives (proximity, sparsity)
- Understand trade-offs in CE generation

# MATHEMATICAL PERSPECTIVE

Terminology:

- $\mathbf{x}$: original/factual data point whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired predi. ($y'$ = “grant credit") or interval ($y' = [1000, \infty[$)

# MATHEMATICAL PERSPECTIVE



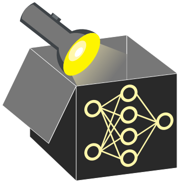
Terminology:

- $\mathbf{x}$: original/factual data point whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired predi. ($y'$ = “grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ satisfies two criteria:

1. **Prediction validity:** CE’s prediction $\hat{f}(\mathbf{x}')$ is equal to the desired pred. $y'$
2. **Proximity:** CE $\mathbf{x}'$ is as close as possible to the original input $\mathbf{x}$

# MATHEMATICAL PERSPECTIVE



Terminology:

- $\mathbf{x}$: original/factual data point whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired predi. ($y'$ = “grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ satisfies two criteria:

1. **Prediction validity:** CE’s prediction $\hat{f}(\mathbf{x}')$ is equal to the desired pred. $y'$
2. **Proximity:** CE $\mathbf{x}'$ is as close as possible to the original input $\mathbf{x}$

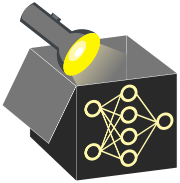
Reformulate these two objectives as optimization problem:

$$\arg\min_{\mathbf{x}'} \lambda_1 o_{target}(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_{proximity}(\mathbf{x}', \mathbf{x})$$

- $\lambda_1$ and $\lambda_2$ balance the two objectives
- $o_{target}$: distance in target space
- $o_{proximity}$: distance in feature space

# OBJECTIVE FUNCTIONS

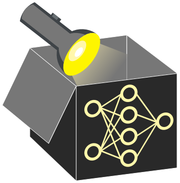
▸ "Dandl et al." 2020



**Distance in target space $o_{target}$:**

- **Regression:** $L_1$ distance $o_{target}(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- **Classification:**
  - For predicted probabilities: $o_{target} = |\hat{f}(\mathbf{x}') - y'|$
  - For predicted hard labels: $o_{target} = \mathbb{I}\{\hat{f}(\mathbf{x}') \neq y'\}$

# OBJECTIVE FUNCTIONS ▸ "Dandl et al." 2020



**Distance in target space $o_{target}$:**

- **Regression:** $L_1$ distance $o_{target}(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- **Classification:**
  - For predicted probabilities: $o_{target} = |\hat{f}(\mathbf{x}') - y'|$
  - For predicted hard labels: $o_{target} = \mathbb{I}\{\hat{f}(\mathbf{x}') \neq y'\}$

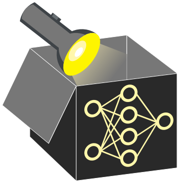
**Distance in input space $o_{proximity}$: Gower distance (mixed feature types)**

$$o_{proximity}(\mathbf{x}', \mathbf{x}) = d_G(\mathbf{x}', \mathbf{x}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x_j', x_j) \in [0, 1], \text{ where}$$

- $\delta_G(x_j', x_j) = \mathbb{I}\left\{x_j' \neq x_j\right\}$ if $x_j$ is categorical
- $\delta_G(x_j', x_j) = \frac{1}{\widehat{R}_j} |x_j' - x_j|$ if $x_j$ is numerical

  $\leadsto$ $\widehat{R}_j$: range of feature $j$ in the training set to ensure $\delta_G(x_j', x_j) \in [0, 1]$

# FURTHER OBJECTIVES: SPARSITY

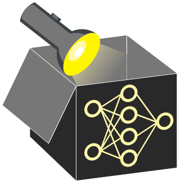


Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include **sparsity** and **plausibility**

**Sparsity** Favor explanations that change few features

- End-users often prefer short over long explanations

# FURTHER OBJECTIVES: SPARSITY

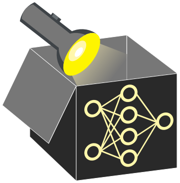


Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include **sparsity** and **plausibility**

**Sparsity** Favor explanations that change few features

- End-users often prefer short over long explanations
- Sparsity could be integrated into $o_{proximity}$
  e.g., using $L_0$-norm (number of changed features) or $L_1$-norm (LASSO)

# FURTHER OBJECTIVES: SPARSITY



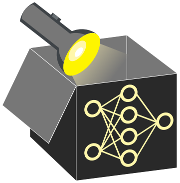
Additional constraints can improve the explanation quality of the corresponding CEs
$\rightsquigarrow$ popular constraints include **sparsity** and **plausibility**

**Sparsity** Favor explanations that change few features

- End-users often prefer short over long explanations
- Sparsity could be integrated into $o_{proximity}$
  e.g., using $L_0$-norm (number of changed features) or $L_1$-norm (LASSO)
- Alternative: Include separate objective measuring sparsity, e.g., via $L_0$-norm

$$o_{sparse}(\mathbf{x}', \mathbf{x}) = \sum_{j=1}^{p} \mathcal{I}_{\{x'_j \neq x_j\}}$$

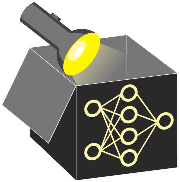
# FURTHER OBJECTIVES: PLAUSIBILITY



**Plausibility:**

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed

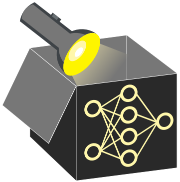
# FURTHER OBJECTIVES: PLAUSIBILITY



**Plausibility:**

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ Avoid unrealistic combinations of feature values

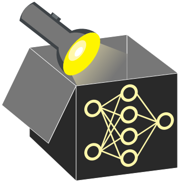
# FURTHER OBJECTIVES: PLAUSIBILITY



**Plausibility:**

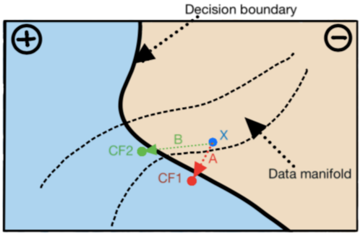
- CEs should suggest realistic (i.e., plausible) alternatives
  $\leadsto$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\leadsto$ Avoid unrealistic combinations of feature values
- Estimating joint distribution is hard, especially for mixed feature spaces
  $\leadsto$ Common proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$

# FURTHER OBJECTIVES: PLAUSIBILITY



**Plausibility:**

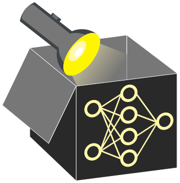
- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ Avoid unrealistic combinations of feature values
- Estimating joint distribution is hard, especially for mixed feature spaces
  $\rightsquigarrow$ Common proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$




**Example from** ▸ "Verma et al." 2020

- Input $\mathbf{x}$ originally classified as $\ominus$
- Two valid CEs in class $\oplus$: CF1 and CF2
- Path **A** (CF1) is shorter (but unrealistic)
- Path **B** (CF2) is longer but in data manifold

# FURTHER OBJECTIVES

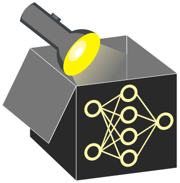


**Plausibility term:** Encourage counterfactuals close to observed data.

- Define $\mathbf{x}^{[1]}$ as the nearest neighbor of $\mathbf{x}'$ in the training set $\mathbf{X}$
- Use Gower distance between $\mathbf{x}'$ and $\mathbf{x}^{[1]}$ to define plausibility objective:

$$o_{plausible}(\mathbf{x}', \mathbf{X}) = d_G(\mathbf{x}', \mathbf{x}^{[1]}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x'_j, x^{[1]}_j)$$

# FURTHER OBJECTIVES



**Plausibility term:** Encourage counterfactuals close to observed data.

- Define $\mathbf{x}^{[1]}$ as the nearest neighbor of $\mathbf{x}'$ in the training set $\mathbf{X}$
- Use Gower distance between $\mathbf{x}'$ and $\mathbf{x}^{[1]}$ to define plausibility objective:

$$o_{plausible}(\mathbf{x}', \mathbf{X}) = d_G(\mathbf{x}', \mathbf{x}^{[1]}) = \frac{1}{p}\sum_{j=1}^{p} \delta_G(x'_j, x_j^{[1]})$$

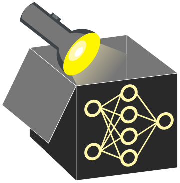
**Extended optimization:** Add sparsity and plausibility terms to the objective

$$\arg\min_{\mathbf{x}'} \lambda_1 o_{\text{target}}(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_{\text{proximity}}(\mathbf{x}', \mathbf{x}) + \lambda_3 o_{\text{sparse}}(\mathbf{x}', \mathbf{x}) + \lambda_4 o_{\text{plausible}}(\mathbf{x}', \mathbf{X})$$

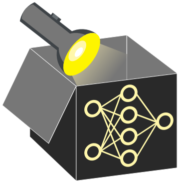
# REMARKS: THE RASHOMON EFFECT

**Issue (Rashomon effect):**

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist



# REMARKS: THE RASHOMON EFFECT



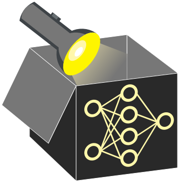
**Issue (Rashomon effect):**

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

**Possible solutions:**

- Present all CEs for $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one/few CEs (but: by which criterion should guide this choice?)

# REMARKS: THE RASHOMON EFFECT



**Issue (Rashomon effect):**

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

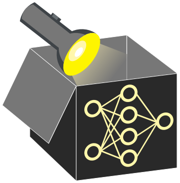
**Possible solutions:**

- Present all CEs for $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one/few CEs (but: by which criterion should guide this choice?)

**Note:**

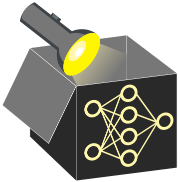
- Nonlinear models can produce diverse and inconsistent CEs
  $\rightsquigarrow$ suggest both increasing and decreasing credit duration
  (confusing for users)
- Handling this **Rashomon effect** remains an open problem in interpretable ML

# REMARKS: MODEL OR REAL-WORLD



- CEs explain model predictions, but may seem to explain real-world users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- **Example:** CE suggests increasing age by 5 to receive a loan
  $\rightsquigarrow$ The applicant waits 5 years and reapplies

# REMARKS: MODEL OR REAL-WORLD



- CEs explain model predictions, but may seem to explain real-world users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- **Example:** CE suggests increasing age by 5 to receive a loan
  $\rightsquigarrow$ The applicant waits 5 years and reapplies
- **Problem:** Other features may change in the meantime (e.g., job status, income) $\rightsquigarrow$ ▸ "Karimi et al." 2020 propose CEs that respect causal structure
- **Model drift:** Bank's algorithm itself may change over time
  $\rightsquigarrow$ Past CEs may become invalid