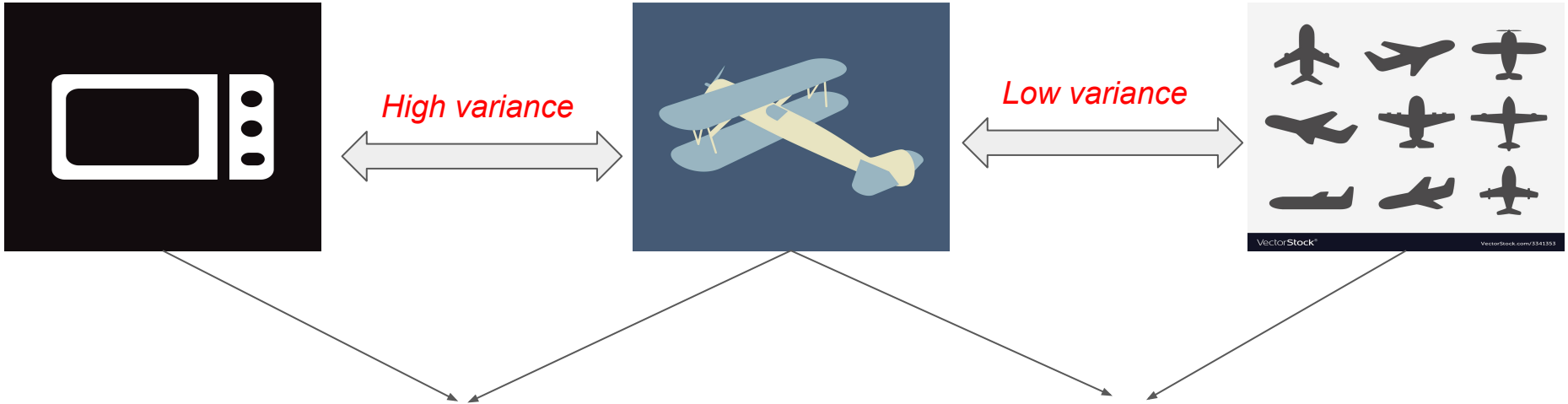
Contrastive Explanations

Limitations

#1: The Curse of Variance



Huge # of PN decreasing interpretability.
Just PP alone suffices.

PP and PN together needed

Begs the question: What is the least amount of information needed to explain a class type?

#2 Performance of real valued datasets

Datasets used in experiments are inherently binary in nature

1. MNIST : Black (0) or White (1) Pixel
2. Procurement Fraud: features are present or not
3. Brain Functional Imaging: functional connectivity between different regions of the brain
4. Coloured Images ? *No experiments performed to comments on results*

the presence of those features/pixels. This idea also applies to colored images where the most prominent pixel value (say median/mode of all pixel values) can be considered as no signal and moving away from this value can be considered as adding signal. One may also argue that there is some information loss in our form of explanation, however we believe that such explanations are lucid

#3 Hyperparameters

There is no concrete method to tune for **β** and ***γ*** and tuning **c** is computationally intensive. Also no mention on the confidence **κ**

$$\min_{\boldsymbol{\delta} \in \mathcal{X} \cap \mathbf{x}_0} c \cdot f_\kappa^{\text{pos}}(\mathbf{x}_0, \boldsymbol{\delta}) + \beta \|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\delta}\|_2^2 + \gamma \|\boldsymbol{\delta} - \text{AE}(\boldsymbol{\delta})\|_2^2,$$

c requires computation on all examples in an iteration and uses 9 iterations for a dataset

Fixed to 0.1

Either {0,100}

#4 Computationally Intensive?

- Looks intensive:

$$f_\kappa^{\text{neg}}(\mathbf{x}_0, \boldsymbol{\delta}) = \max\{[\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_{t_0} - \max_{i \neq t_0}[\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_i,$$

assume $\mathcal{X} = [-1, 1]^p$, $\mathcal{X}/\mathbf{x}_0 = [0, 1]^p$ and let $g(\boldsymbol{\delta}) = f_\kappa^{\text{neg}}(\mathbf{x}_0, \boldsymbol{\delta}) + \|\boldsymbol{\delta}\|_2^2 + \gamma\|\mathbf{x}_0 + \boldsymbol{\delta} - \text{AE}(\mathbf{x}_0 + \boldsymbol{\delta})\|_2^2$ denote the objective function of (1) without the L_1 regularization term. Given the initial iterate $\boldsymbol{\delta}^{(0)} = \mathbf{0}$, projected FISTA iteratively updates the perturbation I times by

3X

$$\boldsymbol{\delta}^{(k+1)} = \Pi_{[0,1]^p}\{S_\beta(\mathbf{y}^{(k)} - \alpha_k \nabla g(\mathbf{y}^{(k)}))\}; \quad (5)$$

$$\mathbf{y}^{(k+1)} = \Pi_{[0,1]^p}\{\boldsymbol{\delta}^{(k+1)} + \frac{k}{k+3}(\boldsymbol{\delta}^{(k+1)} - \boldsymbol{\delta}^{(k)})\}, \quad (6)$$

2*3X for both PP and PN

- No mention of computational costs of optimization.

#5 Just Classification