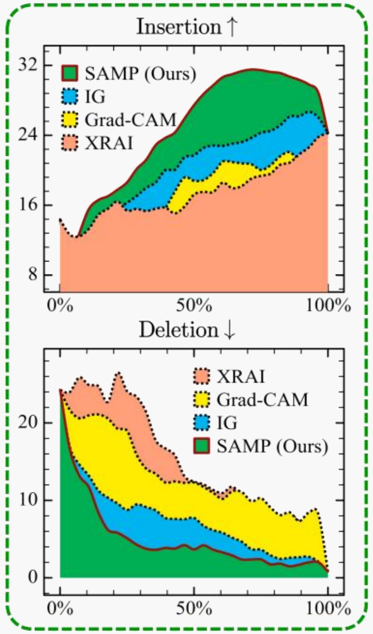
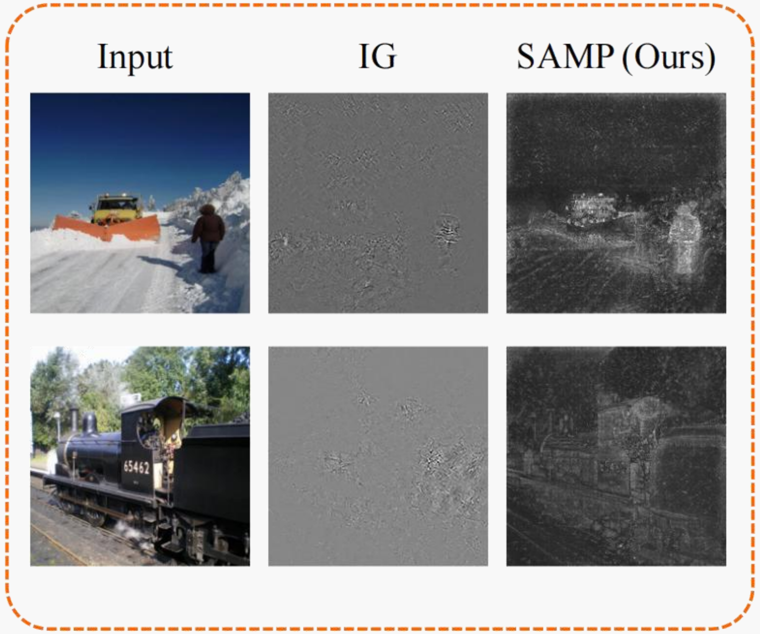
Path Choice Matters for Clear Attributions in Path Methods

Borui Zhang, Wenzhao Zheng, Jie Zhou, Jiwen Lu*
Department of Automation, Tsinghua University, China

ICLR 2024





Interpretable methods need completeness and clarity

- ✓ **Completeness:** Remove the uncertainty of the interpretation itself!
- ✓ **Clarity:** Precise definition of "contribution / significance"!

Visualization

Effect

Data: Noise

Model: Misclassification

~~Interpretation: Uncertainty~~

Cause

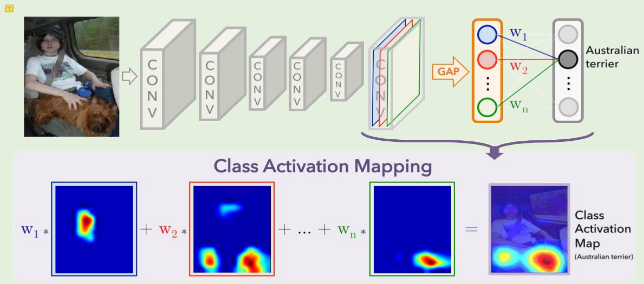
Clear attribution

contribution / significance

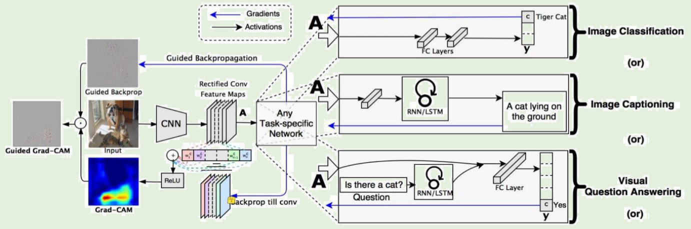
Interpretable methods need completeness and clarity

Lack completeness: e.g., CAM、Grad-CAM

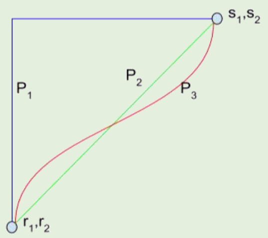
CAM:
Observation-based

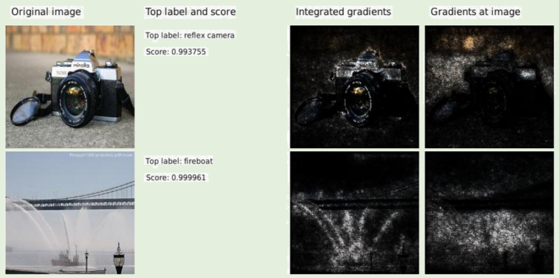


Grad-CAM:
Based on Taylor expansion; Large approximation error!



Lack clarity: e.g., Integrated Gradients





Axioms:
1: Linearity
2: Dummy player
3: Effectiveness

$$\text{IntegratedGrads}_i^{approx}(x) ::= \boxed{(x_i - x'_i) \times \Sigma_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}}$$

BUT path choice influences attribution results

(1) Preliminary: path methods

Line integral from source x^S to target x^T

Line integral expansion

$$\Delta y = y^T - y^0 = \int_\gamma \nabla f(\boldsymbol{x})^T \, \mathrm{d}\boldsymbol{x}$$
$$= \int_{\alpha=0}^{1} \frac{\partial f(\gamma(\alpha))}{\partial \gamma(\alpha)} \frac{\partial \gamma(\alpha)}{\partial \alpha} \mathrm{d}\alpha$$

Definition of attributions

$$a_i \triangleq \int_{\alpha=0}^{1} \frac{\partial f(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} \mathrm{d}\alpha.$$

Completeness

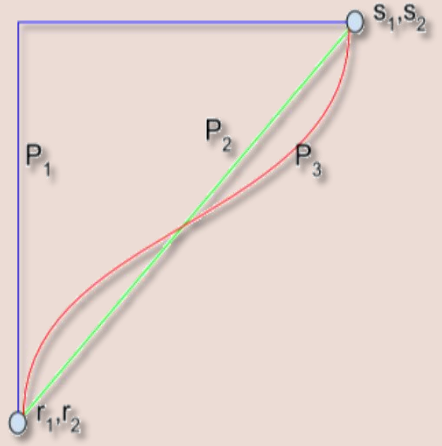
$$\Delta y = \int_{\alpha=0}^{1} \sum_{i=1}^{d} \frac{\partial f(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} \mathrm{d}\alpha = \sum_{i=1}^{d} a_i$$

Path methods

IG (2017)
- Straight line path in spatial domain

BlurIG (2020)
- Straight line path in frequency domain

GuidedIG (2021)
- Adjust paths based on the gradient fluctuation



Different path functions γ lead to **ambiguity** in attribution results!

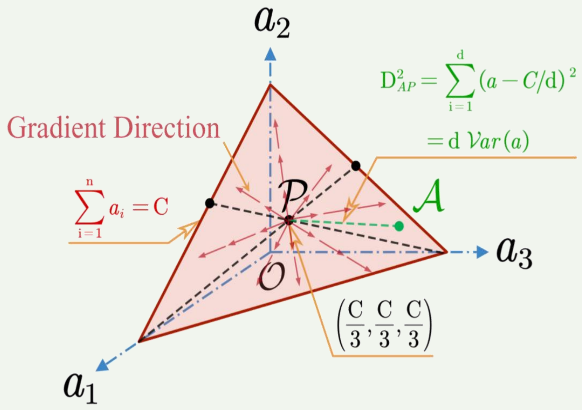
(2) Path selection criterion

Path selection criterion

Definition 1 (Concentration Principle)

A path function γ is said to satisfy **Concentration Principle** if attribution a achieves the max $Var(a) = \frac{1}{d}\sum_{i=1}^{d}(a_i - \bar{a})^2$

- **Intuitive understanding:** **the isotropic field centered on the averaged point $\mathcal{P}$**



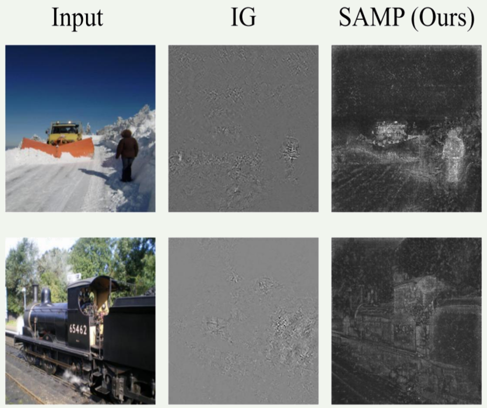
Concrete examples

Example 1:

- **For a 3-feature case, this principle prefers (0.7, 0.2, 0.1) to (0.4, 0.3, 0.3)**

Example 2:

- **For visual data, this principle leads to sparse and aesthetic attributions**



(3) Approximate solution derivation

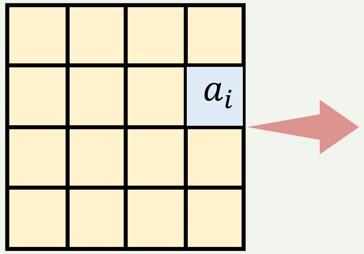
Data distribution assumption → Propositions

Hard to solve: $\gamma^* = \underset{\gamma \in \Gamma}{argmin} \frac{1}{d} \sum_{i=1}^{d} \left(a_i(\gamma) - \bar{a}(\gamma)\right)^2$

Assumption 1 (Allocation as Brownian motion)

We assume the additive process $\{u_t, t \geq 0\}$ as the Brownian motion and $u_t \sim \mathcal{N}(0, \sigma t)$ if without any constraint condition

- **NOT directly assumed the Brownian motion under the condition** $\sum a_i = C$



The key is to solve the joint conditional distribution

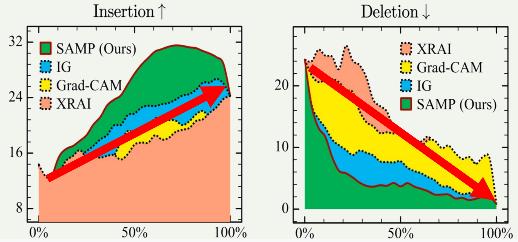
$$P\left(a_1, a_2, \cdots, a_d \middle| \textstyle\sum_i a_i = C\right)$$

Proposition 1 (Joint conditional distribution)

By Brownian motion assumption, the conditional joint distribution $P(\tilde{a}|u_d = C)$ is a multivariate Gaussian distribution as:

$$P(\tilde{\boldsymbol{a}}|C) = \frac{1}{(2\pi)^{\frac{d-1}{2}}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}\left\|\tilde{\boldsymbol{a}} - \frac{C}{d}\mathbf{1}\right\|^2_{\Sigma^{-1}}\right\} \text{ where } \Sigma = \frac{\sigma}{d}\begin{bmatrix} d-1 & -1 & \cdots & -1 \\ -1 & d-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & d-1 \end{bmatrix} \in \mathbb{R}^{(d-1)\times(d-1)}.$$

- **Explanation:** $\left(u_k = \sum_{i=1}^{k} a_i\right)$
  - **For any i,j, conditional covariance** $Cov(a_i, a_j\, | u_d = C) = - \sigma/d$
  - **For any i, conditional expectation** $E(u_k \,|\, u_d = C) = kC/d$**, which indicates that random paths tend to linear perturbations**



(3) Approximate solution derivation

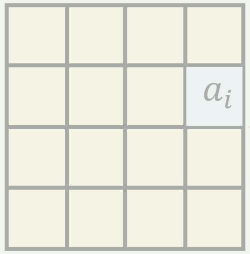
Data distribution assumption → Propositions

Hard to solve: $\gamma^* = \underset{\gamma\in\Gamma}{argmin}\frac{1}{d}\sum_{i=1}^{d}\left(a_i(\gamma) - \bar{a}(\gamma)\right)^2$

Assumption 1 (Allocation as Brownian motion)

We assume the additive process $\{u_t, t \geq 0\}$ as the Brownian motion and $u_t \sim \mathcal{N}(0, \sigma t)$ if without any constraint condition

- **NOT directly assumed the Brownian motion under the condition $\sum a_i = C$**



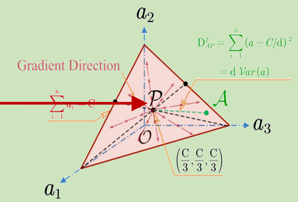
The key is to solve the joint conditional distribution

$$P\left(a_1, a_2, \cdots, a_d \middle| \sum_i a_i = C\right)$$

Proposition 2 (Asymptotic property)

The covariance matrix of the multivariate Gaussian distribution tends to σI as $d \to \infty$

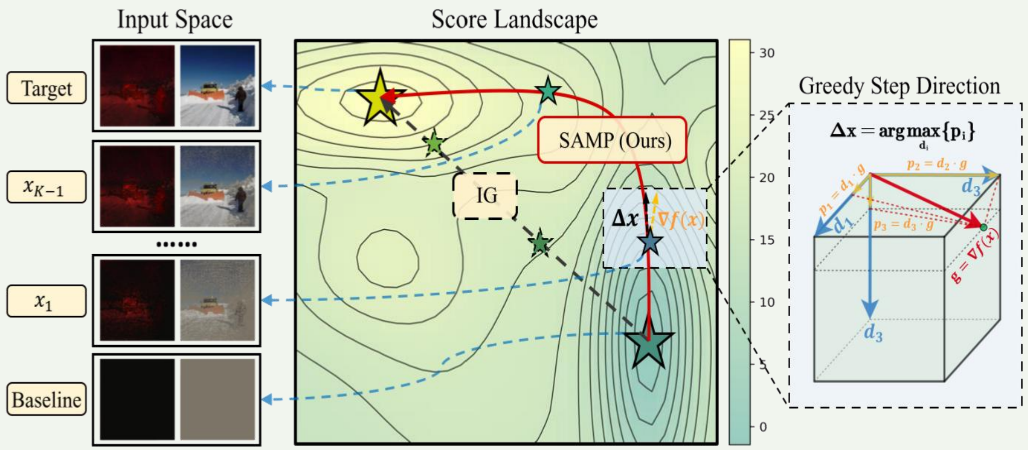
$$\hat{P}(\tilde{\mathbf{a}}|C) = \frac{1}{(2\pi)^{\frac{d-1}{2}}\sqrt{|\Sigma|}}\exp\left(-\boxed{\frac{D_{ap}^2}{2\sigma}}\right)$$



- **Explanation:**
  - **Probability approximation: The original distribution is approximated by $\widehat{P}(\tilde{a}|C) = P(\tilde{a}|C)\, e^{a_d^2/(2\sigma)}$ with tolerable error**
  - **Asymptotically independence: The conditional covariance $Cov(a_i, a_j \,|u_d = C)$ tends to 0**
- **Conclusion:**
  - **Employ the greedy optimization method to assign attributions to pixels in turn!**

(4) Greedy optimization algorithm

Complete algorithm flow



SAMP (analogous for deletion) as follows:

$$(\mathrm{d}\boldsymbol{x}^k)_i = \begin{cases} x_i^E - x_i^k, & i \in \mathbb{M}_k \\ 0, & \text{Otherwise} \end{cases} \qquad (11)$$

where $\mathbb{M}_k = \{i \mid i \in top_s\{\alpha_j\}\}$,

and $\alpha_j = \begin{cases} (\nabla f(\boldsymbol{x}^k))_j (x_j^E - x_j^k), & \text{if } x_j^E \neq x_j^k \\ -\infty, & \text{Otherwise} \end{cases}$

Initialization: model f, data point x_s, baseline point x_0

(1) $x_i \leftarrow x_s$, the target direction is $\Delta \mathrm{x} = \mathrm{x}_0 - \mathrm{x}_s$

(2) Forward propagation: compute prediction scores $y_i = f(x_i)$

(3) Backward propagation: compute gradients $\mathrm{g_i} = \nabla_{\mathrm{x_i}} \mathrm{f}(\mathrm{x_i})$

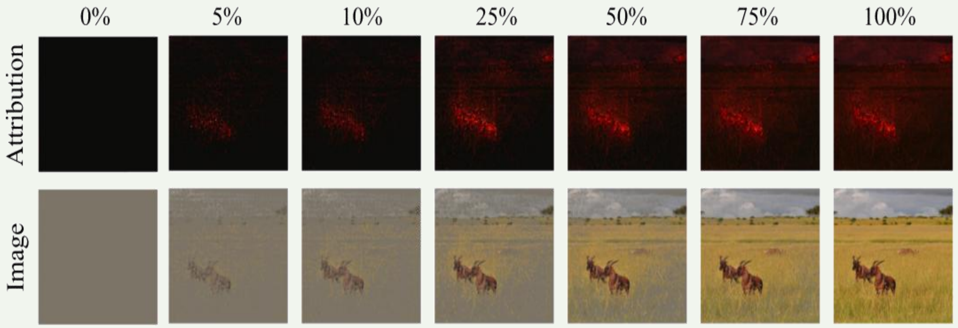
(4) Find the optimal projection gradient $\underset{m_i}{\mathrm{argmax}} |m_i(\Delta \mathrm{x})^{\mathrm{T}} \mathrm{g_i}|$ along Δx, where m_i indicates the mask

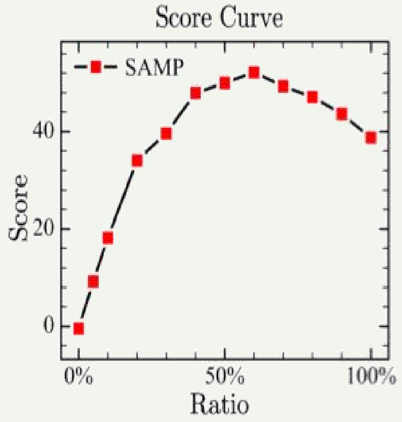
(5) Update $x_i \leftarrow x_i + m_i(\Delta \mathrm{x})$, return step (2), until the end

Output: decomposition path $\Delta \mathrm{y} = \sum_i g_i m_i(\Delta) \rightarrow \int_L \nabla \mathrm{f}(\mathrm{x}) \mathrm{dx}$

Qualitative visualization results

Verification of Concentration Principle





Visualization comparison

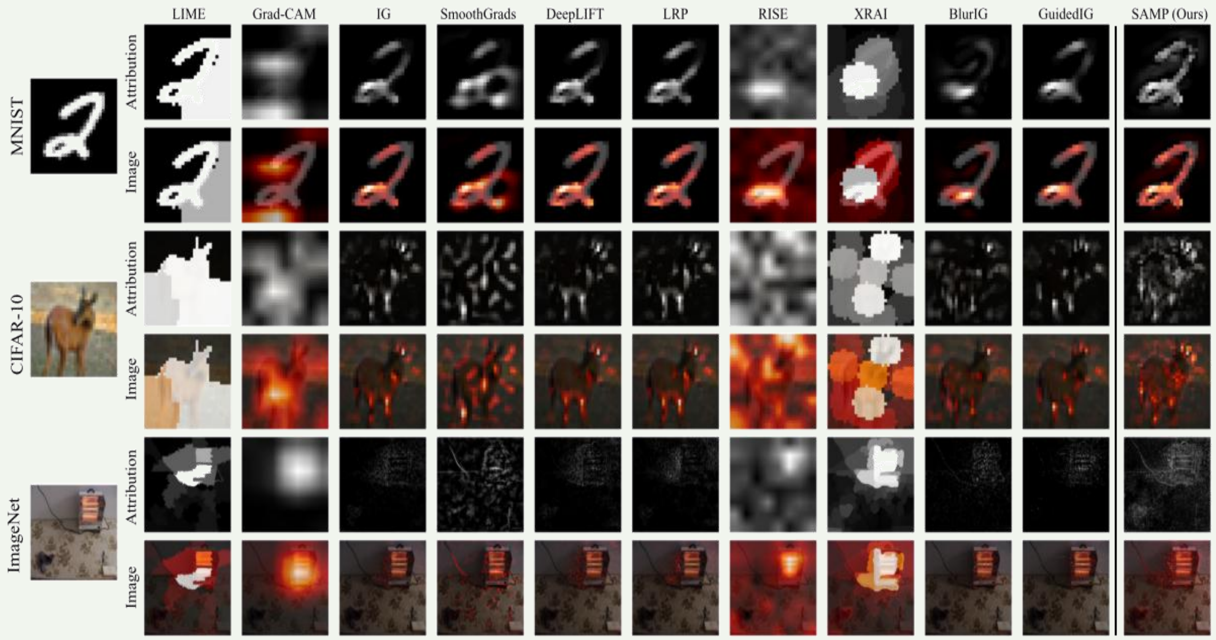


Figure 4. Visualization on MNIST, CIFAR-10, and ImageNet for comparison with other mainstream interpretation methods.

Quantitative result analysis

◻ Deletion/Insertion metrics

Table 1: Deletion/Insertion metrics on MNIST, CIFAR-10, and ImageNet compared with other interpretation methods.

Method	MNIST		CIFAR-10		ImageNet	
	Deletion↓	Insertion↑	Deletion↓	Insertion↑	Deletion↓	Insertion↑
LRP [4]	-0.003 (±0.133)	**0.808 (±0.102)**	-0.257 (±0.485)	**1.452 (±0.373)**	0.210 (±0.133)	0.575 (±0.150)
CAM [40]	0.221 (±0.154)	0.715 (±0.107)	0.314 (±0.307)	0.863 (±0.233)	0.313 (±0.129)	0.897 (±0.130)
LIME [25]	0.282 (±0.139)	0.597 (±0.093)	0.479 (±0.287)	0.722 (±0.235)	0.312 (±0.128)	**0.898 (±0.140)**
Grad-CAM [26]	0.221 (±0.154)	0.715 (±0.107)	0.314 (±0.307)	0.863 (±0.233)	0.313 (±0.129)	0.897 (±0.130)
IG [33]	-0.038 (±0.142)	0.795 (±0.105)	**-0.372 (±0.535)**	**1.452 (±0.400)**	0.197 (±0.129)	0.725 (±0.199)
SmoothGrads [31]	0.003 (±0.127)	0.547 (±0.110)	0.777 (±0.551)	0.517 (±0.283)	0.300 (±0.127)	0.605 (±0.171)
DeepLIFT [28]	-0.025 (±0.135)	0.791 (±0.105)	-0.300 (±0.513)	1.443 (±0.383)	0.216 (±0.122)	0.688 (±0.184)
Grad-CAM++ [7]	0.149 (±0.102)	0.776 (±0.065)	0.386 (±0.346)	0.795 (±0.203)	0.319 (±0.132)	0.890 (±0.128)
RISE [24]	0.059 (±0.111)	0.651 (±0.123)	0.149 (±0.349)	0.904 (±0.272)	0.282 (±0.131)	0.849 (±0.151)
XRAI [14]	0.120 (±0.117)	0.754 (±0.097)	0.248 (±0.330)	0.910 (±0.208)	0.346 (±0.161)	0.865 (±0.141)
Blur IG [36]	0.021 (±0.021)	0.804 (±0.170)	-0.107 (±0.387)	1.407 (±0.464)	0.261 (±0.144)	0.712 (±0.223)
Guided IG [15]	**-0.041 (±0.135)**	0.762 (±0.100)	-0.276 (±0.469)	1.209 (±0.349)	**0.167 (±0.126)**	0.699 (±0.210)
SAMP (ours)	-0.093 (±0.142)	1.074 (±0.176)	-0.733 (±0.671)	1.458 (±0.399)	0.154 (±0.118)	0.984 (±0.195)
SAMP++ (ours)	-0.137 (±0.151)	1.050 (±0.179)	-0.899 (±0.724)	1.514 (±0.425)	0.145 (±0.116)	1.116 (±0.241)

◻ **Significantly outperformed all comparison methods on Deletion/Insertion**

◻ Sensitivity-N check

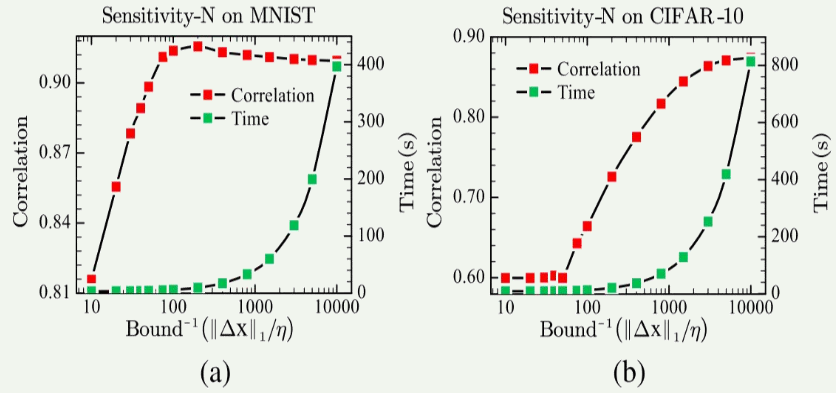


Figure 5. Sensitivity-N check for the infinitesimal constraint.

◻ **The infinitesimal constrant ensures the completeness of attributions**

TAKE-HOME MESSAGE

- **Core thoughts:** Axiomatic design for **unique Path selection** for clarity and **approximation assumption** for fast computation
- **Concentration Principle:** This Heuristic searching target makes attributions **sparse** and **aesthetic**
- **Greedy algorithm:** Greedy algorithm is used to solve the **nearly-optimal solution** under Brownian motion assumption
- **Limitation:** The algorithm cannot guarantee **strict global optimal**, and attributions depend on properties of models!

Thanks