南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect

Kaihua Tang[1], **Jianqiang Huang**[1,2], **Hanwang Zhang**[1]

[1]Nanyang Technological University, [2]Damo Academy, Alibaba Group

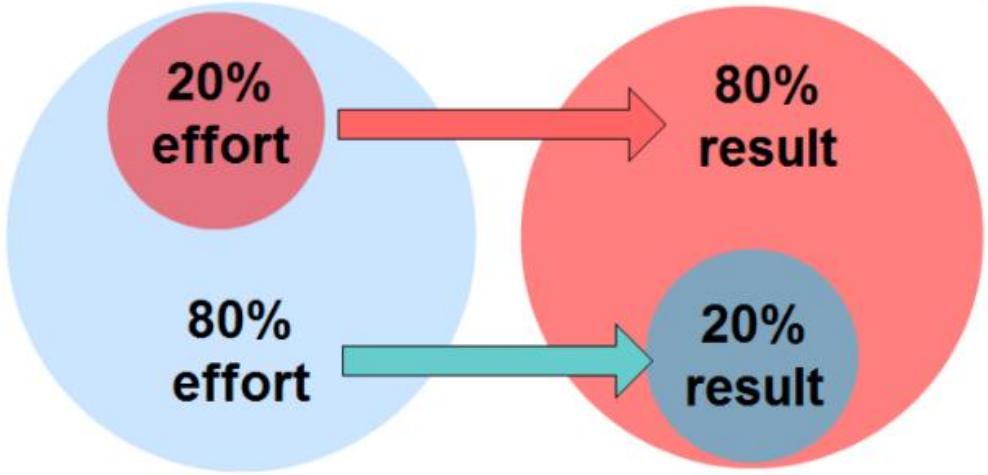
kaihua001@e.ntu.edu.sg, jianqiang.jqh@gmail.com, hanwangzhang@ntu.edu.sg

NIPS 2020

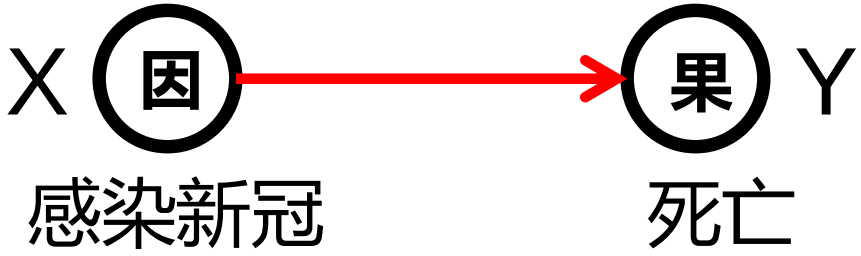
Contents

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Long-Tailed Distribution

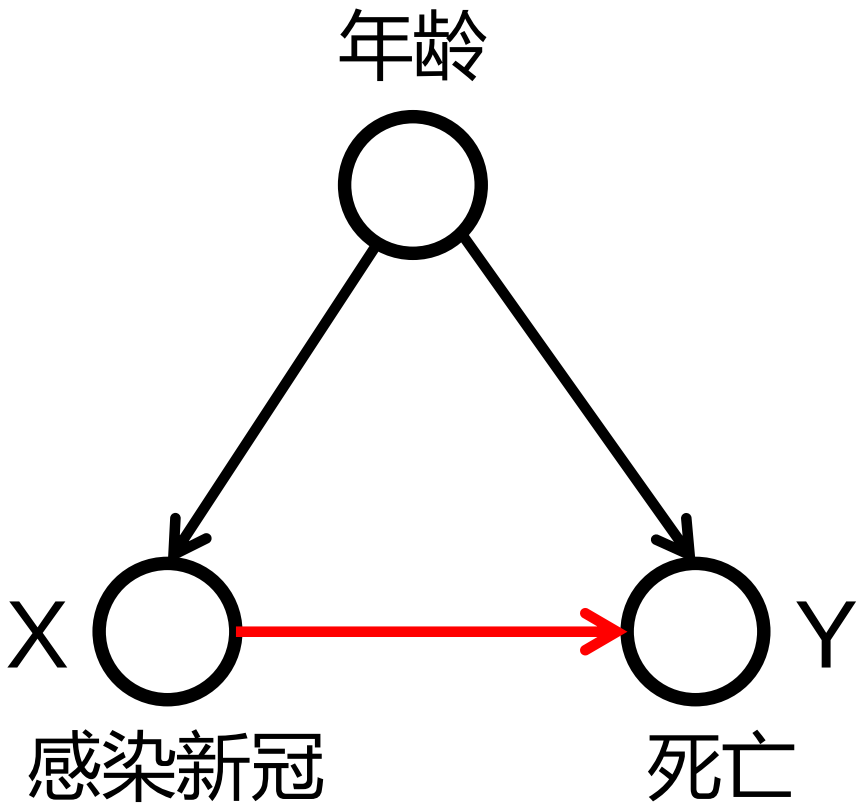


Causal Graph



- **节点**代表**变量**，**边**代表变量之间的**因果关系。**

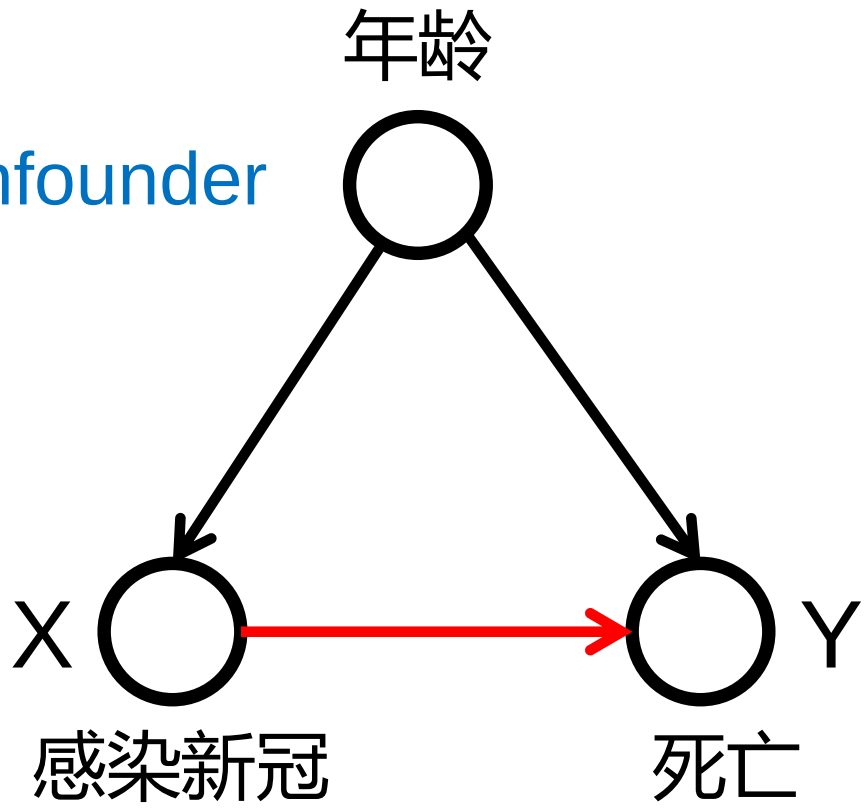
Causal Graph



- **节点**代表**变量**，**边**代表变量之间的**因果关系**。
- 这种因果图被命名为“fork”（**叉子**），特点是有**两条边**从**一个变量**中分岔。

Causal Graph

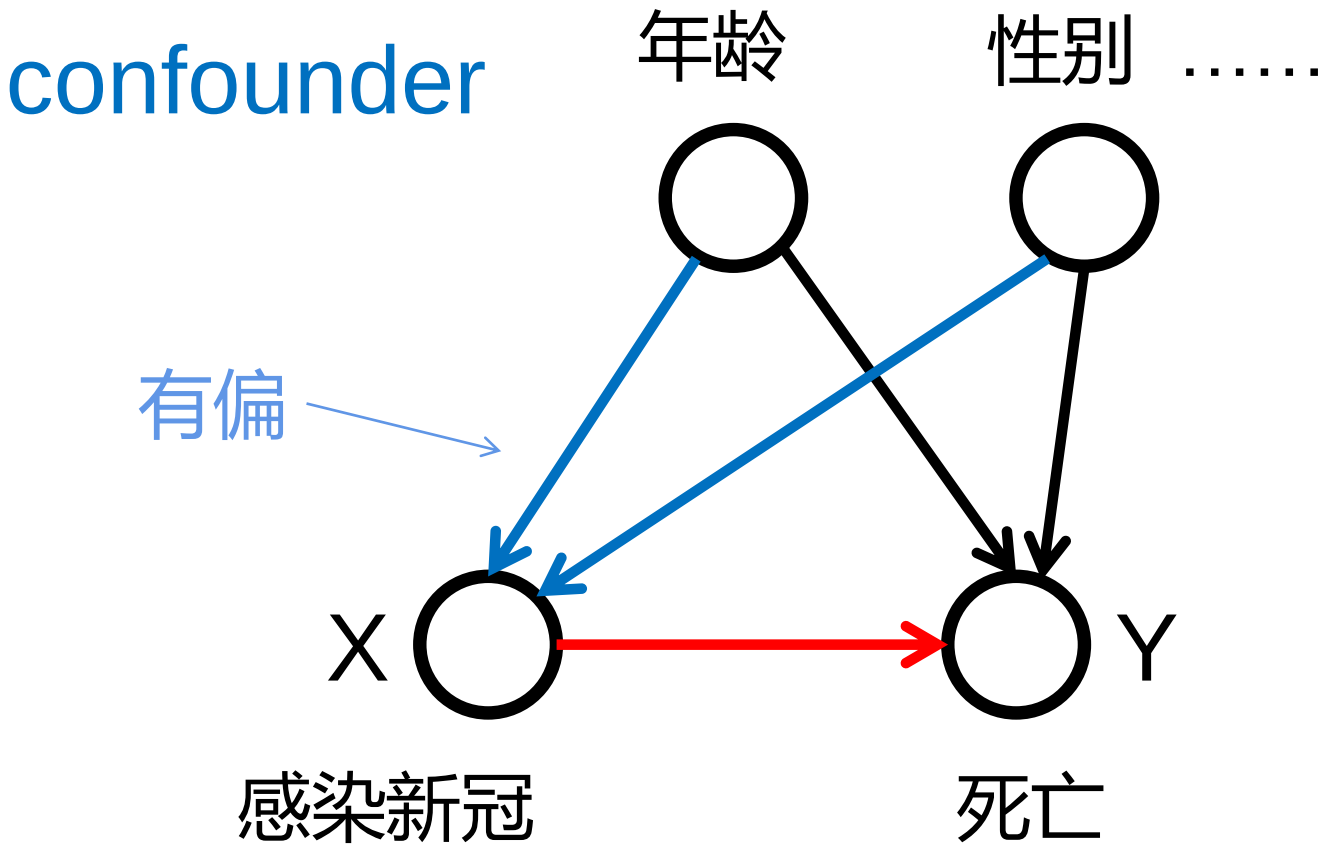
- 共因
- 混杂因素: confounder



- **节点**代表**变量**，**边**代表变量之间的**因果关系**。

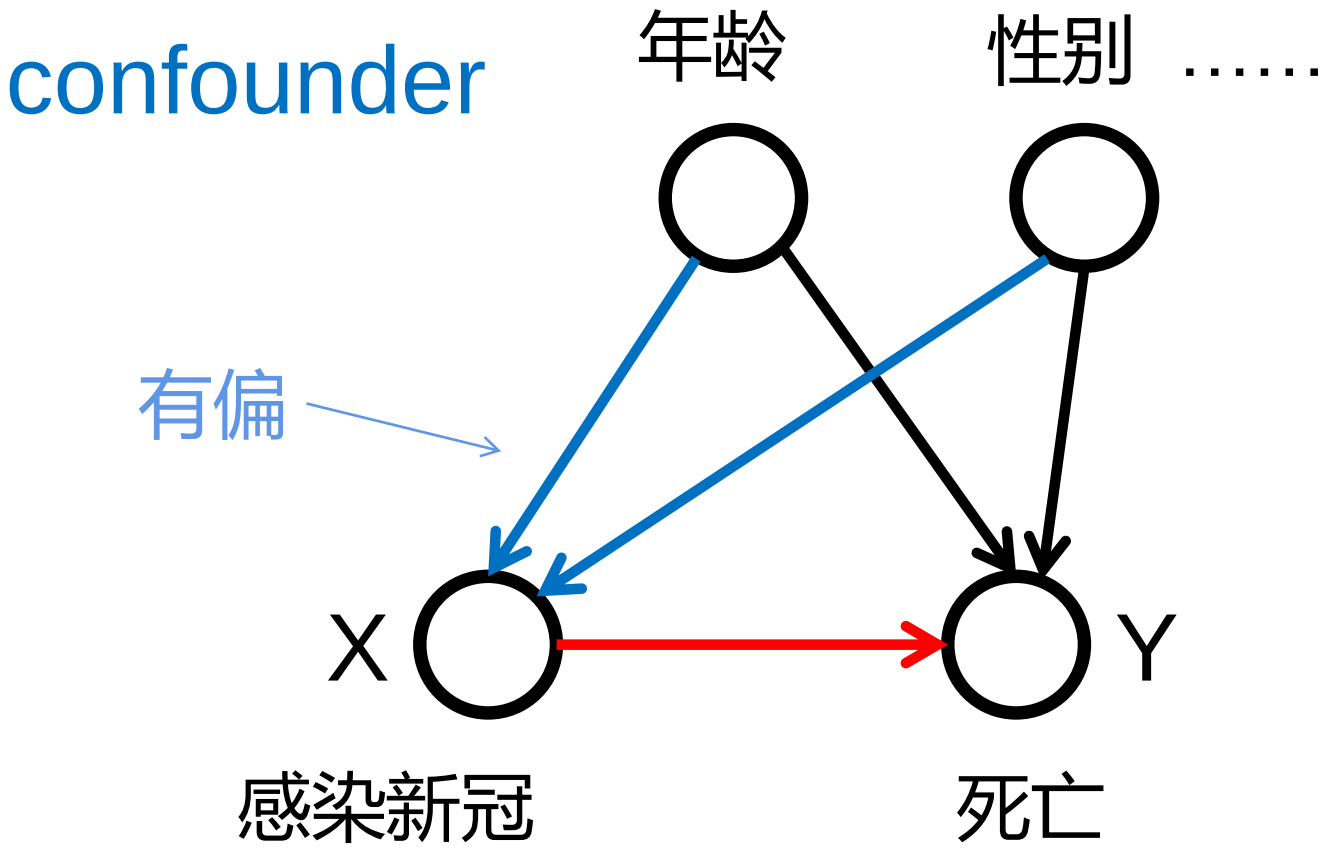
- 这种因果图被命名为“fork”（**叉子**），特点是有**两条边**从**一个变量**中分岔。

Contents



- 混杂因素可以有很多：年龄、性别等等，
  这些混杂因素导致对“**感染新冠**”和“**死亡**”的**直接**观测是有偏的。

Contents

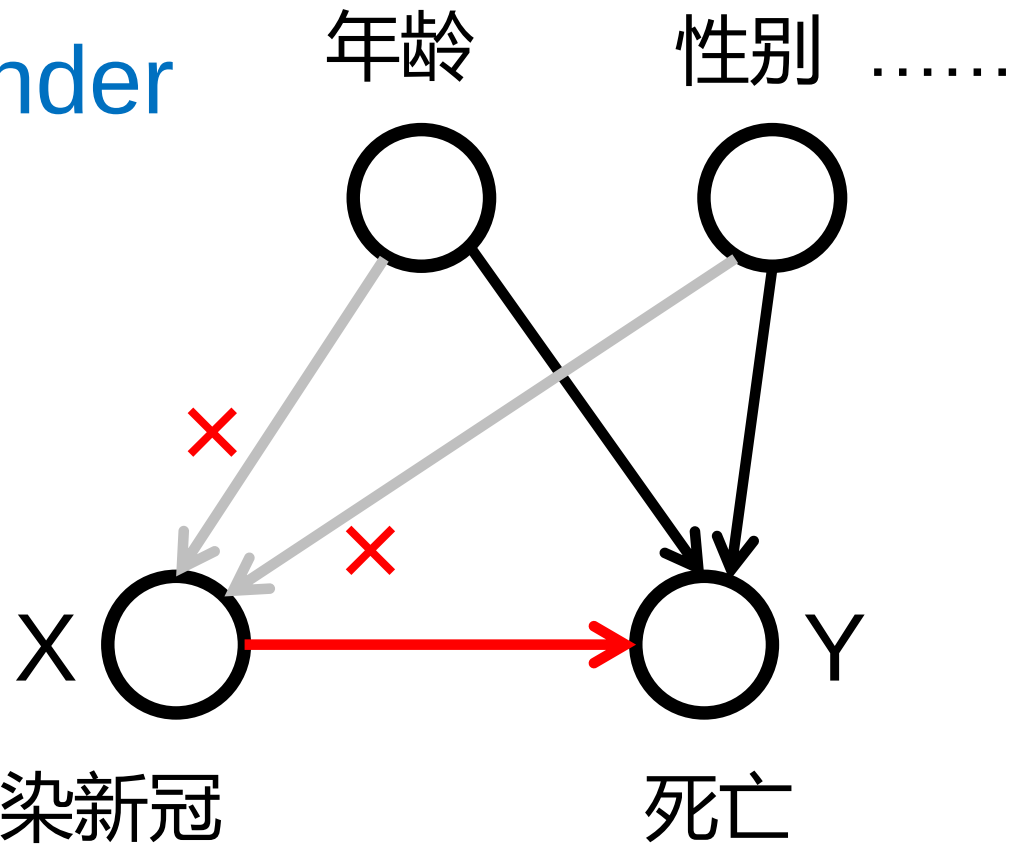


无偏?

- 混杂因素可以有很多：年龄、性别等等，
  这些混杂因素导致对“**感染新冠**”和“**死亡**”的**直接**观测是有偏的。

Causal Graph

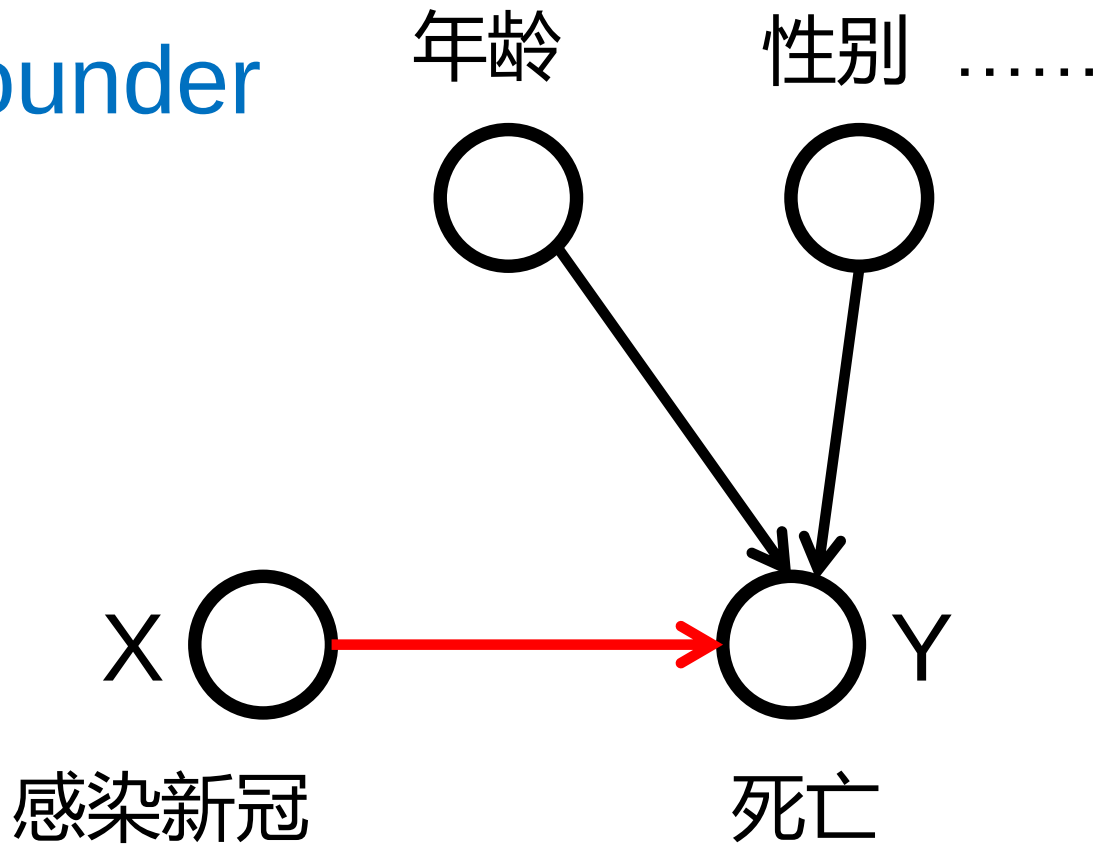
confounder



- 想要对X和Y的直接观测是**无偏**的，即X对Y的**直接影响**，只需删掉**进入**“X”的箭头。

Causal Graph

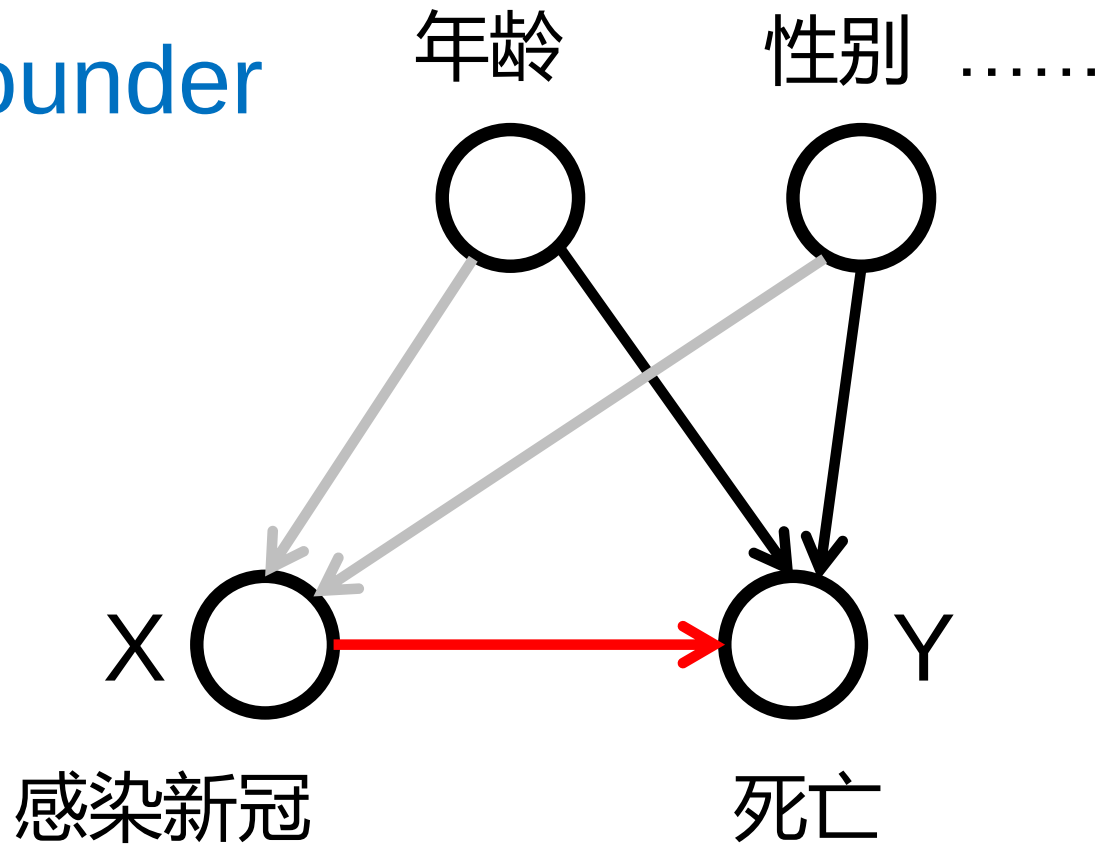
confounder



- 感染新冠、年龄、性别……之间相互**独立**，即
  对“感染新冠”和“死亡”的**直接**观测就是**无偏的**。

Causal Graph

confounder

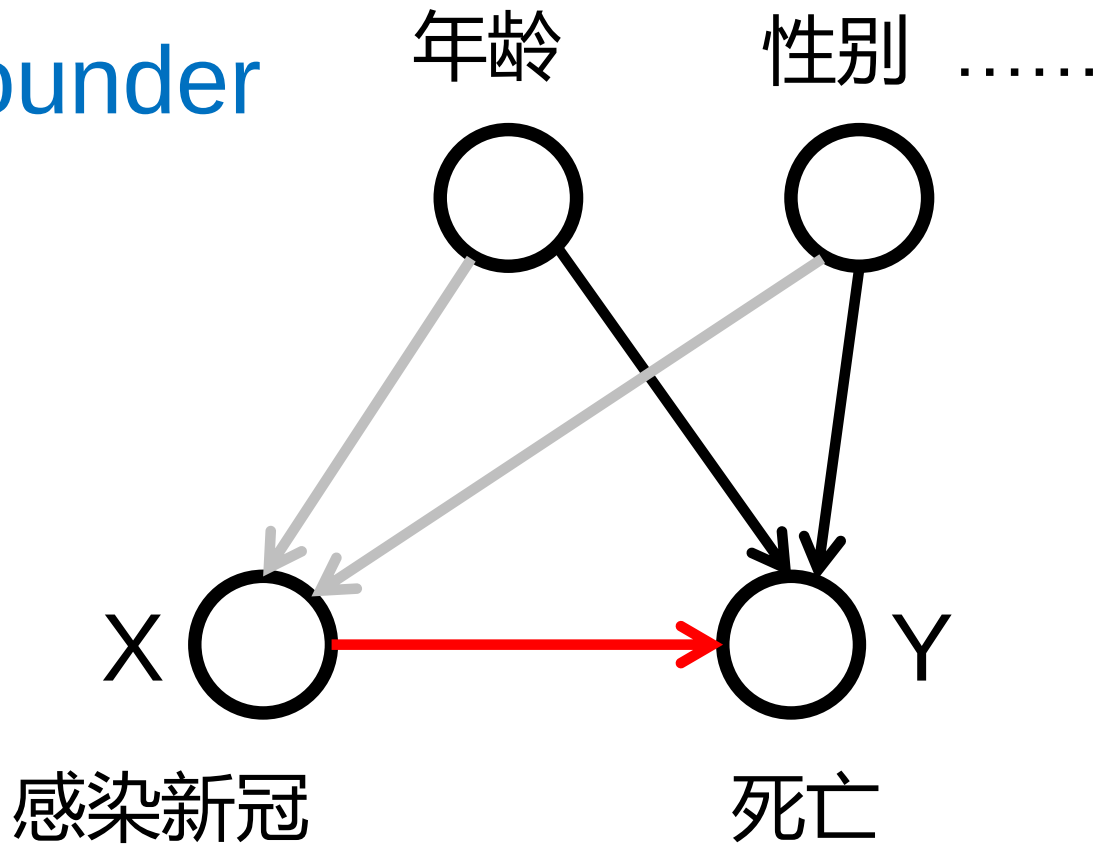


Method：

- 随机对照试验(RCT)
- do 算子

Causal Graph

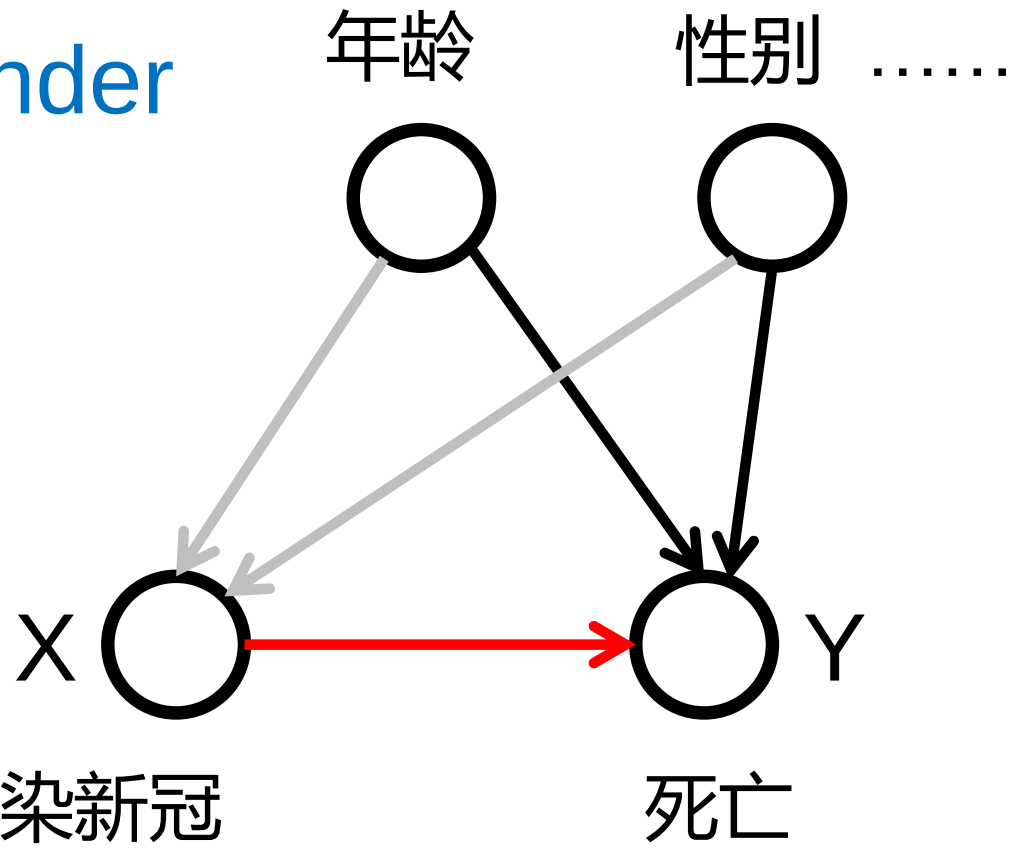
confounder



- 随机对照试验(RCT)：
  随机的从人群中选取样本，人为切断“confounder”到“感染新冠”之间的连接。

Causal Graph

confounder



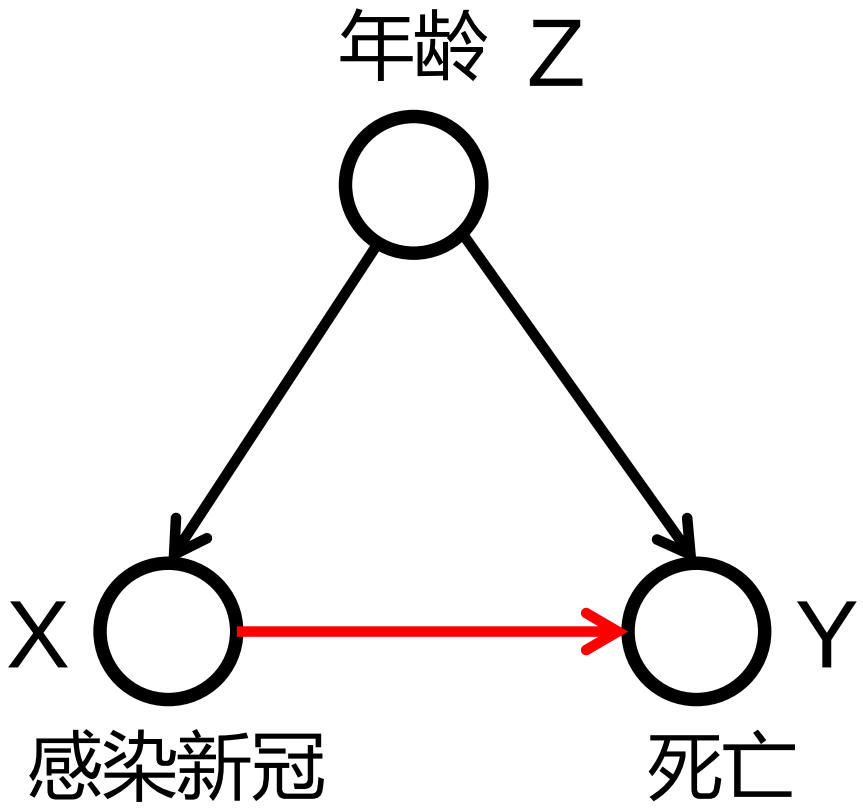
- do算子：简单来说就是“分情况讨论”，被称为后门调整(backdoor adjustment)

  例如：对年龄进行分层，在每个年龄段中分别统计“感染新冠”和“死亡率”的关系。

$$P(\text{死亡率}|\text{感染新冠}) = \sum P(\text{死亡率}|\text{感染新冠}, \text{年龄})P(\text{年龄})$$

De-confounder !

Causal Graph



- 保留混杂因子影响：

$$P(Y|X) = \sum_{i=1}^{n} P(Y|X, Z_i)P(Z_i|X)$$

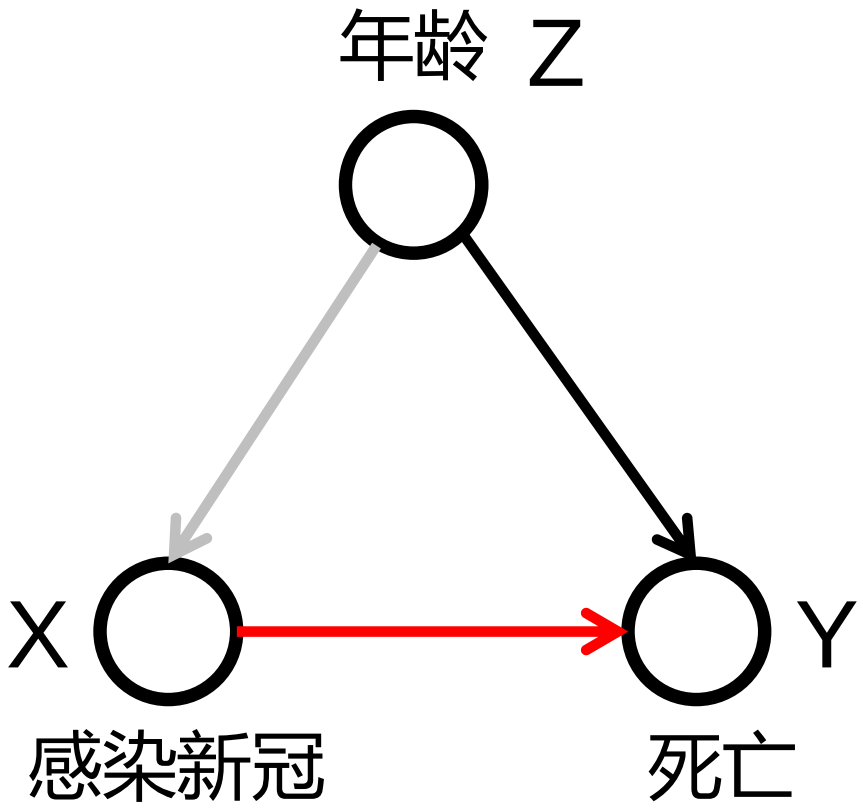
$$Q(Y) = \sum_{i=1}^{n} Q(Y|Z_i)Q(Z_i)$$

条件全概率

↑

全概率

Causal Graph



- 保留混杂因子影响：

$$P(Y|X) = \sum_{i=1}^{n} P(Y|X, Z_i)P(Z_i|X)$$

条件全概率

$$Q(Y) = \sum_{i=1}^{n} Q(Y|Z_i)Q(Z_i)$$

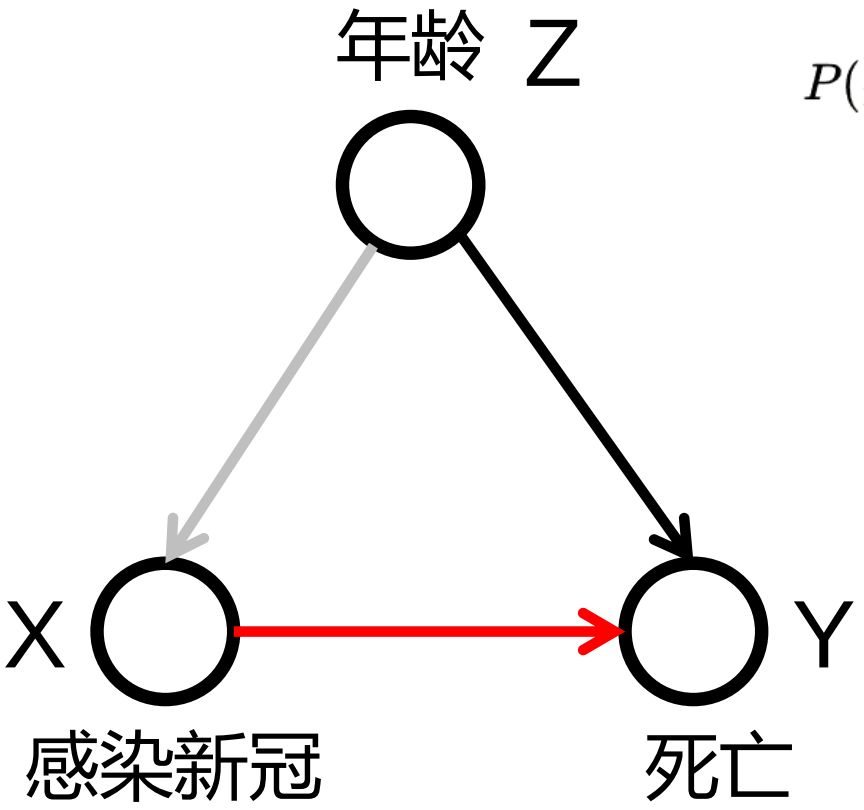
全概率

- 去除混杂因子影响：

$$P(Y|X) = \sum_{i=1}^{n} P(Y|X, Z_i)P(Z_i)$$

$$P(\text{死亡率}|\text{感染新冠}) = \sum P(\text{死亡率}|\text{感染新冠}, \text{年龄})P(\text{年龄})$$

Causal Graph



$$P(Y|X) = \sum_{i=1}^{n} P(Y|X, Z_i)P(Z_i)$$

$$P(\text{死亡率}|\text{感染新冠}) = \sum P(\text{死亡率}|\text{感染新冠}, \text{年龄})P(\text{年龄})$$

年龄Z	感染新冠人数(X)	死亡人数(Y)	$P(Y\|Z=\text{年龄}, X)$
$0 \sim 14$ yr	9	0	$0/9 = 0$
$15 \sim 49$ yr	557	12	$12/557 = 0.022$
$50 \sim 64$ yr	292	21	$21/292 = 0.072$
$>= 65$ yr	153	32	$32/153 = 0.209$
总人数	1011	65	$65/1011 = 0.064$

年龄Z	$0 \sim 14$ yr	$15 \sim 49$ yr	$50 \sim 64$ yr	$>= 65$ yr	汇总
人口数	221322621	773828163	218732927	118927158	1332810869
$P(Z)$	0.166	0.581	0.164	0.089	1.000

代入到后门调整公式可得:

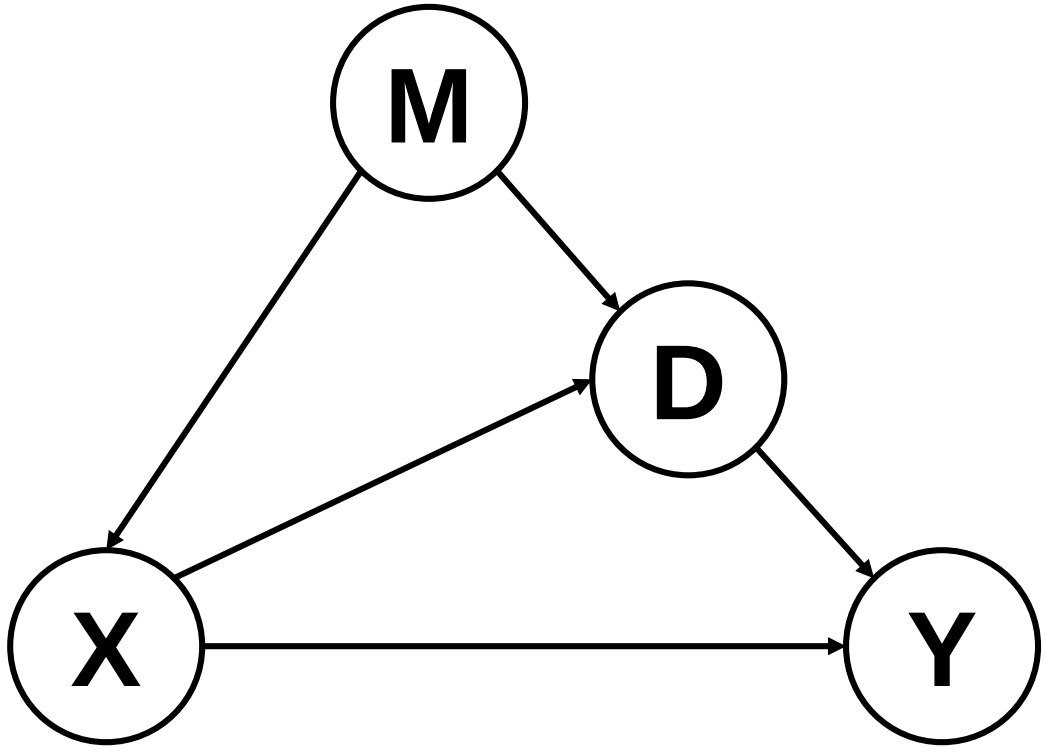
用do算子完成了一次后门调整

$$P(Y|X) = \sum P(Y|Z, X)P(Z) = 0 * 0.166 + , , , + 0.209 * 0.089 = 0.043$$

Contents

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

The Proposed Causal Graph



X : Feature

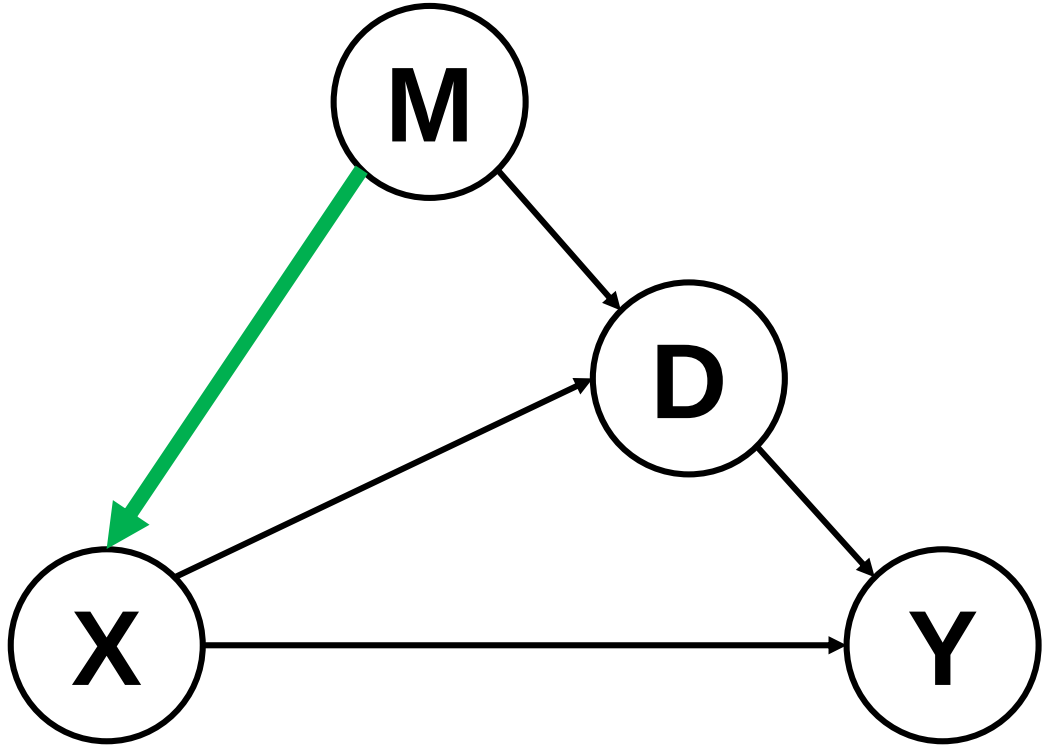
Y : Prediction

M: Momentum

D : Projection on Head

The Proposed Causal Graph

南京航空航天大學
Nanjing University of Aeronautics and Astronautics



X : Feature

Y : Prediction

M: Momentum

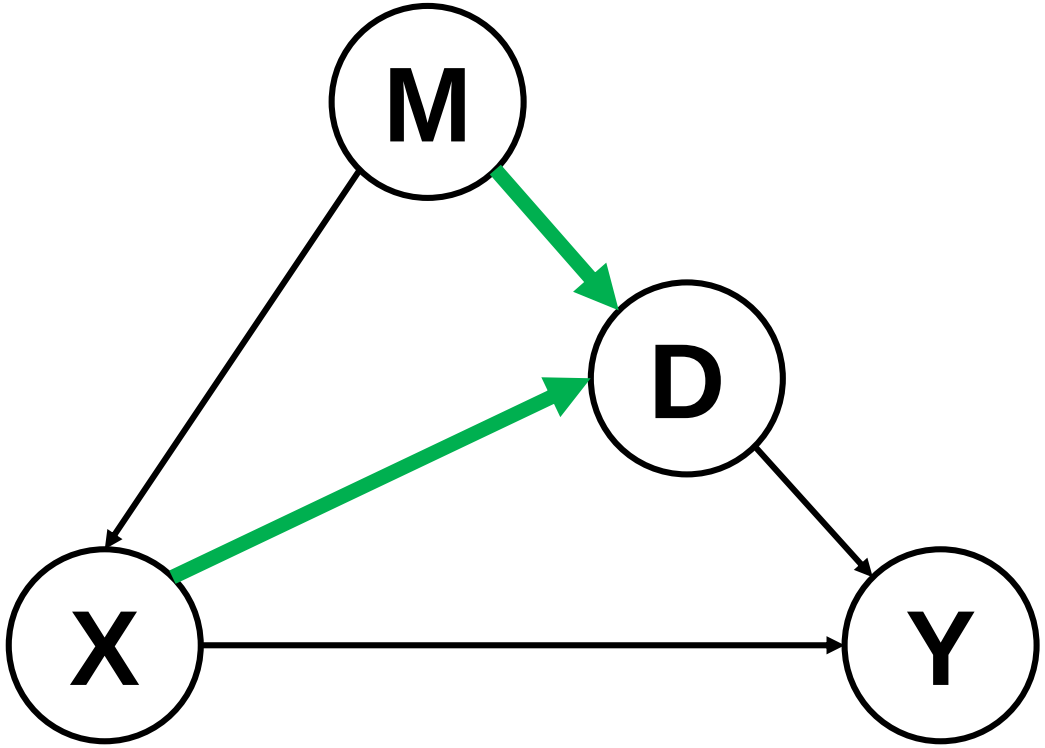
D : Projection on Head

- **M → X**
- (M, X) → D
- X → D → Y & X → Y

- Backbone parameters used to generate feature vectors **X**, are trained under the effect of **M**.
- The momentum **M** also causes feature vector **X** deviates to the head direction D, which is also determined by M.
- The effect of X can be disentangled into an indirect (mediation) and a direct effect.

The Proposed Causal Graph

南京航空航天大學
Nanjing University of Aeronautics and Astronautics



X : Feature

Y : Prediction

M: Momentum

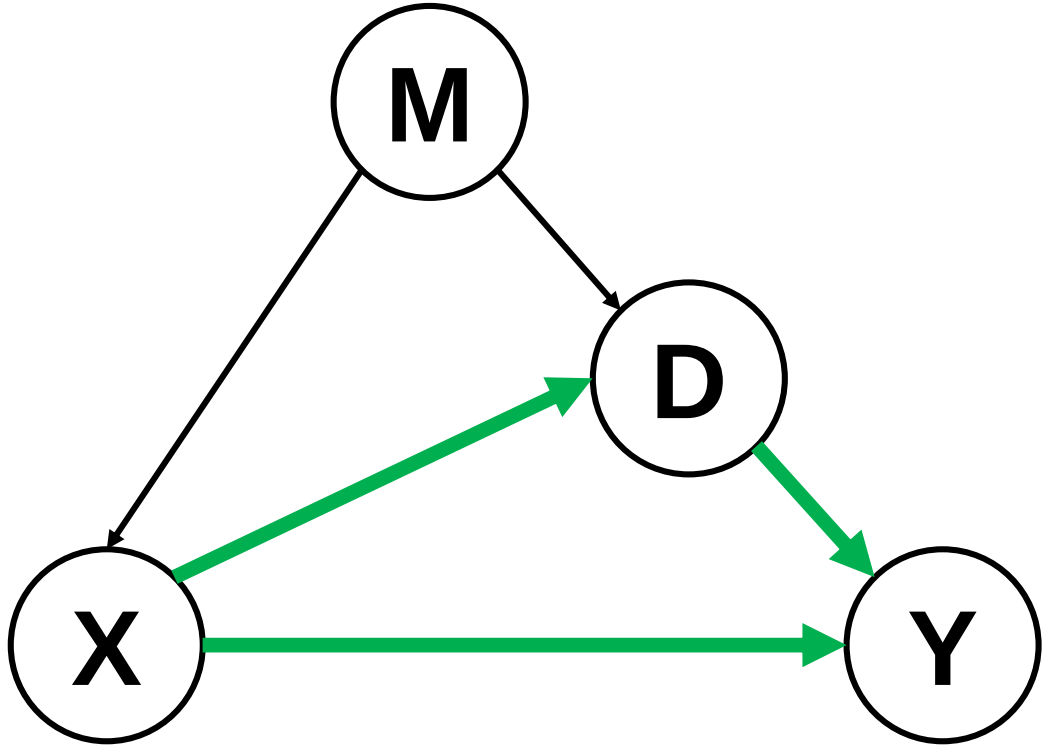
D : Projection on Head

- $M \rightarrow X$
- **$(M, X) \rightarrow D$**
- $X \rightarrow D \rightarrow Y \ \& \ X \rightarrow Y$

- Backbone parameters used to generate feature vectors **X**, are trained under the effect of **M**.
- The momentum **M** also causes feature vector **X** deviates to the head direction D, which is also determined by M.
- The effect of X can be disentangled into an indirect (mediation) and a direct effect.

The Proposed Causal Graph

南京航空航天大學
Nanjing University of Aeronautics and Astronautics



X : Feature

Y : Prediction

M: Momentum

D : Projection on Head

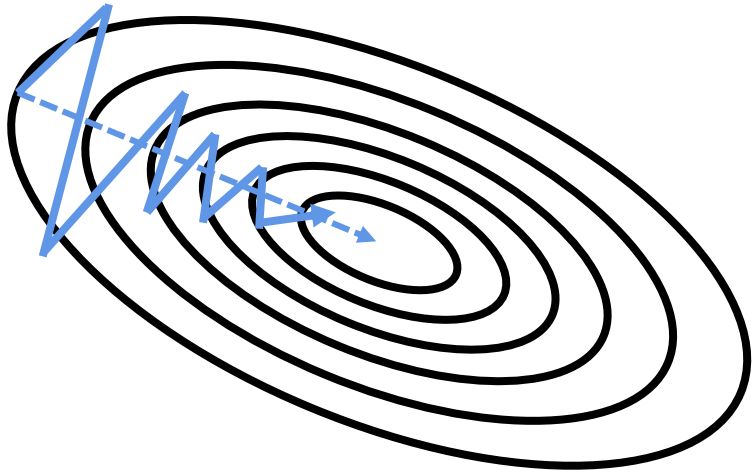
- $M \rightarrow X$
- $(M, X) \rightarrow D$
- **$X \rightarrow D \rightarrow Y$ & $X \rightarrow Y$**

- Backbone parameters used to generate feature vectors X, are trained under the effect of M.
- The momentum **M** also causes feature vector **X** deviates to the head direction D, which is also determined by M.
- The effect of X can be disentangled into an indirect (mediation) and a direct effect.

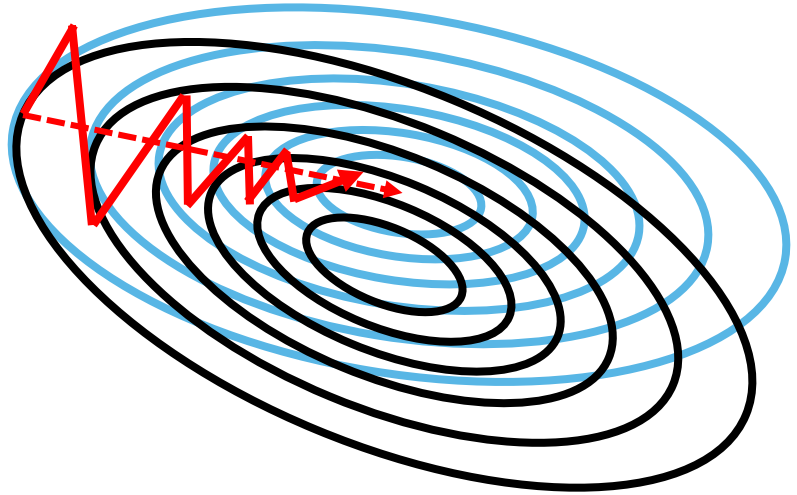
Accumulative Momentum Effect

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Pytorch implementation: $v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$



SGD Momentum in
Balanced Dataset



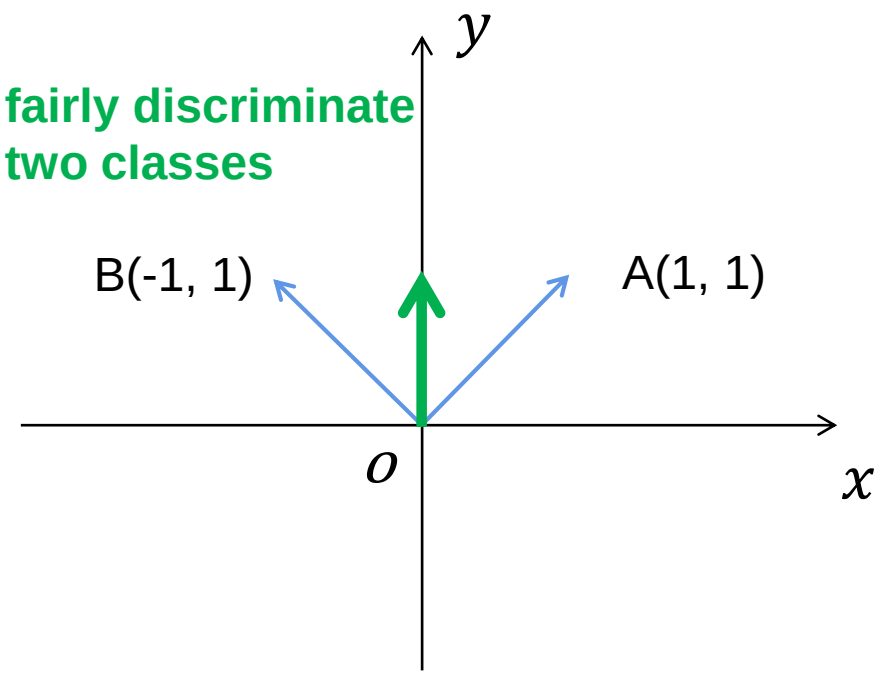
SGD Momentum in
Long-Tailed Dataset

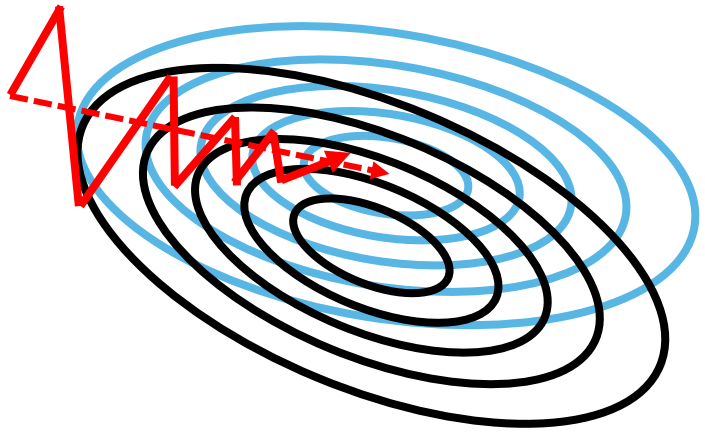
- Global Optima for All Categories
- Local Optima for Head Categories
- Momentum Direction in Balanced Data
- Momentum Direction in Long-Tailed Data

Accumulative Momentum Effect

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Pytorch implementation: $v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$





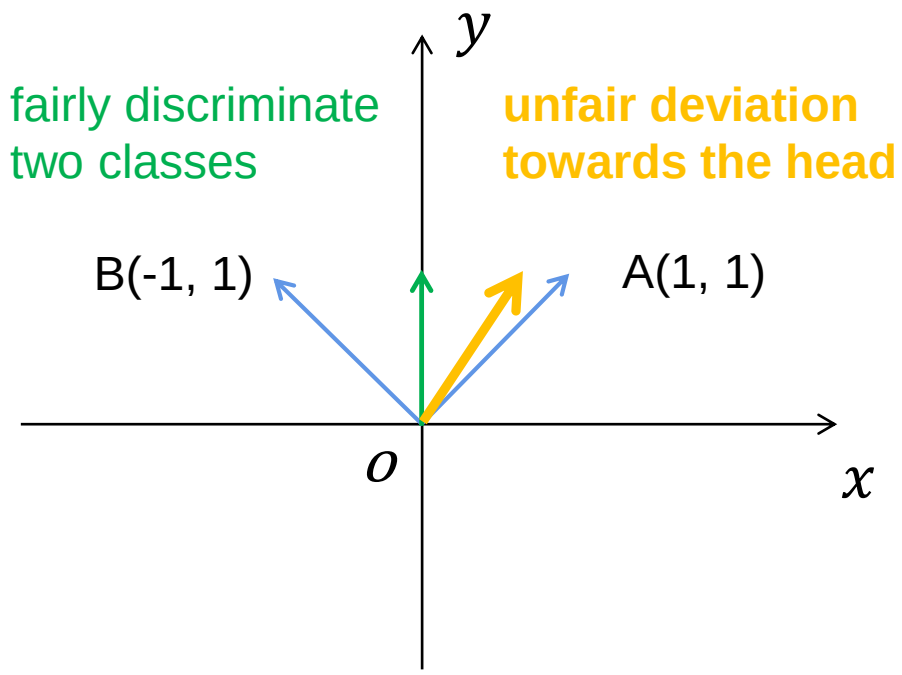
SGD Momentum in
Long-Tailed Dataset

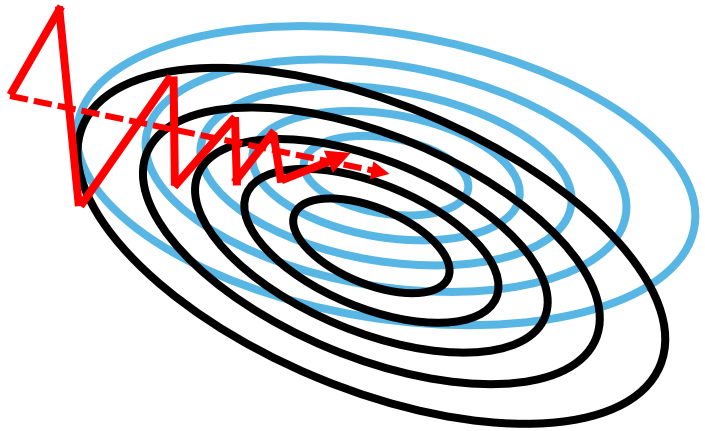
- Global Optima for All Categories
- Local Optima for Head Categories
- Momentum Direction in Balanced Data
- Momentum Direction in Long-Tailed Data

Accumulative Momentum Effect

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Pytorch implementation: $v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$





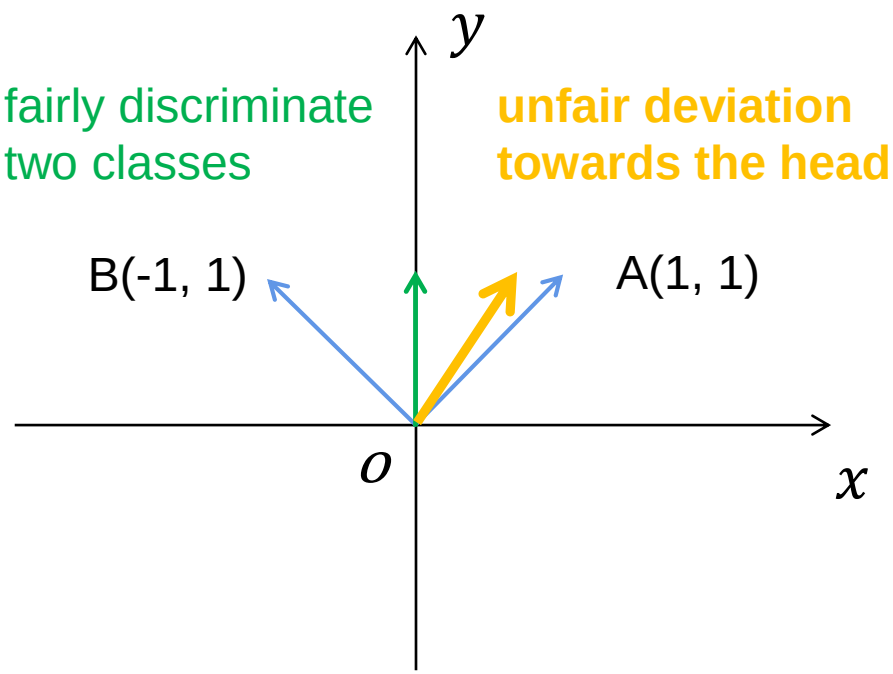
SGD Momentum in
Long-Tailed Dataset

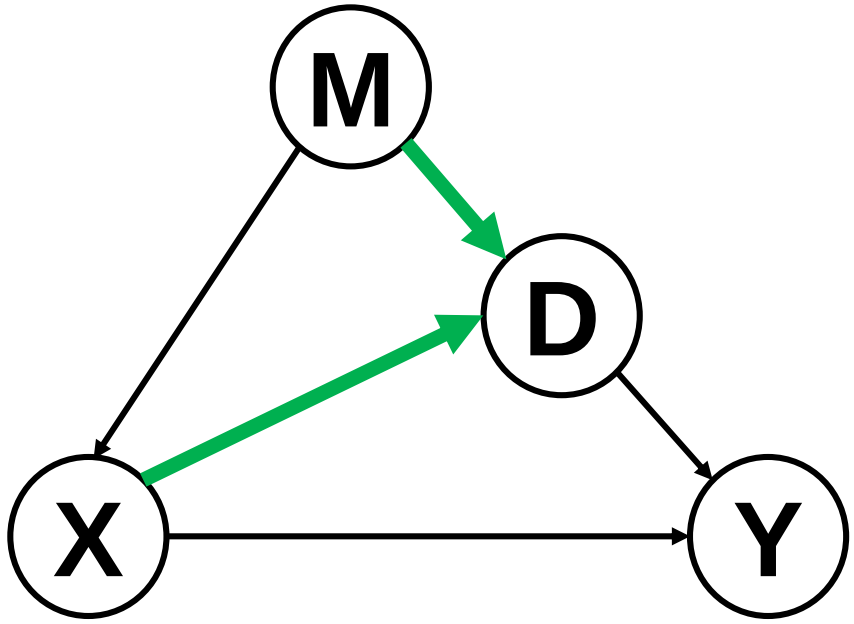
- Global Optima for All Categories
- Local Optima for Head Categories
- Momentum Direction in Balanced Data
- Momentum Direction in Long-Tailed Data

Accumulative Momentum Effect

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Pytorch implementation: $v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$





- **(M, X) → D**

- The momentum **M** also causes feature vector **X** deviates to the head direction D, which is also determined by M.

Accumulative Momentum Effect

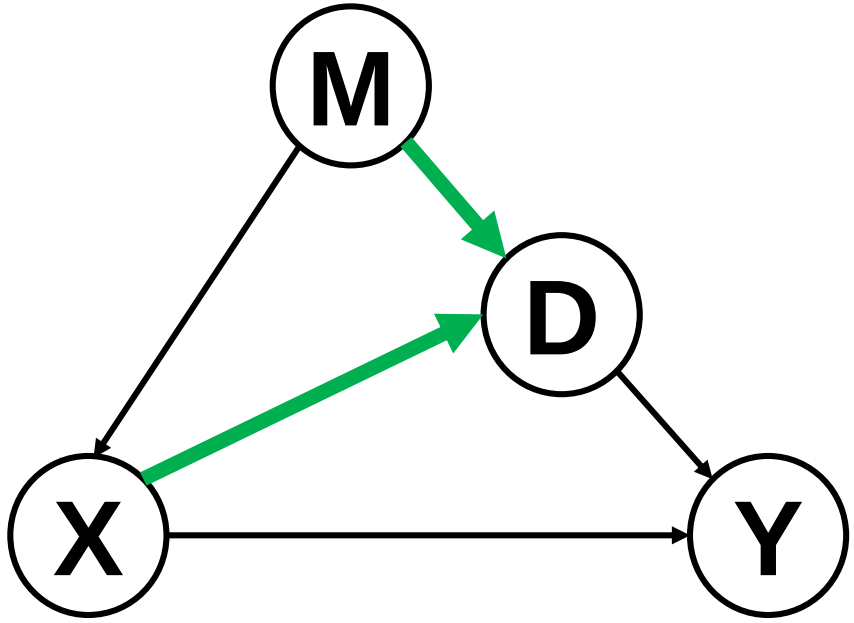
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Pytorch implementation: $v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$

Assumption 1: *The head direction $\hat{d}$ is the unit vector of the exponential moving average features with decay rate μ:*

$$\hat{\boldsymbol{d}} = \overline{\boldsymbol{x}}_T / \|\overline{\boldsymbol{x}}_T\|$$

$$where\ \overline{\boldsymbol{x}}_t = \mu \cdot \overline{\boldsymbol{x}}_{t-1} + \boldsymbol{x}_t$$



- $\hat{d}$ is determined by the sample moving average in the dataset, which does not need the accessibility of the class statistics at all.

- **(M, X) → D**

- The momentum **M** also causes feature vector **X** deviates to the head direction D, which is also determined by M.

Accumulative Momentum Effect

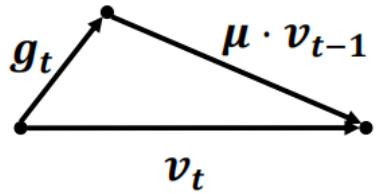
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Pytorch implementation: $v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \quad \theta_t = \theta_{t-1} - lr \cdot v_t,$

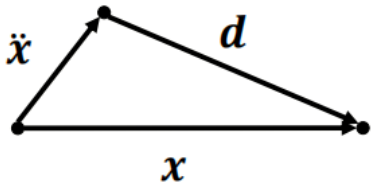
Assumption 1: *The head direction $\hat{d}$ is the unit vector of the exponential moving average features with decay rate μ:*

$$\hat{\boldsymbol{d}} = \overline{\boldsymbol{x}}_T / \|\overline{\boldsymbol{x}}_T\|$$

$$where\ \overline{\boldsymbol{x}}_t = \mu \cdot \overline{\boldsymbol{x}}_{t-1} + \boldsymbol{x}_t$$



(a) Decompose the gradient velocity

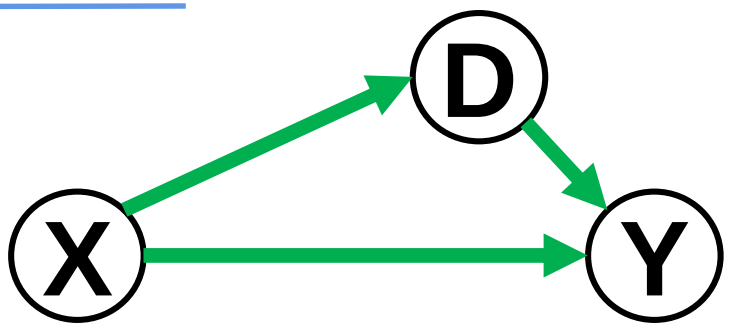


(b) Decompose the biased feature vector

Figure 2: Based on Assumption 1, the feature vector $\boldsymbol{x}$ can be decomposed into a discriminative feature $\ddot{\boldsymbol{x}}$ and a projection on head direction $\boldsymbol{d}$

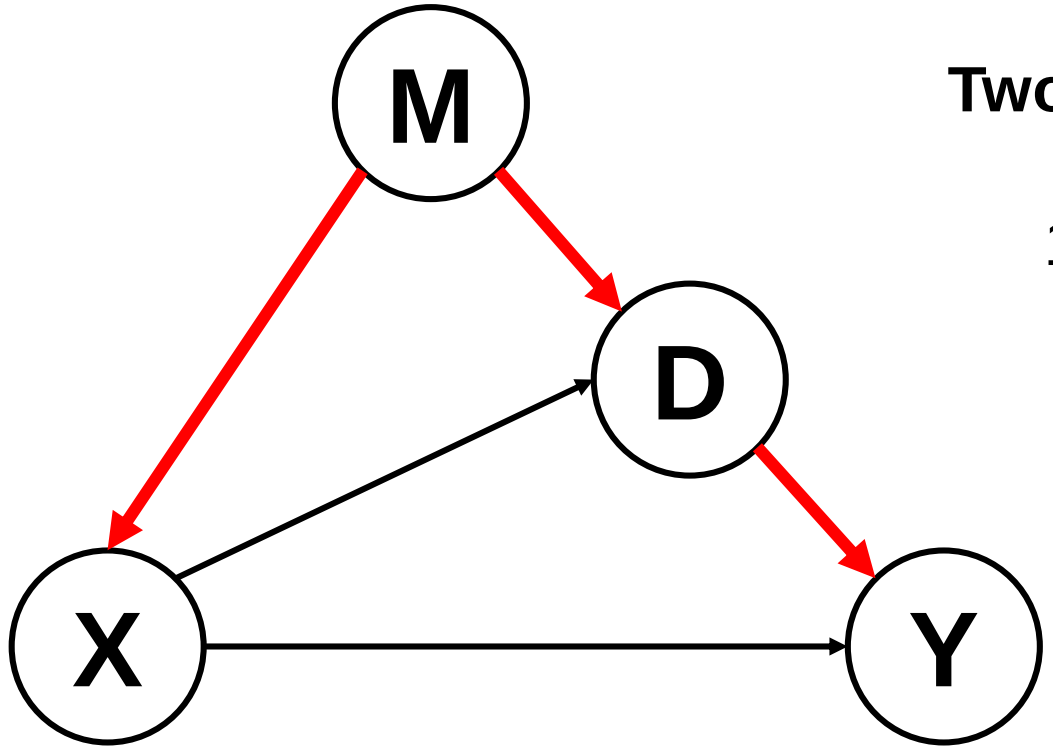
$$\boldsymbol{x} = \ddot{\boldsymbol{x}} + \boldsymbol{d}, \text{ where } D = \boldsymbol{d} = \hat{\boldsymbol{d}} cos(\boldsymbol{x}, \hat{\boldsymbol{d}}) \|\boldsymbol{x}\|$$

direct effect　　indirect effect



X → D → Y & X → Y

confounder



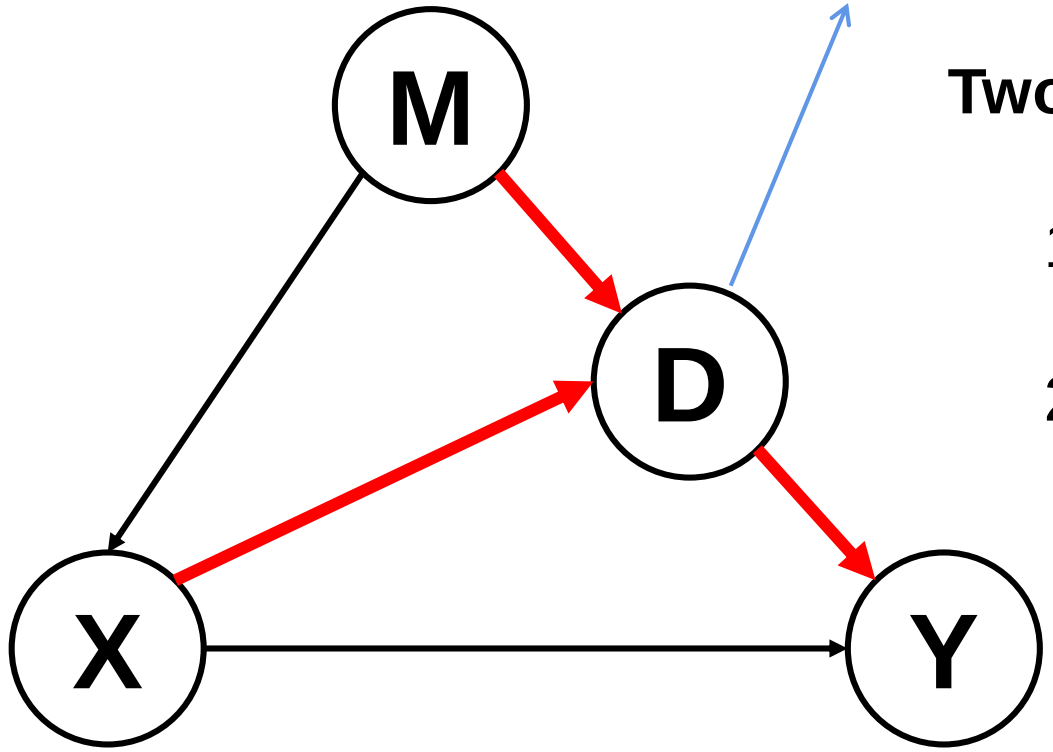
Two Undesired Causal Effects of Momentum:

1. Backdoor shortcut

Two Undesired Causal Effects of Momentum

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

mediator:中介



Two Undesired Causal Effects of Momentum:

1. Backdoor shortcut
2. Indirect Mediator Effect

Contents

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Long-Tailed Classification
- The Proposed Causal Graph
- De-confound TDE
- Experiments

De-confound TDE Classifier

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

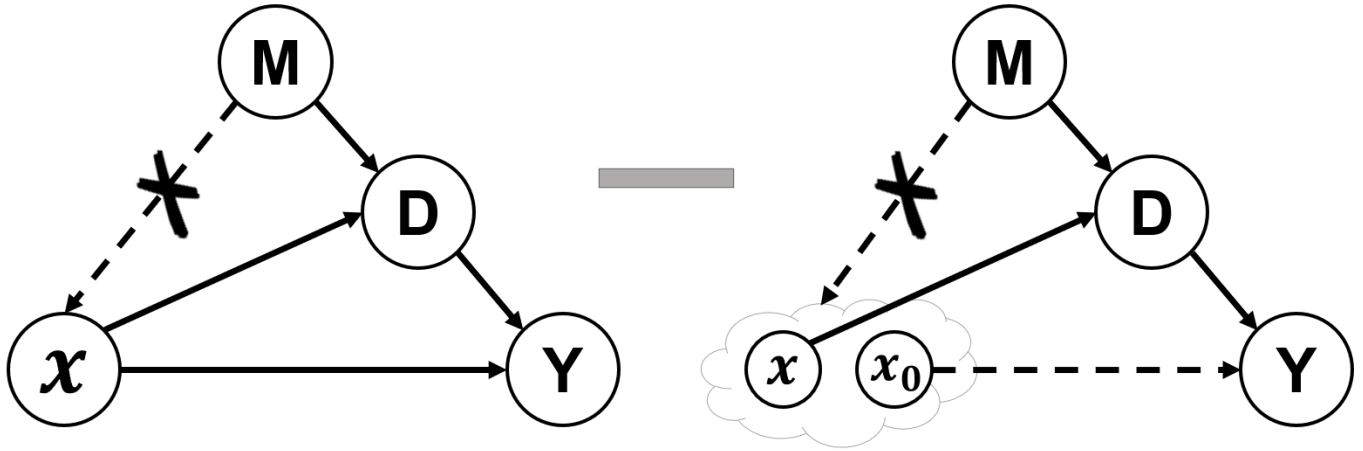
The definition of Total Direct Effect (TDE):

$$argmax_{i \in C}\, TDE(Y_i) = [Y_d = i | do(X = x)] - [Y_d = i | do(X = x_0)]$$

- Subscript d denotes that the mediator D always takes the value d, x_0=0

The proposed de-confound TDE:

$$TDE(Y_i) = \frac{\tau}{K} \sum_{k=1}^{K} \left(\frac{\left(w_i^k\right)^T x^k}{\left(\|w_i^k\| + \gamma\right)\|x^k\|} - \alpha \cdot \frac{\cos\left(x^k, \hat{d}^k\right) \cdot \left(w_i^k\right)^T \hat{d}^k}{\left(\|w_i^k\| + \gamma\right)} \right)$$



De-confound Training

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

$$P(Y=i|do(X=\boldsymbol{x}))=\sum_{\boldsymbol{m}}P(Y=i|X=\boldsymbol{x},M=\boldsymbol{m})P(M=\boldsymbol{m}) \tag{3}$$

infinite number of M = m

Inverse Probability Weighting

$$=\sum_{\boldsymbol{m}}\frac{P(Y=i,X=\boldsymbol{x}|M=\boldsymbol{m})P(M=\boldsymbol{m})}{P(X=\boldsymbol{x}|M=\boldsymbol{m})}. \tag{4}$$

no matter how many *m* there are, we can only observe one (*i*, *x*) given one *m*.

$$\approx\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}(Y=i,X=\boldsymbol{x}^k|M=\boldsymbol{m}), \tag{5}$$

where $\widetilde{P}$ is the inverse weighted probability

K times more fine-grained sampling

$$\widetilde{P}(Y=i,X=\boldsymbol{x}^k)\propto E(i,\boldsymbol{x}^k;\boldsymbol{w}_i^k)=\tau\frac{f(i,\boldsymbol{x}^k;\boldsymbol{w}_i^k)}{g(i,\boldsymbol{x}^k;\boldsymbol{w}_i^k)},$$

energy-based model

$$P(Y=i|do(X=\boldsymbol{x}))=\frac{\tau}{K}\sum_{k=1}^{K}\frac{(\boldsymbol{w}_i^k)^\top(\ddot{\boldsymbol{x}}^k+\boldsymbol{d}^k)}{(\|\boldsymbol{w}_i^k\|+|\gamma)\|\boldsymbol{x}^k\|}=\frac{\tau}{K}\sum_{k=1}^{K}\frac{(\boldsymbol{w}_i^k)^\top\boldsymbol{x}^k}{(\|\boldsymbol{w}_i^k\|+\gamma)\|\boldsymbol{x}^k\|}. \tag{7}$$

Total Direct Effect Inference

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

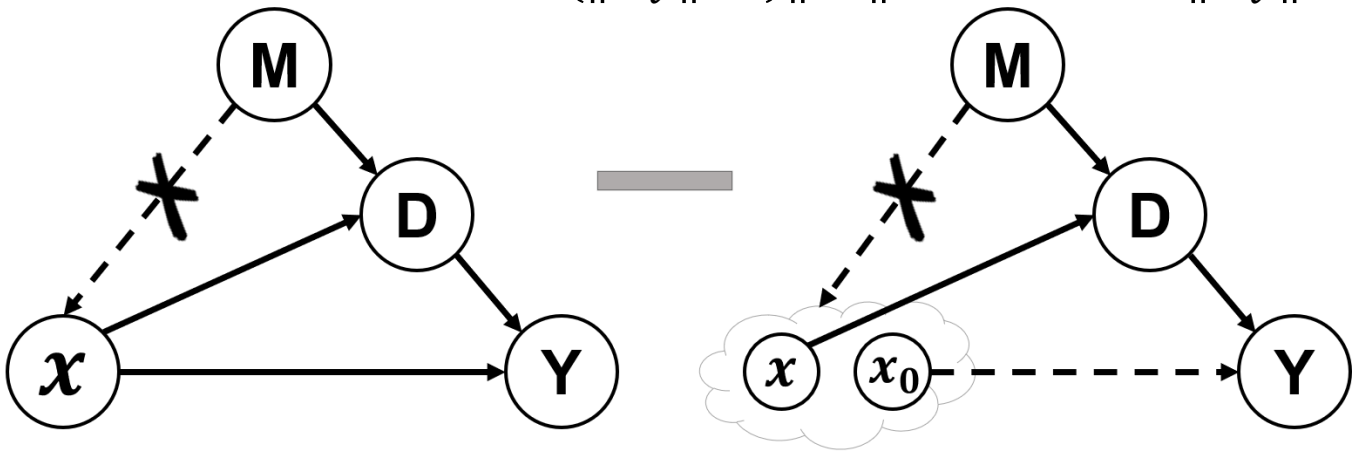
$$P(Y=i|do(X=\boldsymbol{x})) = \frac{\tau}{K}\sum_{k=1}^{K}\frac{(\boldsymbol{w}_i^k)^\top(\ddot{\boldsymbol{x}}^k+\boldsymbol{d}^k)}{(\|\boldsymbol{w}_i^k\|+|\gamma)\|\boldsymbol{x}^k\|} = \frac{\tau}{K}\sum_{k=1}^{K}\frac{(\boldsymbol{w}_i^k)^\top\boldsymbol{x}^k}{(\|\boldsymbol{w}_i^k\|+\gamma)\|\boldsymbol{x}^k\|}. \quad (7)$$

The definition of Total Direct Effect (TDE):

$$argmax_{i\in C}\, TDE(Y_i) = [Y_d = i|do(X=x)] - [Y_d = i|do(X=x_0)]$$

The proposed de-confound TDE:

$$TDE(Y_i) = \frac{\tau}{K}\sum_{k=1}^{K}\left(\frac{\left(w_i^k\right)^T x^k}{(\|w_i^k\|+\gamma)\|x^k\|} - \alpha\cdot\frac{\cos(x^k,\hat{d}^k)\cdot\left(w_i^k\right)^T\hat{d}^k}{(\|w_i^k\|+\gamma)}\right)$$



$$\boldsymbol{x_0} = \boldsymbol{0}$$
$$\boldsymbol{x} = \ddot{\boldsymbol{x}} + \boldsymbol{d},$$
$$\boldsymbol{d} = \|\boldsymbol{d}\|\cdot\hat{\boldsymbol{d}} = cos(\boldsymbol{x},\hat{\boldsymbol{d}})\|\boldsymbol{x}\|\cdot\hat{\boldsymbol{d}}.$$

Contents

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Long-Tailed Classification
- The Proposed Causal Graph
- De-confound TDE
- Experiments

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Methods	Many-shot	Medium-shot	Few-shot	Overall
Focal Loss† [28]	64.3	37.1	8.2	43.7
OLTR† [9]	51.0	40.8	20.8	41.9
Decouple-OLTR† [9, 11]	59.9	45.8	27.6	48.7
Decouple-Joint [11]	65.9	37.5	7.7	44.4
Decouple-NCM [11]	56.6	45.3	28.1	47.3
Decouple-cRT [11]	61.8	46.2	27.4	49.6
Decouple-τ-norm [11]	59.1	46.9	30.7	49.4
Decouple-LWS [11]	60.2	47.2	30.3	49.9
Baseline	66.1	38.4	8.9	45.0
Cosine† [50, 51]	67.3	41.3	14.0	47.6
Capsule† [9, 54]	67.1	40.0	11.2	46.5
(Ours) De-confound	**67.9**	42.7	14.7	48.6
(Ours) Cosine-TDE	61.8	47.1	30.4	50.5
(Ours) Capsule-TDE	62.3	46.9	30.6	50.6
(Ours) De-confound-TDE	62.7	**48.8**	**31.6**	**51.8**

Table 2: The performances on ImageNet-LT test set [9]. All models were using the ResNeXt-50 backbone. The superscript † denotes being re-implemented by our framework and hyper-parameters.

Experiments

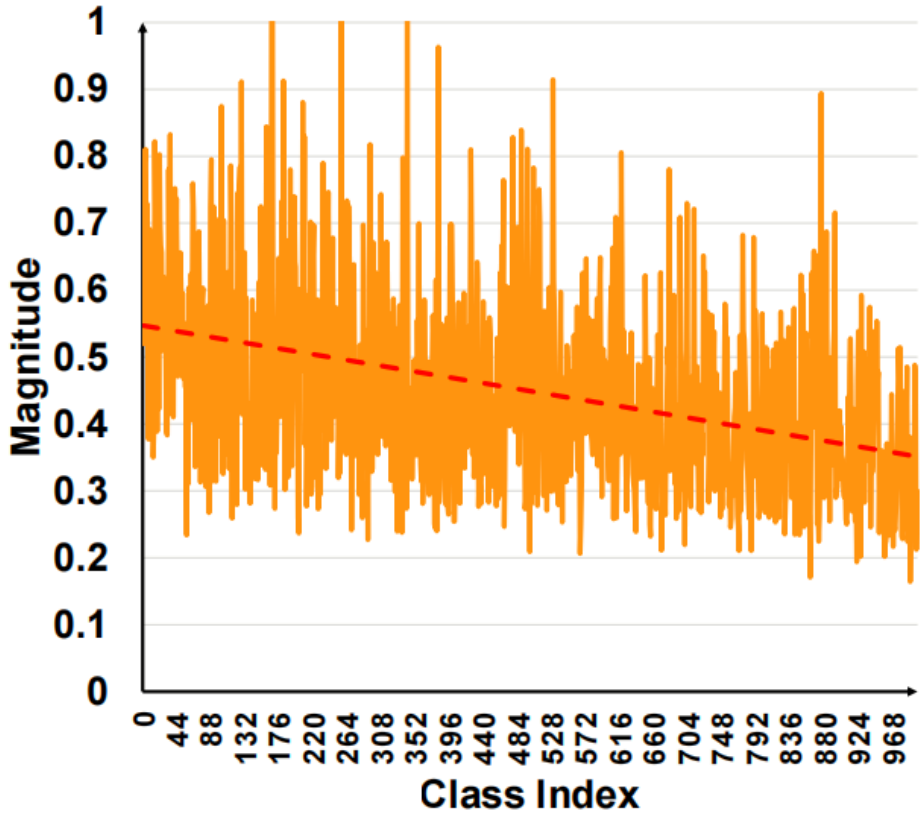
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Dataset	Long-tailed CIFAR-100			Long-tailed CIFAR-10		
Imbalance ratio	100	50	10	100	50	10
Focal Loss [28]	38.4	44.3	55.8	70.4	76.7	86.7
Mixup [56]	39.5	45.0	58.0	73.1	77.8	87.1
Class-balanced Loss [13]	39.6	45.2	58.0	74.6	79.3	87.1
LDAM [12]	42.0	46.6	58.7	77.0	81.0	88.2
BBN [10]	42.6	47.0	59.1	79.8	82.2	88.3
(Ours) De-confound	40.5	46.2	58.9	71.7	77.8	86.8
(Ours) De-confound-TDE	**44.1**	**50.3**	**59.6**	**80.6**	**83.6**	**88.5**

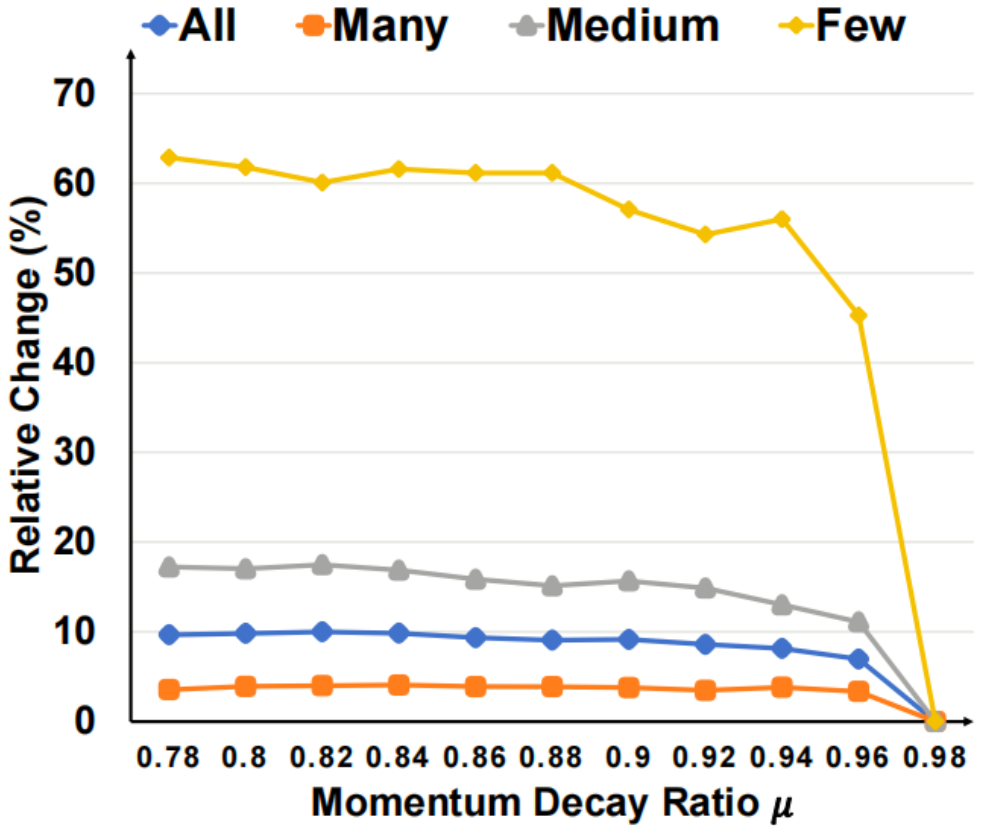
Table 3: Top-1 accuracy on Long-tailed CIFAR-10/-100 with different imbalance ratios. All models are using the same ResNet-32 backbone. We further adopted the same warm-up scheduler from BBN [10] for fair comparisons.

How momentum M affects X and Y?

南京航空航天大學
Nanjing University of Aeronautics and Astronautics



(b) Mean magnitude of x for each class i



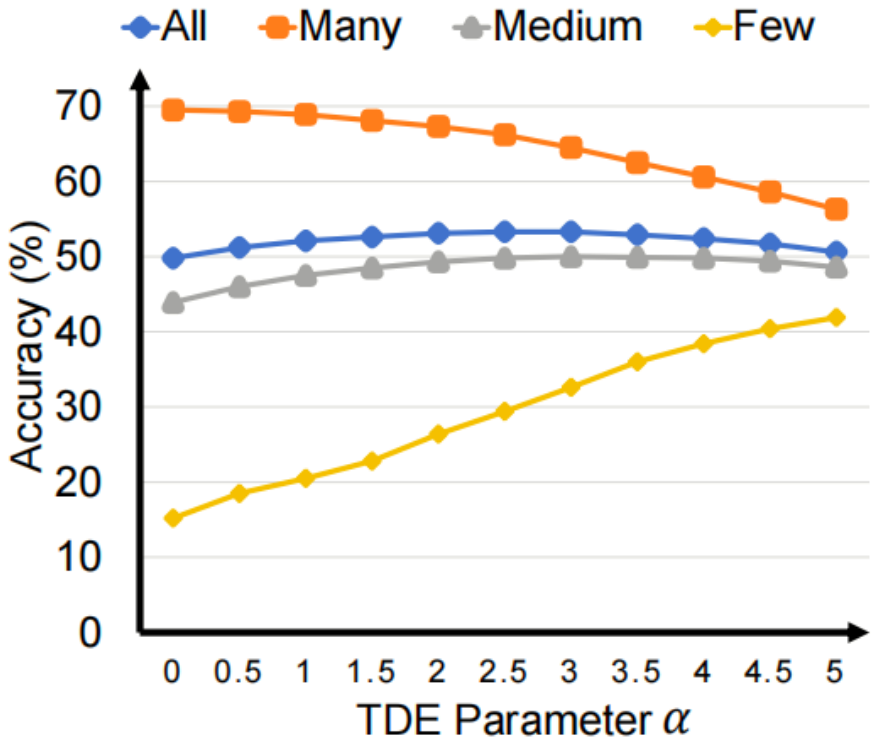
(c) Relative Change of Accuracy from μ=0.98

(b) The mean magnitudes of **feature vectors** for each class i after training with momentum μ = 0.9, where i is ranking from head to tail.

(c) The relative change of the performance on the basis of μ = 0.98 shows that the few-shot tail is more vulnerable to the momentum.

Ablation of α

南京航空航天大學
Nanjing University of Aeronautics and Astronautics



(a) Accuracy for different TDE parameter α

Figure 4: The influence of parameter α in Eq. (8) on ImageNet-LT val set [9] shows how D controls the head/tail preference.

THANKS