Appendices

Acknowledgements

We would like to thank Hamed Jalali, Charlie Marx and the anonymous reviewers for insightful comments and suggestions.

Appendix A Further experimental evaluations

On semantics. Consider table 2. While GS' recommendations tend to have low costs, they often take on ambiguous values. We have marked the critical values in red. AR's recommendations tend to make sense, if one inspects them input value by input value. However, often they run into logical inconsistencies, which we highlighted in blue in table 2. So does it make sense to tell someone to be more often on time for the 30-59 days range while demanding that the person should be paying more often 60-89 days late? Probably not. The *data support* counterfactual recommendations from OURS, in contrast, appear to make sense and seem consistent. There exist multiple of these examples in the generated explanations and many of them follow a similar pattern. *Most importantly, this demonstrates that end users could find it troublesome to comprehend what causes a classifier to behave a certain way, versus what causes the world to behave in a certain way.*

Model	set of mutable inputs							
	rev.util.	#30-59 d. late	debt ratio	income	# credit	$>$ 90 d. late	# r. est. loans	# 60-89 d. l.
$x \in H_f^-$	1.00	3.00	0.19	2700.00	3.00	4.00	0.00	0.00
GS	1.12	2.77	0.24	2699.92	3.03	4.08	-0.13	0.25
AR	1.00	2.00	0.19	2700.00	3.00	4.00	0.00	2.00
OURS	0.97	0.00	0.18	2753.82	3.00	0.00	0.00	0.00

Table 2: **Illustrative example, comparing semantics of recommendations from GS, OURS and AR**. The instance $x \sim p_{data}$ was negatively classified by the prediction model f. For this individual, the immutable inputs are fixed at age = 36 and # dependents = 3. red: ambiguous values. blue: inconsistent values.

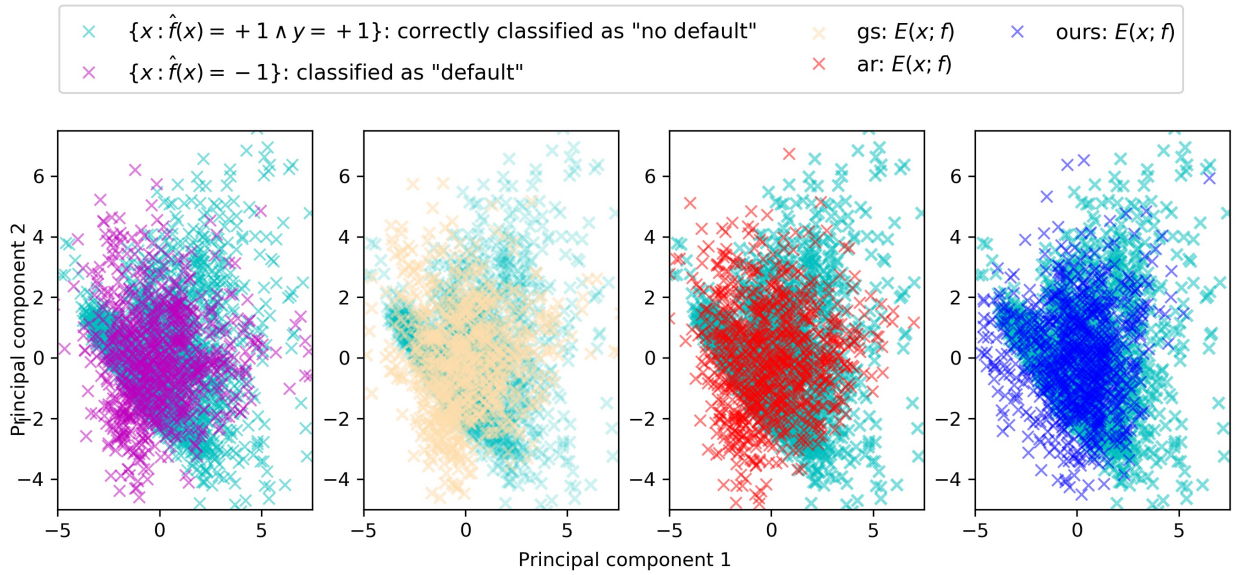
More on robustness. Recall that we wish to find recommendations for the negative predicted individuals, H_f^-. As opposed to the other methods, the OURS method *pushes the negative predicted individuals towards data points from the correctly classified individuals*, $H_f^+ \cap D^+$. To show this for all explanations, we compute the first two principal components of $E_{ar}(\boldsymbol{x}; f)$, $E_{gs}(\boldsymbol{x}; f)$ and $E_{ours}(\boldsymbol{x}; f)$ and compare them to H_f^-, $H_f^+ \cap D^+$ (see figure 6).

Appendix B Data and Implementations

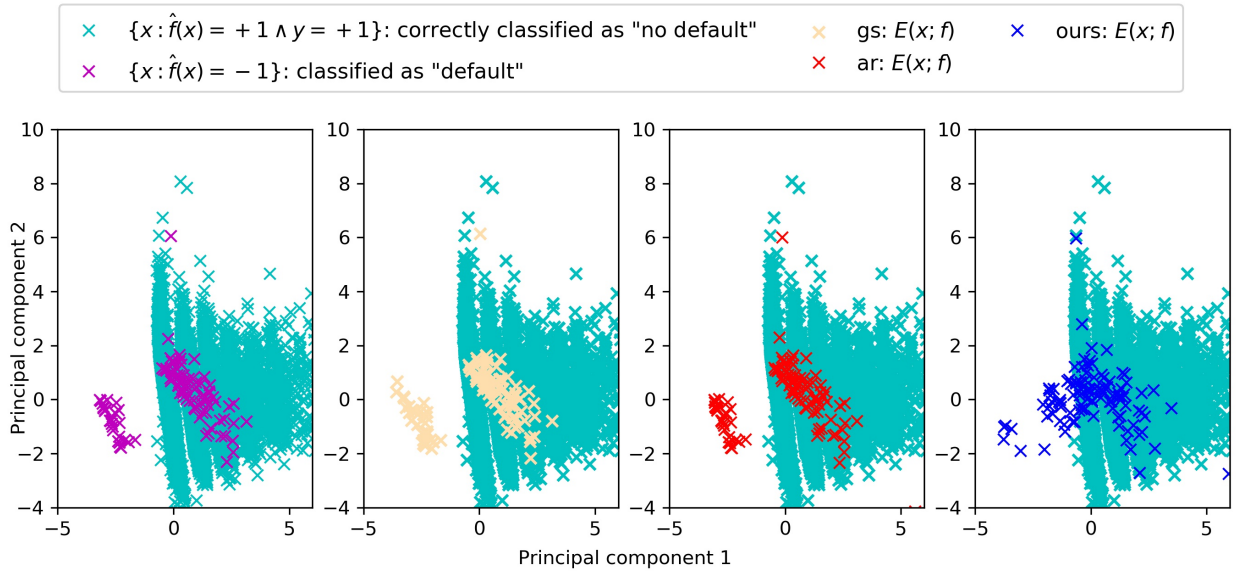
B.1 Real world example: "`Give Me Some Credit`"

In the following, we list the specified pretrained classification models as well as the parameter specification used for the experiments. We use 80 percent of the data as our training set and the remaining part is used as the holdout test set. Additionally, we allow f and g access to all features, i.e.. to the mutable and immutable ones. The state of features can be found in table 3.

AR (Ustun et al., 2019). The AR algorithm requires to choose both an action set and free and immutable features. The implementation can be found here: `https://github.com/ustunb/actionable-recourse`. We specify that the *DebtRatio* feature can only move downward (Ustun et al., 2019). The AR implementation has a default decision boundary at 0 and therefore one needs to shift the boundary. We choose $p_{AR} = 0.50$, adjusting the boundary appropriately. Finally, we set the linear programming optimizer to *cbc*, which is based on an open-access `python` implementation.



(a) HELOC



(b) Give Me Some Credit

Figure 6: **First and second principal components** of $H_f^+ \cap D^+$ (cyan), H^- (magenta) and counterfactual recommendations $E(\boldsymbol{x}; f)$, where f denotes the pretrained regularized linear regression classifier (in this case). Recall that $f(E(\boldsymbol{x}; f)) = +1$. AR's (red) and GS' (yellow) latent space representation of the generated counterfactual recommendations remain very close to the incorrectly classified representation (purple). OURS (blue, right most) rotates and pushes the latent space closer to the one of the correctly classified observations $H_f^+ \cap D^+$ (cyan).

GS (Laugel et al., 2017). GS is based on a version of the YPHL algorithm. As such we have to choose appropriate step sizes in our implementation to generate new observations from the sphere around $\boldsymbol{x}$. We choose a step size of 0.1.

OURS (Pawelczyk et al., 2020). We used the (H)VAE implemention as described here: `https://github.com/probabilistic-learning/HI-VAE` (Nazabal et al., 2018). Random search in the latent space was conducted to find counterfactual recommendations, using the YPHL algorithm (Laugel et al., 2017). We made the following choices. We set the latent space dimension of both $\boldsymbol{s}$ and $\boldsymbol{z}$ to 3 and 6, respectively. For training, we used 15 epochs.

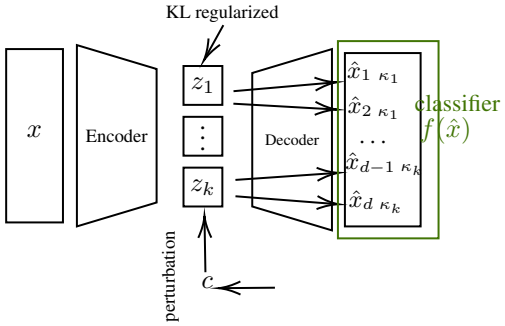


Figure 7: Schematic figure for counterfactual search from the OURS model (Pawelczyk et al., 2020). The latent representation ideally learns independent concepts denoted by $\kappa_1, ..., \kappa_k$ (e.g. timeliness, overall financial situation, etc.).

Table 3 gives details about the chosen likelihood model for each input. For count inputs, we use the Poisson likelihood model, while for inputs with a support on the positive part of the real line we choose log normal distributions.

Inputs	Mutable	Model
Revolving Utilization Of Unsecured Lines	Y	log Normal
Age	N	Poisson
Number Of Times 30-59 Days Past Due Not Worse	Y	Poisson
Debt Ratio	Y	log Normal
Monthly Income	Y	log Normal
Number Open Credit Lines And Loans	Y	Poisson
Number Of Times 90 days Late	Y	Poisson
Number Real Estate Loans Or Lines	Y	Poisson
Number Of Times 60-89 Days Past Due Not Worse	Y	Poisson
Number Of Dependents	N	Poisson

Table 3: "`Give Me Some Credit`": State of inputs and likelihood models.

B.2 Real world example: `HELOC`

The *Home Equity Line of Credit (HELOC)* data set consists of credit applications made by homeowners in the US, which can be obtained from the FICO community.[1] The task is to use the applicant's information within the credit report to predict whether they will repay the HELOC account within 2 years. Table 4 gives an overview of the available inputs and the corresponding assumed likelihood models.

AR and GS As before. Additionally, we do not specify how features have to move.

OURS We set the latent space dimension of both $\boldsymbol{s}$ and $\boldsymbol{z}$ to 12 and 10, respectively. For training, we used 60 epochs. Table 4 gives details about the chosen likelihood model for each feature. The rest remains as before.

Appendix C Proof of proposition 1

Proof. Let us consider an $\boldsymbol{x}_1 \in H_f^+$, i.e. $f(\boldsymbol{x}_1) = +1 = f(h(\tilde{\boldsymbol{z}}))$. By the assumption of the generative model in the main text, we know that $h(\boldsymbol{z}) = \boldsymbol{x}$. We have $\boldsymbol{c}_D(\tilde{z}) = \|\boldsymbol{x} - h(\tilde{\boldsymbol{z}})\| = \|(\boldsymbol{x} - \boldsymbol{x}_1) + (\boldsymbol{x}_1 - h(\tilde{\boldsymbol{z}}))\| \leq \|\boldsymbol{x} - \boldsymbol{x}_1\| + \|\boldsymbol{x}_1 - h(\tilde{\boldsymbol{z}})\|$, where we used the triangle inequality. By (1), we have that $\|\boldsymbol{x}_1 - h(\tilde{\boldsymbol{z}})\| \leq \|\boldsymbol{x}_1 - h(\boldsymbol{z})\|$. Hence, we can write

[1] https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2.

Input	Mutable	Model
MSinceOldestTradeOpen	N	Poisson
AverageMInFile	N	Poisson
NumSatisfactoryTrades	Y	Poisson
NumTrades60Ever/DerogPubRec	Y	Poisson
NumTrades90Ever/DerogPubRec	Y	Poisson
NumTotalTrades	Y	Poisson
PercentInstallTrades	Y	Poisson
MSinceMostRecentInqexcl7days	Y	Poisson
NumInqLast6M	Y	Poisson
NetFractionRevolvingBurden	Y	Poisson
NumRevolvingTradesWBalance	Y	Poisson
NumBank/NatlTradesWHighUtilization	Y	Poisson
ExternalRiskEstimate	N	Poisson
MPercentTradesNeverDelq	Y	Poisson
MaxDelq2PublicRecLast12M	Y	Poisson
MaxDelqEver	Y	Poisson
NumTradesOpeninLast12M	Y	Poisson
NumInqLast6Mexcl7days	Y	Poisson
NetFractionRevolvingBurden	Y	Poisson
NumInstallTradesWBalance	Y	Poisson
NumBank2NatlTradesWHighUtilization	Y	Poisson
PercentTradesWBalance	Y	Poisson

Table 4: HELOC: State of inputs and likelihood models.

$\boldsymbol{c}_D(\tilde{z}) \leq 2\|\boldsymbol{x}_1 - h(\boldsymbol{z})\| = 2\|\boldsymbol{x}_1 - \boldsymbol{x}\|$. Now, minimizing $\boldsymbol{x}$ over $\mathcal{E}_S$ (recall definition 1 from the main text) gives the desired result. □

Appendix D Proof of proposition 2

From now on, we suppress the dependence of $f(x) := f$ and $g(x) := g$ on $\boldsymbol{x} := x$. For brevity, we sometimes say $A := H_f^- \cup H_g^-$ and $\pi_- = 1 - \pi$. $\pi = Pr_{H_f^- \cap H_g^-}(y = 1)$; $\pi_f = Pr_{H_f^-}(y = 1)$; $\pi_g = Pr_{H_g^-}(y = 1)$

D.1 Main argument

Proof. We first expand the $\overline{cost}(f, g)_{H_f^- \cup H_g^-}$.

$$
\begin{aligned}
\overline{cost}(f,g)_{H_f^- \cup H_g^-} &= \mathbb{E}_{H_f^- \cup H_g^-}[c^*(f,g,x)] \\
&= \pi \cdot [\mathbb{E}_{A\cap D^+}[c^*|f \leq 0, g \leq 0]P_{A\cap D^+}(f \leq 0, g \leq 0) + \mathbb{E}_{A\cap D^+}[c^*|f \leq 0, g > 0]P_{A\cap D^+}(f \leq 0, g > 0) \\
&+ \mathbb{E}_{A\cap D^+}[c^*|f > 0, g > 0]P_{A\cap D^+}(f > 0, g > 0) + \mathbb{E}_{A\cap D^+}[c^*|f > 0, g \leq 0]P_{A\cap D^+}(f > 0, g \leq 0)] \\
&+ \pi_-[\mathbb{E}_{A\cap D^-}[c^*|f \leq 0, g \leq 0]P_{A\cap D^-}(f \leq 0, g \leq 0) + \mathbb{E}_{A\cap D^-}[c^*|f \leq 0, g > 0]P_{A\cap D^-}(f \leq 0, g > 0) \\
&+ \mathbb{E}_{A\cap D^-}[c^*|f > 0, g > 0]P_{A\cap D^-}(f > 0, g > 0) + \mathbb{E}_{A\cap D^-}[c^*|f > 0, g \leq 0]P_{A\cap D^-}(f > 0, g \leq 0)]
\end{aligned}
$$

Moreover, note that $|\mathbb{E}_{H_f^- \cap D^+}[f|f \leq 0]| \leq c^{max}_{H_f^-}(f)$, and $|\mathbb{E}_{H_f^- \cap D^-}[f|f > 0]| \leq c^{max}_{H_f^-}(f)$. Analogously for the classifier g. And hence we can write for the classifier f:

$$\begin{aligned}
-\pi_f P_{H_f^- \cap D^+}(f \leq 0)\mathbb{E}_{H_f^- \cap D^+}[f|f \leq 0] &= 2\pi_f P_{H_f^- \cap D^+}(f \leq 0)|\mathbb{E}_{H_f^- \cap D^+}[f|f \leq 0]| \\
&\quad + \pi_f P_{H_f^- \cap D^+}(f \leq 0)\mathbb{E}_{H_f^- \cap D^+}[f|f \leq 0], \\
&\leq 2P_{H_f^- \cap D^+}(f \leq 0)\pi_f c^{max}_{H_f^-}(f) + \pi_f P_{H_f^- \cap D^+}(f \leq 0)\mathbb{E}_{H_f^- \cap D^+}[f|f \leq 0];
\end{aligned} \tag{6}$$

$$\begin{aligned}
\pi_f P_{H_f^- \cap D^-}(f > 0)\mathbb{E}_{H_f^- \cap D^-}[f|f > 0] &= 2\pi_f P_{H_f^- \cap D^-}(f > 0)|\mathbb{E}_{H_f^- \cap D^-}[f|f > 0]| \\
&\quad - \pi_f P_{H_f^- \cap D^-}(f > 0)\mathbb{E}_{H_f^- \cap D^-}[f|f > 0]. \\
&\leq 2P_{H_f^- \cap D^-}(f > 0)\pi c^{max}_{H_f^-}(f) - \pi_f P_{H_f^- \cap D^-}(f > 0)\mathbb{E}_{H_f^- \cap D^-}[f|f > 0].
\end{aligned} \tag{7}$$

By assumption 1 from the main text, fact 1 and fact 3 we have:

$$\begin{aligned}
&\mathbb{E}_{A \cap D^{+/-}}[c^*|f \leq 0, g \leq 0] \leq \alpha \mathbb{E}_{A \cap D^{+/-}}[\max(-f, -g)^\gamma | f \leq 0, g \leq 0] \\
&= \left(\frac{\alpha}{2^\gamma}\right)(\mathbb{E}_{A \cap D^{+/-}}[(-f - g + |-f + g|)^\gamma | f \leq 0, g \leq 0]) \leq \left(\frac{\alpha}{2^\gamma}\right)(\mathbb{E}_{A \cap D^{+/-}}[(-f - g + |-f + g|)|f \leq 0, g \leq 0])^\gamma.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{A \cap D^{+/-}}[c^*|f \leq 0, g > 0] \leq \alpha \mathbb{E}_{A \cap D^{+/-}}[\max(-f, +g)^\gamma | f \leq 0, g > 0] \\
&= \left(\frac{\alpha}{2^\gamma}\right)(\mathbb{E}_{A \cap D^{+/-}}[(-f + g + |-f - g|)^\gamma | f \leq 0, g > 0]) \leq \left(\frac{\alpha}{2^\gamma}\right)(\mathbb{E}_{A \cap D^{+/-}}[(-f + g + |-f - g|)|f \leq 0, g > 0])^\gamma.
\end{aligned}$$

Analogously for the remaining 2 terms. Next, note that

$$\mathbb{E}_{H_f^- \cup H_g^- \cap D^{+/-}}[f|f \leq 0, g \leq 0] \leq \mathbb{E}_{H_f^- \cup H_g^- \cap D^{+/-}}[f|f \leq 0] = \mathbb{E}_{H_f^- \cap D^{+/-}}[f|f \leq 0], \tag{8}$$

$$\mathbb{E}_{H_f^- \cup H_g^- \cap D^{+/-}}[g|f \leq 0, g \leq 0] \leq \mathbb{E}_{H_f^- \cup H_g^- \cap D^{+/-}}[g|g \leq 0] = \mathbb{E}_{H_g^- \cap D^{+/-}}[g|g \leq 0], \tag{9}$$

where the first equality follows by assuming that f and g do not assign widely different predictions to the same input and the second equality follows since H_g^- is not restricting $H_f^- \cup H_g^-$ and vice versa. Similarly, for the remaining terms. Now, we go back to the expanded cost, use linearity of expectations, (8), (9), facts 2 (second inequality) and 4 (first inequality) and upper bound it by:

$$\begin{aligned}
&\overline{cost}(f, g)_{H_f^- \cup H_g^-} \leq \pi^\gamma \left(\frac{\alpha}{2^\gamma}\right) \Big[\mathbb{E}_{A \cap D^+}[(-f - g + |-f + g||f \leq 0, g \leq 0]^\gamma P_{A \cap D^+}(f \leq 0, g \leq 0)^\gamma \\
&+ \mathbb{E}_{A \cap D^+}[(-f + g + |-f - g||f \leq 0, g \leq 0]^\gamma P_{A \cap D^+}(f \leq 0, g > 0)^\gamma \\
&+ \mathbb{E}_{A \cap D^+}[(f + g + |f - g||f \leq 0, g \leq 0]^\gamma P_{A \cap D^+}(f > 0, g > 0)^\gamma \\
&+ \mathbb{E}_{A \cap D^+}[(f - g + |f + g||f \leq 0, g \leq 0]^\gamma P_{A \cap D^+}(f > 0, g \leq 0)^\gamma \Big] \\
&+ \pi_-^\gamma \left(\frac{\alpha}{2^\gamma}\right) \Big[\mathbb{E}_{A \cap D^-}[(-f - g + |-f + g||f \leq 0, g \leq 0])^\gamma P_{A \cap D^-}(f \leq 0, g \leq 0)^\gamma \\
&+ \mathbb{E}_{A \cap D^-}[(-f + g + |-f - g||f \leq 0, g \leq 0]^\gamma P_{A \cap D^-}(f \leq 0, g > 0)^\gamma \\
&+ \mathbb{E}_{A \cap D^-}[(f + g + |f - g||f \leq 0, g \leq 0]^\gamma P_{A \cap D^-}(f > 0, g > 0)^\gamma \\
&+ \mathbb{E}_{A \cap D^-}[(f - g + |f + g||f \leq 0, g \leq 0]^\gamma P_{A \cap D^-}(f > 0, g \leq 0)^\gamma \Big]
\end{aligned}$$

$$
\begin{aligned}
&\leq \pi^{\gamma}\left(\frac{\alpha}{2^{\gamma}}\right)\Bigg[\Bigg(\mathbb{E}_{H_f^-\cap D^+}[-f|f\leq 0]P_{H_f^-\cap D^+}(f\leq 0)\mathbb{E}_{H_g^-\cap D^+}[-g|g\leq 0]P_{H_g^-\cap D^+}(g\leq 0)]\\
&+\mathbb{E}_{A\cap D^+}[|-f+g||f\leq 0,g\leq 0]P_{A\cap D^+}(f\leq 0,g\leq 0)\Bigg)^{\gamma}\\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^+}[-f|f\leq 0]P_{H_f^-\cap D^+}(f\leq 0)+\mathbb{E}_{H_g^-\cap D^+}[g|g>0]P_{H_g^-\cap D^+}(g>0)\\
&+\mathbb{E}_{A\cap D^+}[|-f-g||f\leq 0,g>0]P_{A\cap D^+}(f\leq 0,g>0)\Bigg)^{\gamma}\\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^+}[f|f>0]P_{H_f^-\cap D^+}(f>0)+\mathbb{E}_{H_g^-\cap D^+}[g|g>0]P_{H_g^-\cap D^+}(g>0)\\
&+\mathbb{E}_{H_g^-\cap D^+}[|f-g||f>0,g>0]P_{A\cap D^+}(f>0,g>0)\Bigg)^{\gamma}\\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^+}[f|f>0]P_{H_f^-\cap D^+}(f>0)+\mathbb{E}_{H_g^-\cap D^+}[-g|g\leq 0]P_{H_g^-\cap D^+}(g\leq 0)\\
&+\mathbb{E}_{A\cap D^+}[|f+g||f>0,g\leq 0]P_{A\cap D^+}(f>0,g\leq 0)\Bigg)^{\gamma}\Bigg]\\
&+\pi_-^{\gamma}\left(\frac{\alpha}{2^{\gamma}}\right)\Bigg[\Bigg(\mathbb{E}_{H_f^-\cap D^+}[-f|f\leq 0]P_{H_f^-\cap D^+}(f\leq 0)+\mathbb{E}_{H_g^-\cap D^+}[-g|g\leq 0]P_{H_g^-\cap D^+}(g\leq 0)\\
&+\mathbb{E}_{A\cap D^+}[|-f+g||f\leq 0,g\leq 0]P_{A\cap D^+}(f\leq 0,g\leq 0)\Bigg)^{\gamma}\\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^-}[-f|f\leq 0]P_{H_f^-\cap D^-}(f\leq 0)+\mathbb{E}_{H_g^-\cap D^-}[g|g>0]P_{H_g^-\cap D^+}(g>0)\\
&+\mathbb{E}_{A\cap D^-}[|-f-g||f\leq 0,g>0]P_{A\cap D^+}(f\leq 0,g>0)\Bigg)^{\gamma}\\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^-}[f|f>0]P_{H_f^-\cap D^-}(f>0)+\mathbb{E}_{H_g^-\cap D^-}[g|g>0]P_{H_g^-\cap D^-}(g>0)\\
&+\mathbb{E}_{H_g^-\cap D^-}[|f-g||f>0,g>0]P_{A\cap D^-}(f>0,g>0)\Bigg)^{\gamma}\\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^-}[f|f>0]P_{H_f^-\cap D^-}(f>0)+\mathbb{E}_{H_g^-\cap D^-}[-g|g\leq 0]P_{H_g^-\cap D^-}(g\leq 0)\\
&+\mathbb{E}_{A\cap D^-}[|f+g||f>0,g\leq 0]P_{A\cap D^-}(f>0,g\leq 0)\Bigg)^{\gamma}\Bigg]
\end{aligned}
$$

Next we apply lemma 2 with $n = 8$, use (6) and (7) for the last equality and obtain:

$$
\begin{aligned}
&\leq 8^{1-\gamma}\left(\frac{\alpha}{2^\gamma}\right)\Bigg[\pi_+\Bigg(\Bigg(\mathbb{E}_{H_f^-\cap D^+}[-f|f\leq 0]P_{H_f^-\cap D^+}(f\leq 0)+\mathbb{E}_{H_g^-\cap D^+}[-g|g\leq 0]P_{H_g^-\cap D^+}(g\leq 0)] \\
&+\mathbb{E}_{A\cap D^+}[|-f+g||f\leq 0,g\leq 0]P_{A\cap D^+}(f\leq 0,g\leq 0)\Bigg) \\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^+}[-f|f\leq 0]P_{H_f^-\cap D^+}(f\leq 0)+\mathbb{E}_{H_g^-\cap D^+}[g|g>0]P_{H_g^-\cap D^+}(g>0) \\
&+\mathbb{E}_{A\cap D^+}[|-f-g||f\leq 0,g>0]P_{A\cap D^+}(f\leq 0,g>0)\Bigg) \\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^+}[f|f>0]P_{H_f^-\cap D^+}(f>0)+\mathbb{E}_{H_g^-\cap D^+}[g|g>0]P_{H_g^-\cap D^+}(g>0) \\
&+\mathbb{E}_{H_g^-\cap D^+}[|f-g||f>0,g>0]P_{A\cap D^+}(f>0,g>0)\Bigg) \\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^+}[f|f>0]P_{H_f^-\cap D^+}(f>0)+\mathbb{E}_{H_g^-\cap D^+}[-g|g\leq 0]P_{H_g^-\cap D^+}(g\leq 0) \\
&+\mathbb{E}_{A\cap D^+}[|f+g||f>0,g\leq 0]P_{A\cap D^+}(f>0,g\leq 0)\Bigg)\Bigg) \\
&+\pi_-\Bigg(\Bigg(\mathbb{E}_{H_f^-\cap D^+}[-f|f\leq 0]P_{H_f^-\cap D^+}(f\leq 0)+\mathbb{E}_{H_g^-\cap D^+}[-g|g\leq 0]P_{H_g^-\cap D^+}(g\leq 0)] \\
&+\mathbb{E}_{A\cap D^+}[|-f+g||f\leq 0,g\leq 0]P_{A\cap D^+}(f\leq 0,g\leq 0)\Bigg) \\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^-}[-f|f\leq 0]P_{H_f^-\cap D^-}(f\leq 0)+\mathbb{E}_{H_g^-\cap D^-}[g|g>0]P_{H_g^-\cap D^+}(g>0) \\
&+\mathbb{E}_{A\cap D^-}[|-f-g||f\leq 0,g>0]P_{A\cap D^+}(f\leq 0,g>0)\Bigg) \\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^-}[f|f>0]P_{H_f^-\cap D^-}(f>0)+\mathbb{E}_{H_g^-\cap D^-}[g|g>0]P_{H_g^-\cap D^-}(g>0) \\
&+\mathbb{E}_{H_g^-\cap D^-}[|f-g||f>0,g>0]P_{A\cap D^-}(f>0,g>0)\Bigg) \\
&+\Bigg(\mathbb{E}_{H_f^-\cap D^-}[f|f>0]P_{H_f^-\cap D^-}(f>0)+\mathbb{E}_{H_g^-\cap D^-}[-g|g\leq 0]P_{H_g^-\cap D^-}(g\leq 0) \\
&+\mathbb{E}_{A\cap D^-}[|f+g||f>0,g\leq 0]P_{A\cap D^-}(f>0,g\leq 0)\Bigg)\Bigg]^\gamma \\
&= 8^{1-\gamma}\left(\frac{\alpha}{2^\gamma}\right)\Bigg[\pi_+\Bigg(2\mathbb{E}_{H_f^-\cap D^+}[-f|f\leq 0]P_{H_f^-\cap D^+}(f\leq 0)+2\mathbb{E}_{H_g^-\cap D^+}[-g|g\leq 0]P_{H_g^-\cap D^+}(g\leq 0)] \\
&+2\mathbb{E}_{H_f^-\cap D^+}[f|f>0]P_{H_f^-\cap D^+}(f>0)+2\mathbb{E}_{H_g^-\cap D^+}[g|g>0]P_{H_g^-\cap D^+}(g>0)\Bigg) \\
&+\pi_-\Bigg(2\mathbb{E}_{H_f^-\cap D^-}[-f|f\leq 0]P_{H_f^-\cap D^-}(f\leq 0)+2\mathbb{E}_{H_g^-\cap D^-}[-g|g\leq 0]P_{H_g^-\cap D^-}(g\leq 0)] \\
&+2\mathbb{E}_{H_f^-\cap D^-}[f|f>0]P_{H_f^-\cap D^-}(f>0)+2\mathbb{E}_{H_g^-\cap D^-}[g|g>0]P_{H_g^-\cap D^-}(g>0)\Bigg)+\mathbb{E}_A[|f-g|]\Bigg]^\gamma \\
&= \alpha 8^{1-\gamma}\Bigg[2\Bigg(\frac{2[R_{H_f^-}(f)c^{max}_{H_f^-}(f)+R_{H_g^-}(g)c^{max}_{H_g^-}(g)]}{2} \\
&+\frac{\pi_f\mathbb{E}_{H_f^-\cap D^+}[f]+\pi_g\mathbb{E}_{H_g^-\cap D^+}[g]}{2}-\frac{(1-\pi_f)\mathbb{E}_{H_f^-\cap D^-}[f]+(1-\pi_g)\mathbb{E}_{H_g^-\cap D^-}[g]}{2}\Bigg)+\mathbb{E}_{H_f^-\cup H_g^-}[|f-g|]\Bigg]^\gamma,
\end{aligned}
$$

where

$$\begin{aligned}\mathbb{E}_A[|f-g|] := \pi\Big(&\mathbb{E}_{A\cap D^+}[|-f+g||f\leq 0, g\leq 0]P_{A\cap D^+}(f\leq 0, g\leq 0))\\ &+\mathbb{E}_{A\cap D^+}[|-f-g||f\leq 0, g>0]P_{A\cap D^+}(f\leq 0, g>0))\\ &+\mathbb{E}_{A\cap D^+}[|f-g||f>0, g>0]P_{A\cap D^+}(f>0, g>0))\\ &+\mathbb{E}_{A\cap D^+}[|f+g||f>0, g\leq 0]P_{A\cap D^+}(f>0, g\leq 0))\Big)\\ &+\pi_-\Big(\mathbb{E}_{A\cap D^-}[|-f+g||f\leq 0, g\leq 0]P_{A\cap D^-}(f\leq 0, g\leq 0))\\ &+\mathbb{E}_{A\cap D^-}[|-f-g||f\leq 0, g>0]P_{A\cap D^-}(f\leq 0, g>0))\\ &+\mathbb{E}_{A\cap D^-}[|f-g||f>0, g>0]P_{A\cap D^-}(f>0, g>0))\\ &+\mathbb{E}_{A\cap D^-}[|f+g||f>0, g\leq 0]P_{A\cap D^-}(f>0, g\leq 0))\Big).\end{aligned}$$

and (using (6) and (7)) to rewrite the first line and fact 5 to establish the inequality.

$$\begin{aligned}&2\pi\Big[-\mathbb{E}_{H_f^-\cap D^+}[f|f\leq 0]P_{H_f^-\cap D^+}(f|f\leq 0)+\mathbb{E}_{H_f^-\cap D^+}[f|f>0]P_{H_f^-\cap D^+}(f|f>0)\Big]\\ &+2\pi_-\Big[-\mathbb{E}_{H_f^-\cap D^-}(f|f\leq 0)P_{H_f^-\cap D^-}(f|f\leq 0)+\mathbb{E}_{H_f^-\cap D^-}[f|f>0]P_{H_f^-\cap D^-}(f|f>0)\Big]\\ &\leq 2\Big[2\pi_f P_{H_f^-\cap D^+}(f|f\leq 0)c^{max}_{H_f^-}(f)-\pi_f P_{H_f^-\cap D^+}(f\leq 0)\mathbb{E}_{H_f^-\cap D^+}[f|f>0]+\pi_f\mathbb{E}_{H_f^-\cap D^+}(f|f>0)P_{H_f^-\cap D^+}(f>0)\\ &-\Big(-(1-\pi_f)2P_{H_f^-\cap D^-}(f|f>0)c^{max}_{H_f^-}(f)+(1-\pi_f)P_{H_f^-\cap D^-}(f|f>0)\mathbb{E}_{H_f^-\cap D^-}[f|f>0]\\ &+(1-\pi_f)P_{H_f^-\cap D^-}(f|f\leq 0)\mathbb{E}_{H_f^-\cap D^-}[f|f\leq 0]\Big)\Big]\\ &=2\Big[2R_{H_f^-}(f)c^{max}_{H_f^-}(f)+\pi_f\mathbb{E}_{H_f^-\cap D^+}[f]-(1-\pi_f)\mathbb{E}_{H_f^-\cap D^-}[f]\Big].\end{aligned}$$

and

$$R_{H_f^-}(f) := \Big[\pi_f P_{H_f^-\cap D^+}(f|f\leq 0)+(1-\pi_f)P_{H_f^-\cap D^-}(f|f>0)\Big].$$

□

Appendix E Proof of proposition 3

In essence, we wish to identify the conditions under which $\overline{s}(f,g)_S \lesseqgtr \overline{s}(f,g)_D$.

We first note that the following result is immediate from proposition 1.

Corollary 1.

$$\mathbb{E}_{H_f^-}[c^*(f)]_S \leq \mathbb{E}_{H_f^-}[c^*(f)]_D.$$

Using lemma 1, we can lower bound (3) from the main text as follows:

$$\Bigg(\mathbb{E}_{H_f^-}[c^*(f)]_M + \mathbb{E}_{H_g^-}[c^*(g)]_M \\ + \alpha\mathbb{E}_{H_f^- \cup H_g^-}[|f(x) - g(x)|]_M\Bigg)^{\gamma}\Bigg(8 \cdot \alpha\Bigg)^{1-\gamma} \\ \leq \mathbb{E}_{H_f^- \cup H_g^-}[c^*(f,g)]_M.$$

For simplicity of the statement we assume that $\gamma = 1$. We can now find an upper bound for the expected inverse cost of negative surprise under method $M = \{S, D\}$:

$$\overline{s}(f,g)_M = \frac{\mathbb{E}_{H_f^-}[c^*(f)]_M}{\mathbb{E}_{H_f^- \cup H_g^-}[c^*(f,g)]_M} \leq \frac{\mathbb{E}_{H_f^-}[c^*(f)]_M}{\mathbb{E}_{H_f^-}[c^*(f)]_M + \mathbb{E}_{H_g^-}[c^*(g)]_M + \alpha\mathbb{E}_{H_f^- \cup H_g^-}[|f(x) - g(x)|]_M}.$$

For simplicity, by corollary 1 we can set:

$$\mathbb{E}_{H_f^-}[c^*(f)]_S + \delta_f = \mathbb{E}_{H_f^-}[c^*(f)]_D,$$
$$\mathbb{E}_{H_g^-}[c^*(g)]_S + \delta_g = \mathbb{E}_{H_g^-}[c^*(g)]_D,$$

for some $\delta_f, \delta_g > 0$.

Proposition. Suppose $\overline{s}(f,g)_S > \overline{s}(f,g)_D$ and $\mathbb{E}_{H_f^- \cup H_g^-}[|f(x) - g(x)|]_S = \mathbb{E}_{H_f^- \cup H_g^-}[|f(x) - g(x)|]_D := k$ hold, then we must have:

$$\frac{\mathbb{E}_{H_g^-}[c^*(g)]_D}{\mathbb{E}_{H_f^-}[c^*(f)]_D} < \frac{\mathbb{E}_{H_g^-}[c^*(g)]_S}{\mathbb{E}_{H_f^-}[c^*(f)]_S}.$$

Using the definition $\overline{s}(f,g)_M$ and the stated assumptions we can write $\overline{s}(f,g)_S > \overline{s}(f,g)_D$ as follows:

$$1 + \underbrace{\frac{\mathbb{E}_{H_g^-}[c^*(g)]_S + \delta_g}{\mathbb{E}_{H_f^-}[c^*(f)]_S + \delta_f}}_{b_1} + \underbrace{\frac{\alpha k}{\mathbb{E}_{H_f^-}[c^*(f)]_S + \delta_f}}_{a_1} < 1 + \underbrace{\frac{\mathbb{E}_{H_g^-}[c^*(g)]_S}{\mathbb{E}_{H_f^-}[c^*(f)]_S}}_{b_2} + \underbrace{\frac{\alpha k}{\mathbb{E}_{H_f^-}[c^*(f)]_S}}_{a_2}$$

Note that $a_1 < a_2$ for $\delta_f > 0$. So the terms that remain to be checked are b_1 and b_2. Hence, we obtain

$$\frac{\mathbb{E}_{H_g^-}[c^*(g)]_D}{\mathbb{E}_{H_f^-}[c^*(f)]_D} < \frac{\mathbb{E}_{H_g^-}[c^*(g)]_S}{\mathbb{E}_{H_f^-}[c^*(f)]_S},$$

as desired.

Appendix F Other Prerequisites

Fact 1. $max(a,b) = \frac{1}{2}(a + b + |a - b|)$.
Fact 2. $P(A \cap B) \leq P(A); P(A \cap B) \leq P(B)$.
Fact 3. $E[X^\gamma] \leq E[X]^\gamma$ *for* $0 \leq \gamma \leq 1$ *(by Jensen's inequality).*
Fact 4. $P(X) \leq P(X)^\gamma$ *for* $0 \leq \gamma \leq 1$.
Fact 5. *Note that* $\pi \leq \pi_f$ *and* $\pi \leq \pi_g$.

We state the following lemmata without proof.

Lemma 1 (Ustun et al. (2019))**.** *For* $\gamma = 1$*, the expected cost of counterfactual explanations under model* f*,* $\mathbb{E}_{H_f^-}[c^*(f,x)]$*, is bounded from above such that:*

$$\overline{cost}_{H_f^-}(f) \leq \alpha\Bigg(\pi_f \cdot c_{D^+}(f) - (1 - \pi_f) \cdot c_{D^-}(f) + 2 \cdot c_{H_f^-}^{max} \cdot R_{H_f^-}(f)\Bigg).$$

Lemma 2 (Fawzi et al. (2018))**.** *Let $z_1, ..., z_n$ be non-negative real numbers, and let $0 \leq \gamma \leq 1$. Then*

$$\sum_{i=1}^{n} z_i^{\gamma} \leq n^{1-\gamma} \cdot \left(\sum_{i=1}^{n} z_i \right)^{\gamma}.$$