论文地址：

<https://arxiv.org/abs/1712.09491>

项目地址：

<https://github.com/sunblaze-ucb/blackbox-attacks>

# **1.论文核心思想**：

目前存在对DNN的黑盒攻击都是基于对抗攻击的迁移性的。所谓对抗攻击迁移性是指基于一个“本地”模型导出的对抗样本同样会使得其他模型失效。在本文中，作者提出一个基于梯度估计的黑盒攻击方法，这种方法不依赖与迁移性，而是不断的询问目标分类器的分类概率，通过这个概率来估计梯度。

# 2.论文核心贡献：

1. 提出基于Query的梯度估计黑盒攻击

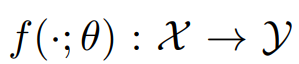
2. 使用PCA和特征分组的方法将所需的query 降低

3. 攻击真实的AI系统：Clarifai，内容监管系统

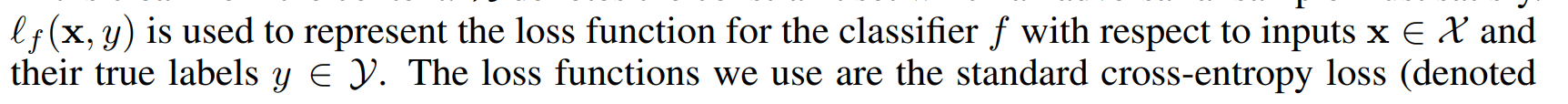
4. 跟其他黑盒攻击做对比；基于query的实验效果最好，同时达到和白盒攻击类似的攻击成功率。

5. 基于query的黑盒攻击，对对抗训练也依然有很大的攻击性。

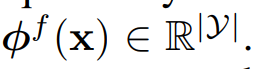
# 3.符号标记

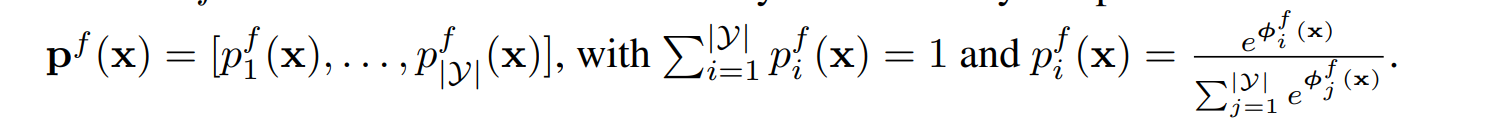
：分类器，X是输入空间，y是标签空间。是模型参数。



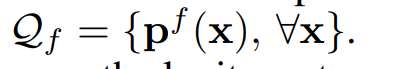
损失函数lf(x,y),表示在分类器f下，输入x和真是标签y之间的误差。

Xadv :对抗样本，是往一个benign样本X中加入一个微小的扰动产生的。这种Xadv要么会误导f把它分类为目标标签，或者任意一个不是正确的标签。

：对于输入x,神经网络f倒数第二层的logit。

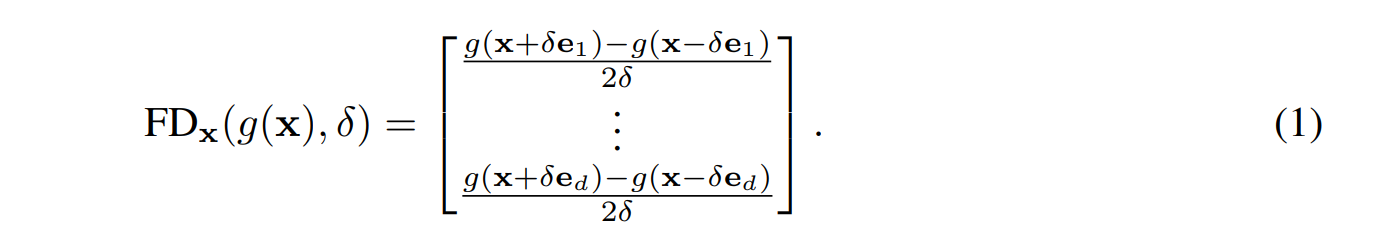
:最后一层是基于softmax的概率输出层。

# 4.威胁模型：

假设攻击者能够获得任意输入x的输出概率向量，攻击者能够询问的请求集合为：，通过取softmax输出概率的对数，攻击者也能够求出logits.

# 5.基于有限差分方法估计梯度：

假设需要估计梯度的目标函数为g(x),这个函数的输入是一个d维的向量。ei是x的基向量。g关于输入x的两步有限差分梯度估计可以表示为：

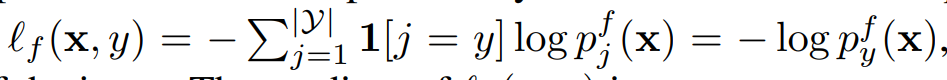


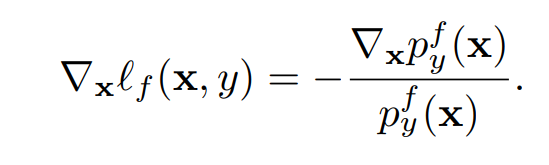
是控制梯度估计精度的超参数。如果g关于x的导数是存在的，那么必然有：



通过这种有限差分方法，攻击者可以通过访问x对应的函数值就可以得到对函数梯度的估计。

## 5.1 使用有限差分估计FGS

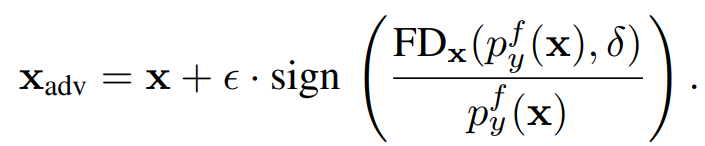
在FGS中，一般损失函数为交叉熵函数，其对应的损失函数为即取出真实标签那个分量的概率值，然后取对数。于是损失函数的梯度为：



 是攻击者通过query就可以拿到的，但是它关于x的梯度却是要估计的：。它的有限差分估计为：

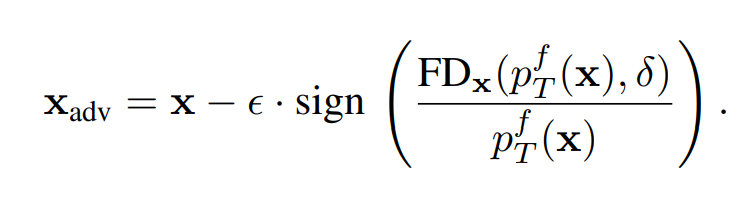


于是对抗样本Xadv:



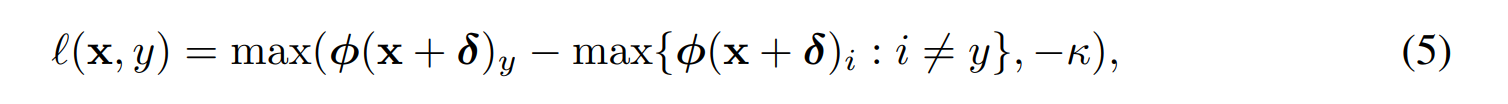
注意 是一个数，而不是向量，这种无目标生成对抗样本的方法叫做FD-xent。

有目标的梯度估计方法为，这种方法让Xadv的标签越来越接近T。

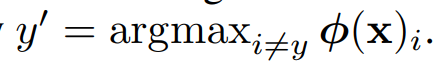
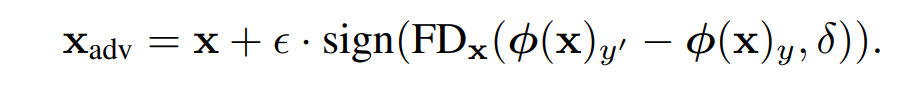


## 5.2 估计Logit 损失

Carlin和Wagner提出一个用于抵御白盒攻击的logit损失，据说特别好使：



其中 表示往输入添加一点小噪声后，输出logit里面除了真实标签最大那个logit。这个i对应的logit意味着最有可能分错。表示真实的标签距离分错的位置的logit的程度。 控制这个偏离程度最大为-k。

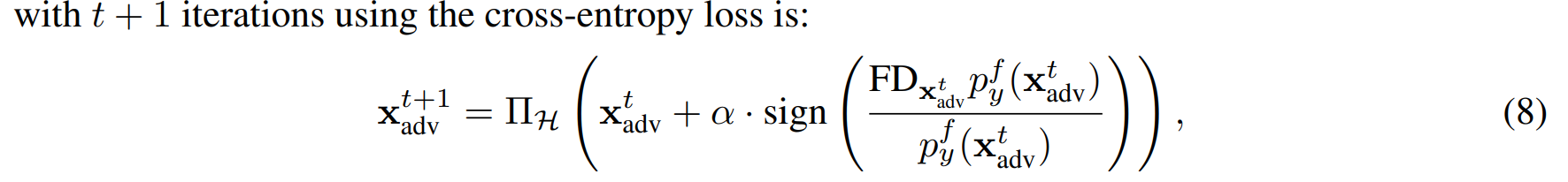
奇怪，为啥？要添加一个符号。

其次，这个损失到底是用于抵抗白盒攻击？还是干什么的？

这种攻击叫做FD-logit攻击。

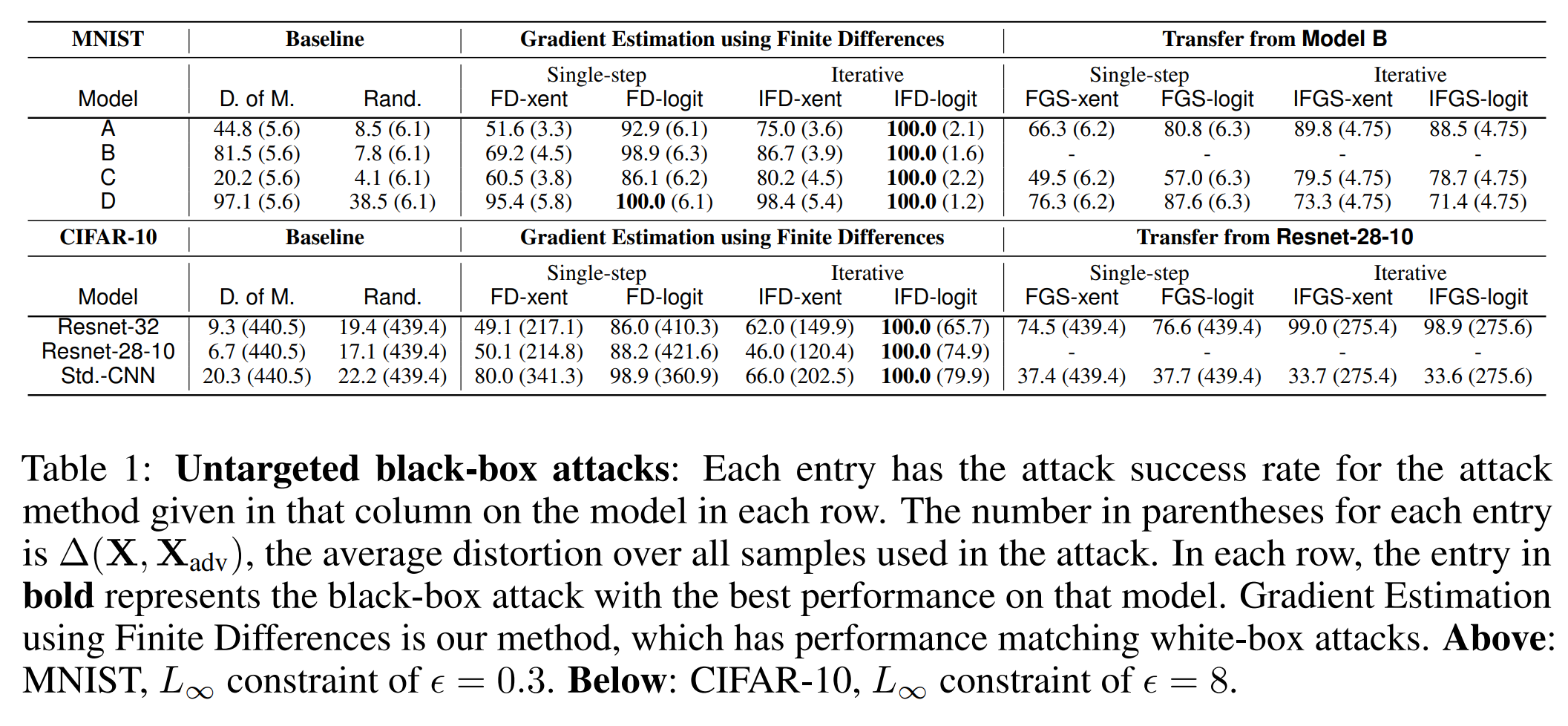
# 5.3 迭代攻击

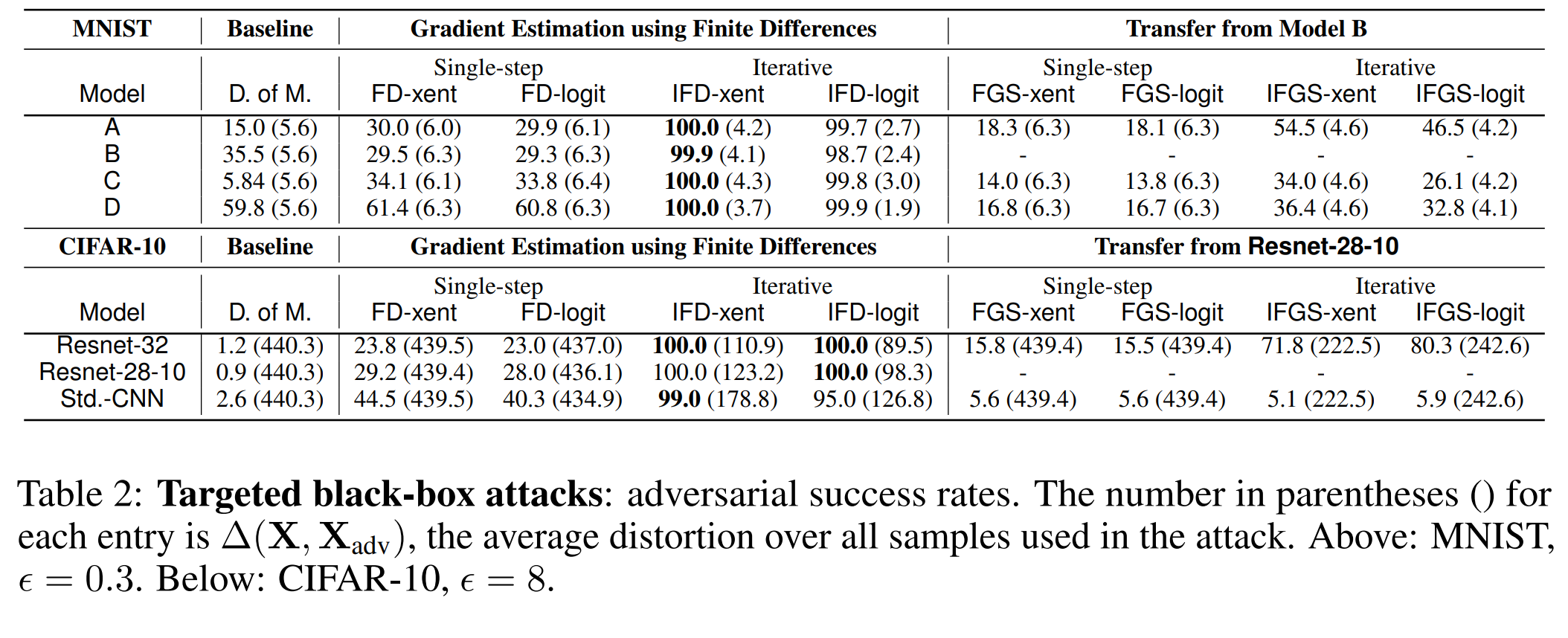
基于交叉熵的迭代攻击为：

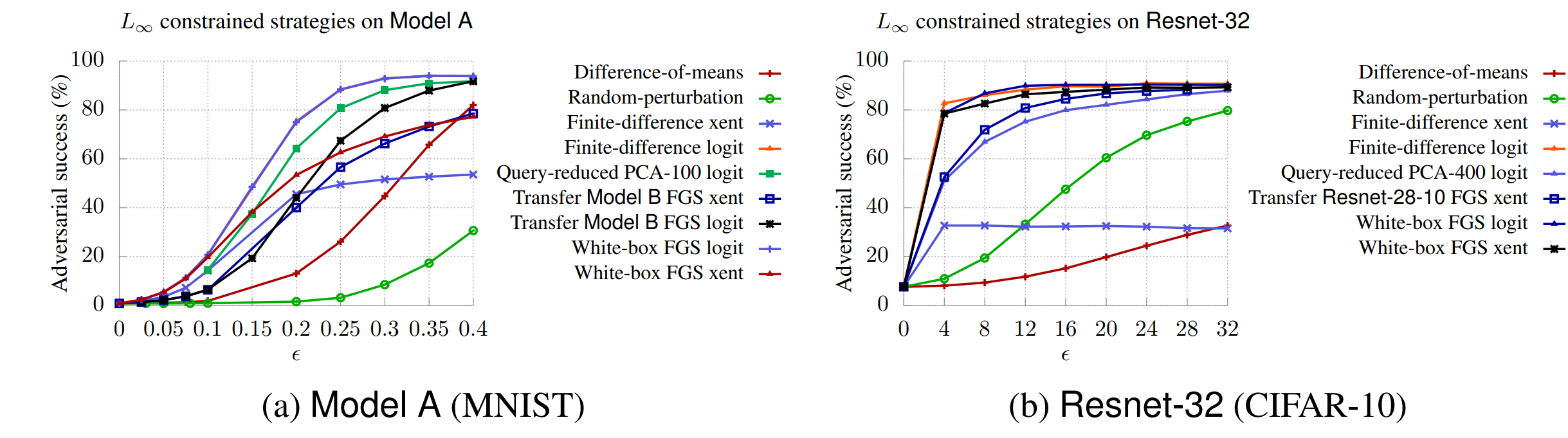


也就是对于同一个Benign样本一直往它上面**迭乘**。为啥是乘法？？？

# 5.4 有限差分的实验结果







结论：

1. 基于Logit的黑盒攻击要远远好于基于交叉熵的。因为基于logit不需要考虑求导后加上的常数。

2. 迭代攻击的效果好远远好于单次攻击。

3. 有目标攻击要比无目标攻击更加困难

4. 基于迁移性的攻击能力很有限。

# 6.减少差分攻击的query次数

基于有限差分估计梯度的方法需要询问2d次query(每一维度 各一次）。对于图片来说，这个次数太多了。两种方法都是多个维度一起改，然后统一query。

6.1 基于随机分组：

6.2 基于PCA

两种方法都是基于方向导数来估计梯度的，区别在于PCA选择信息损失最小的投影方向。随机分组是任意选。实际上，两种方法没有啥多大的区别。

# 收货

1. 作者使用是否训练一个本地模型来对黑盒攻击进行分类。

依赖知识：

白盒攻击Whit-box：白盒攻击应该是指Fast gradient sign method,这类方法需要拿到模型的梯度。但现实情况是，我们几乎很少能够拿到目标模型的梯度、参数等等。

基于迁移的黑盒攻击： 攻击样本是从一个本地的代理模型(surrogate model )产生，这种样本可以用于攻击攻击者不能直接访问的模型。

对抗训练 adversarial training：