

## A Analytical $\alpha$ -value, $\alpha_0$

Define  $y(\alpha) = z + \alpha\Delta_{p,q}$  to be the line intersecting  $z$  with direction  $\Delta_{p,q}$ . We wish to identify the intersection between  $y(\alpha)$  and the hyperplane that constitutes the decision boundary between the two normal distributions  $\mathcal{N}(\mu_p, \mathbb{I})$  and  $\mathcal{N}(\mu_q, \mathbb{I})$ . Due to the identity covariance matrices of the Gaussians, we can define  $w = \mu_q - \mu_p$  and  $b = -\left(\frac{\mu_p + \mu_q}{2}\right)^\top w$  to form the decision boundary

$$w^\top x + b = 0. \quad (4)$$

Equation (4) corresponds to the blue line in Figure 2.

To find the  $\alpha$ -value which corresponds to the intersection, set  $x = z + \alpha\Delta_{p,q}$  and solve for  $\alpha$  in Equation (4):

$$w^\top (z + \alpha\Delta_{p,q}) + b = 0 \quad (5)$$

$$\Rightarrow \alpha w^\top \Delta_{p,q} = -(w^\top z + b) \quad (6)$$

$$\Rightarrow \alpha = -\frac{w^\top z + b}{w^\top \Delta_{p,q}}. \quad (7)$$

**Choice of  $\alpha_1$  value** As described, we found  $\alpha_1 = \alpha_0 + \frac{4}{5}(1 - \alpha_0)$  to be an appropriate value for generating convincing counterfactuals across the three datasets covered in this work. That said,  $\alpha_1 = 1$  would probably also have worked out fine. However, the goal was to stay as close as possible to  $\alpha_0$  to change as little as possible, while still generating convincing counterfactuals.

To give the reader an idea of the effect, we plot counterfactuals for five different inputs for varying values of  $t$  in Figure 7. In the plot,  $\alpha_1$  is determined as a function of  $t$  and  $\alpha_0$ :

$$\alpha_1 = \alpha_0 + t(1 - \alpha_0). \quad (8)$$

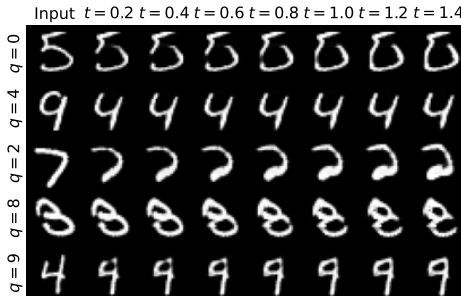


Figure 7: Effect of varying  $t$ , when generating counterfactuals using ECINN.

## B Tipping-point Counterfactuals

Here, we prove that ECINN produces tipping-point values according to Definition 1. Recall that we wish to find a constant  $C \in ]0; 1[$  and  $h : [0; 1] \rightarrow \mathcal{X}$  such that  $p(y|h(c)) > p(q|h(c))$  whenever  $c < C$ , similarly  $p(y|h(c)) < p(q|h(c))$  whenever  $c > C$ , and finally  $p(y|h(c)) = p(q|h(c))$  when  $c = C$ . In the following, we choose  $C$  and  $h$  in such a way that they comply with Definition 1.

*Proof.* Let  $C = \frac{1}{2}$  and define  $\bar{c}$  as

$$\bar{c} = \begin{cases} 2\alpha_0 & \text{if } c \leq \frac{1}{2} \\ \alpha_0 + 2c - 1 & \text{otherwise} \end{cases}. \quad (9)$$

This way,  $\bar{c} < \alpha_0$  when  $c < C$  and  $\bar{c} > \alpha_0$  when  $c > C$ .

Now define  $h(c) = f^{-1}(z + \bar{c}\Delta_{y,q})$ , where  $f$  is the INN,  $z = f(x)$ , and  $\Delta_{y,q}$  is as defined in Section 3.3. Assume further that  $G(x) = y$  and  $G(f^{-1}(x + \Delta_{y,q})) = q$ , i.e., the input sample is correctly classified and the counterfactual is classified as class  $q$ .

*Sketch of proof:* we use the property of Equation (2) to show that  $\|\mu_y - z + c\Delta_{y,q}\| < \|\mu_q - z + c\Delta_{y,q}\|$  when  $c < C$  and vice versa.

By the change-of-variable formula (Equation (1)), Bayes' theorem, and the assumption  $p(y) = \frac{1}{K}$ , we have the relation

$$p(y|x) = \frac{p(f(x)|y)p(y)}{\sum_{y'} p(f(x)|y')p(y')} \quad (10)$$

$$= \frac{p(f(x)|y)}{\sum_{y'} p(f(x)|y')} \quad (11)$$

$$\Rightarrow \log p(y|x) = \log p_{z|y}(f(x)) + \log |det(J)| - \log \left[ \sum_{y'} p_{z|y}(f(x)|y') \right] - \log |det(J)| \quad (12)$$

$$= \log p_{z|y}(f(x)) - \log \left[ \sum_{y'} p_{z|y}(f(x)|y') \right] \quad (13)$$

and an identical relation holds for  $p(q|x)$

$$\log p(q|x) = \log p_{z|q}(f(x)) - \log \left[ \sum_{y'} p_{z|q}(f(x)) \right] \quad (14)$$

For a fixed  $x$ , we see that for  $\log p(y|x)$  to be greater than  $\log p(q|x)$ , only the first term matters. As  $p_{z|y} = \mathcal{N}(\mu_y, \mathbb{1})$ , we have that

$$\log p_{z|y}(z) = -\frac{d}{2} \log 2\pi - \frac{1}{2} \|\mu_y - z\|^2 \propto \|\mu_y - z\|^2, \quad (15)$$

and similarly for  $p_{z|q}$ . As such, by injecting  $h(c)$  into  $\log p_{z|y}$  it suffices to prove that  $\|\mu_y - z + \bar{c}\Delta_{y,q}\| = \|\mu_q - z + \bar{c}\Delta_{y,q}\|$  when  $c = C$ ,  $\|\mu_y - z + \bar{c}\Delta_{y,q}\| < \|\mu_q - z + \bar{c}\Delta_{y,q}\|$  when  $c < C$  and vice versa.

First, note that when  $c = C$ , then  $\bar{c} = \alpha_0$  so  $z + \bar{c}\Delta_{y,q} = z + \alpha_0\Delta_{y,q}$  and thus  $\|\mu_y - z + \bar{c}\Delta_{y,q}\| = \|\mu_q - z + \bar{c}\Delta_{y,q}\|$  holds by construction of  $\alpha_0$ .

Second, from the assumption that  $G(x) = y$ , we know that  $\|\mu_y - z + 0\Delta_{y,q}\| < \|\mu_q - z + 0\Delta_{y,q}\|$ . Similarly, from the assumption that  $G(f^{-1}(x + \Delta_{y,q})) = q$ , we know that  $\|\mu_y - z + \Delta_{y,q}\| > \|\mu_q - z + \Delta_{y,q}\|$ . It follows, that when  $c < C$ ,  $\|\mu_y - z + c\Delta_{y,q}\| < \|\mu_q - z + c\Delta_{y,q}\|$  and vice versa.  $\square$

## C Experimental Details

In Table 2, we provide an overview of hyperparameters and performances of the networks used in this work.

**IB-INN.** We have trained IB-INN models “as-is”<sup>5</sup> and adjusted only the  $\beta$ -value of the loss function. On FakeMNIST and MNIST, the IB-INN models were trained for 60 epochs with stochastic gradient descent and a milestone scheduler stepping from learning rate 0.07 to 0.007 after 50 epochs. On CelebA-HQ, the IB-INN models were trained for 800 epochs with the Adam optimizer [18] and a milestone scheduler stepping with a factor  $\frac{1}{10}$  after every 200 epochs.

## D IB-INN Model and Loss

The model architecture and loss function used in this work were proposed by [9]. The loss was derived from an information bottleneck formulation with a hyperparameter,  $\beta$ , that allows trading off generative and classification capabilities. The loss function is based on mutual information  $I$ :

$$\mathcal{L}_{IB} = I(X, Z) - \beta I(Z, Y). \quad (16)$$

Mutual information quantifies the amount of information which is shared between variables.<sup>6</sup> As such, by minimizing  $\mathcal{L}_{IB}$ , the mutual information between the input and the latent vector is minimized while the mutual information between the latent vector and class label is maximized. In practice, the first term,  $I(X, Z)$ , can be thought of as a generative loss, which results in a good performance on generating images. The second term,  $I(Z, Y)$ , is closely

<sup>5</sup>IB-INN code: <https://github.com/VLL-HD/IB-INN>

<sup>6</sup>For an invertible mapping  $f$  and  $Z = f(X)$ ,  $\mathcal{L}_{IB}$  is, in fact, ill-defined, and the authors [9] add noise to  $X$  to overcome the issue.

Dataset	$\beta$	BPD	Err.
FakeMNIST	1.4265	1.77	0%
MNIST	1.4265	1.89	0.85%
CelebA-HQ			
Smile	1	3.32	7.42%
High cheekbones	1	3.09	14.38%
Lipstick	1	3.06	4.87%
Heavy makeup	1	3.08	12.68%

Table 2: Hyperparameters, negative log-likelihood measured in bits per dimension (BPD), and error rates for the models used in this work.

related to the categorical cross-entropy loss, thus promoting high accuracy. Throughout our experiments, we use models trained with the IB-INN loss,  $\mathcal{L}_{IB}$ .

For simplicity, we do not include experiments across multiple values of  $\beta$  in the main paper. Overall, we find that values close to one strike a good balance between counterfactual examples and model accuracy in our experiments. We do, however, include Figure 8 which demonstrates the conflicting effect of  $\beta$  on the quality of counterfactuals and the accuracy of the model.

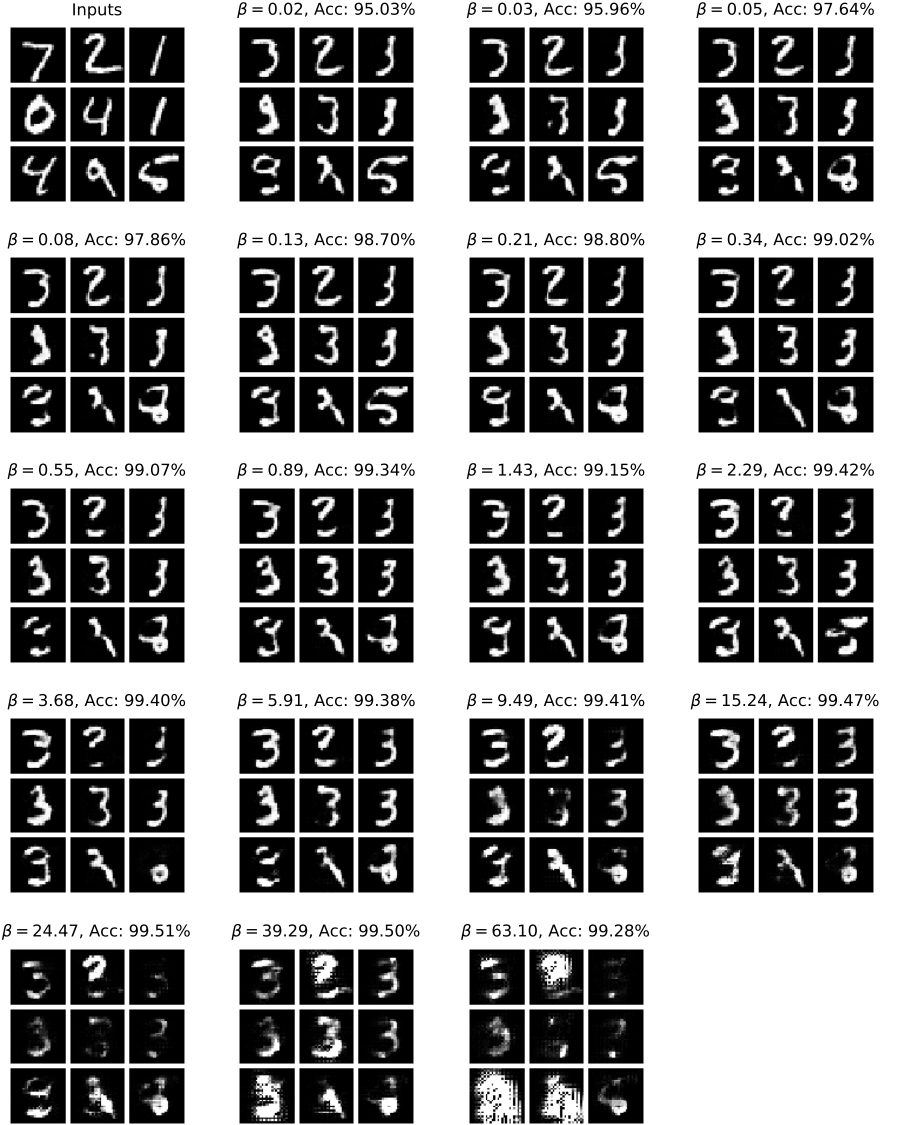


Figure 8: Counterfactual examples for MNIST models trained with different values of  $\beta$ . The top left square represents the input images that are all changed with target  $q = 3$ . Above plots are  $\beta$ -values in ascending order and corresponding test set accuracies.



(a) Label: High cheekbones.



(b) Label: Wearing libstick.



(c) Label: Heavy makeup.

Figure 9: CelebA-HQ counterfactual examples. First five columns are inputs with negative labels and counterfactuals with positive labels and vice versa for the last five columns.

## E Additional Samples

In Figure 9, we include counterfactual examples similar to Figure 6 for three additional labels. We also include pdfs with extra samples of all figures from our experiments. For each figure, there is a corresponding pdf in the related work zip-file. For example, Figure 9a has a corresponding pdf in the supplementary material named `figure9a.pdf` with additional samples.

## F Hardware Specifications

All experiments were run on a single machine learning server with 128GB system memory, an Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz processor, and 5 NVIDIA RTX 2080 Ti. See all details below.

```
$ nvidia-smi -L
GPU 0: GeForce RTX 2080 Ti (UUID: ...)
GPU 1: GeForce RTX 2080 Ti (UUID: ...)
GPU 2: GeForce RTX 2080 Ti (UUID: ...)
GPU 3: GeForce RTX 2080 Ti (UUID: ...)
GPU 4: GeForce RTX 2080 Ti (UUID: ...)

$ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                48
On-line CPU(s) list:   0-47
Thread(s) per core:    2
Core(s) per socket:    12
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz
Stepping:              7
CPU MHz:               1000.777
CPU max MHz:           2201.0000
CPU min MHz:           1000.0000
BogoMIPS:              4400.00
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              1024K
L3 cache:              16896K
NUMA node0 CPU(s):     0-11,24-35
NUMA node1 CPU(s):     12-23,36-47

$ lsmem

```

RANGE	SIZE	STATE	REMOVABLE	BLOCK
0x0000000000000000-0x0000000007ffffffff	2G	online	no	0
0x0000000010000000-0x0000000027ffffffff	6G	online	yes	2-4
0x0000000028000000-0x0000000067ffffffff	18G	online	no	5-13
0x0000000070000000-0x0000000077ffffffff	4G	online	yes	14-15
0x0000000080000000-0x00000000f7ffffffff	30G	online	no	16-30
0x00000000f8000000-0x00000000ffffffff	2G	online	yes	31
0x0000001000000000-0x0000002077ffffffff	66G	online	no	32-64

```

Memory block size:      2G
Total online memory:    128G
Total offline memory:    0B
```