A Frank-Wolfe Framework for Effcient and Effective Adversarial Attacks

文章链接： [<https://arxiv.org/pdf/1811.10828.pdf>（published](https://arxiv.org/pdf/2004.06660.pdf（published) by AAAI 2020）

**0 Abstract**

解决的问题：白盒攻击会产生接近扰动集上限的对抗样本，导致失真；而黑盒攻击同样会有high query complexities，导致实用性下降。

针对上述问题，本文提出了基于Frank-Wolfe算法的性能优化对抗攻击算法，以提高攻击性能。

**1 Introduction**

白盒攻击中，Optimization-based方法通常需要梯度迭代非常多次来优化对抗样本的失真度，因而效率低。而黑盒攻击中，需要大量的问询，尤其是在数据维度很高时。

本文贡献：

1. 提出基于Frank-Wolfe的projection-free的动力攻击框架。包含迭代first-order白盒攻击算法，将FGSM作为one-step特例；黑盒攻击则选择两种感应矢量（欧几里得单元或标准高斯分布）的zeroth-order优化。
2. 证明了白盒和黑盒攻击都有收敛率，黑盒攻击的问询复杂度为数据维度d的线性增长。

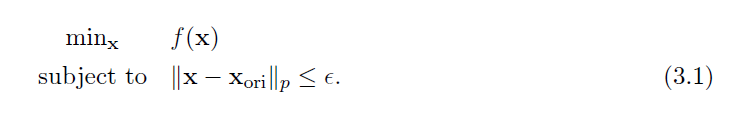
Frank-Wolfe Algorithms：是条件梯度方法，对约束性优化问题的迭代优化方法

**2 Method**

1. 问题公式化

攻击目标分为targeted attack和untargeted attack。Untargeted attack是指将模型预测为任何不正确的标签；targeted attack则是误导分类器预测为特定的目标类。本文针对目标攻击。

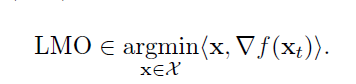
目标攻击问题：



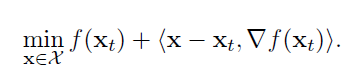
为损失函数的简化，为当时的bounded convex set。

1. 使用Frank-Wolfe的原因

PGD虽然可以实现高成功率的攻击，但多步更新公式需要在每次迭代进行额外的projection step来keep迭代依然在约束条件内。而Frank-Wolfe则是projection-free的线性最小化Orale（LMO），即



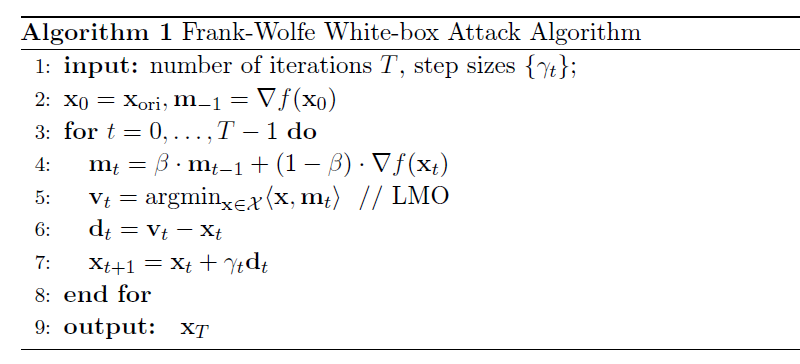
LMO可以看成是f(.)在点一阶展开的minimization：



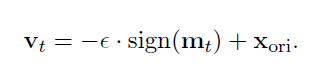
可以保证在约束集内进行迭代，避免了projection，有更好的真度。

**Frank-Wolfe White-box Attacks**

该方法基于经典点Frank-Wolfe算法，如算法1，关键性改进在第4行，引入了一个附加的动力形式，可以稳定LMO deirection，加速收敛。



解LMO是很expensive的，但根据3.1定义的约束条件，LMO可以closed-form解决。因而第5行针对可以改写为（证明在文章附录A）：



带入算法1可以得到更新：

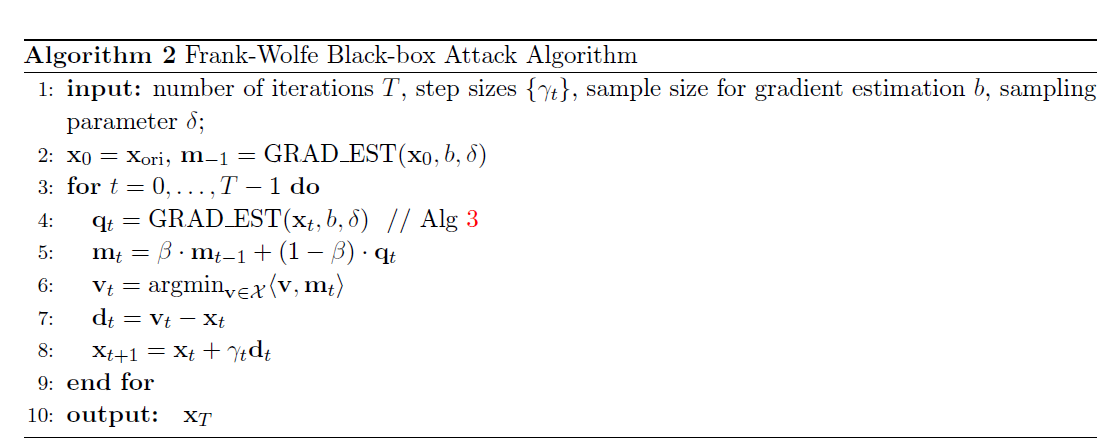
C:\Users\LISHUANG\AppData\Local\Temp\1587706986(1).png

该算法的关键在于强制*xt*逼近*xori*，使得对抗样本失真很小。

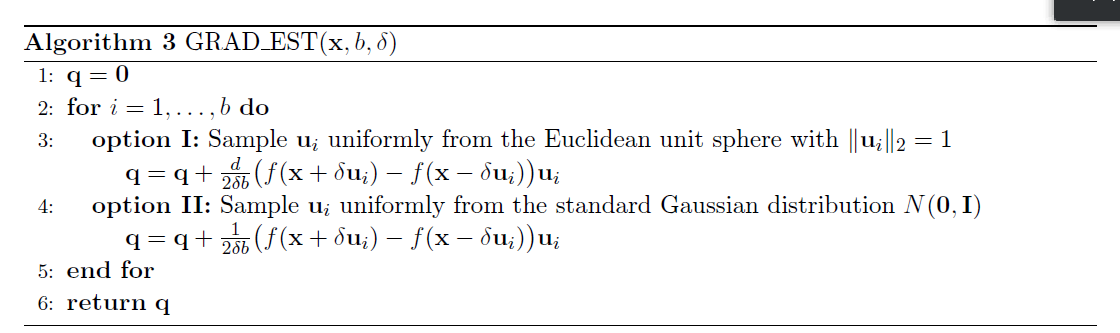
**Frank-Wolfe Black-box Attacks**

基于zeroth-order Frank-Wolfe based algorithm，如算法2：

相比于白盒攻击，关键改进为第4行，增加了一步梯度估计。而第5行则改为原来梯度估计的指数平均，这可以减少zeroth-order梯度估计的方差，快速收敛。



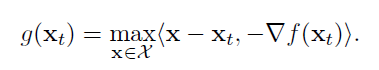
为避免使用白盒中的反向传播计算，这里使用两个选项来估测，Euclidean unit sphere和standard multivariate Gaussian distribution，提高梯度估计的有效性。



**Main Theory**

收敛判别

DNN model的损失函数通常是凹函数（nonconvex），将Frank-Wolfe gap作为收敛判别：

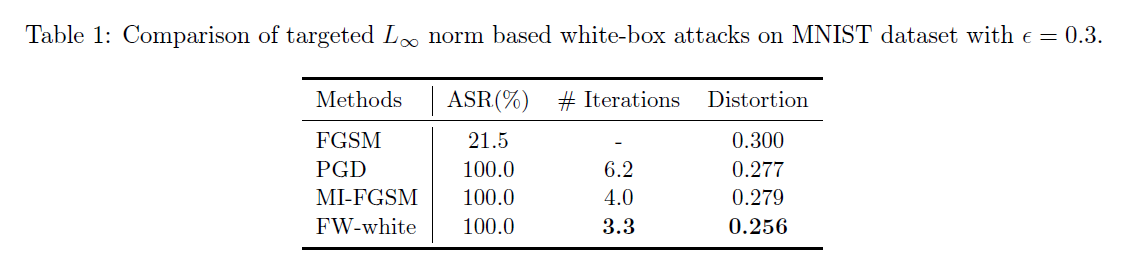


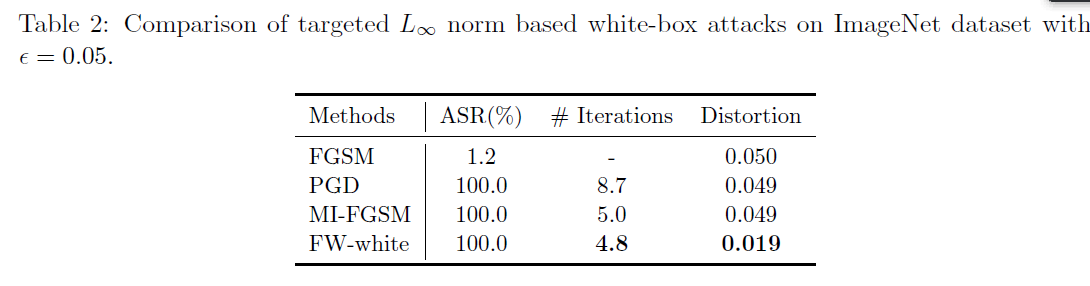
当时，当且仅当时是constrainted optimization问题的一个稳定点，才是一个完美的收敛。

**Experimental**

**训练数据集：MNIST（6层预训练过的CNN，4层卷积+2层max-pooling和Relu激活）/ImageNet（pre-trained Inception V3 model）**

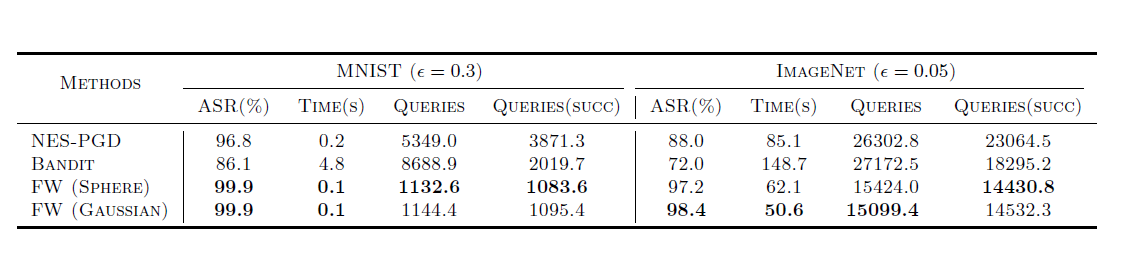
**实验1 白盒攻击实验**

****

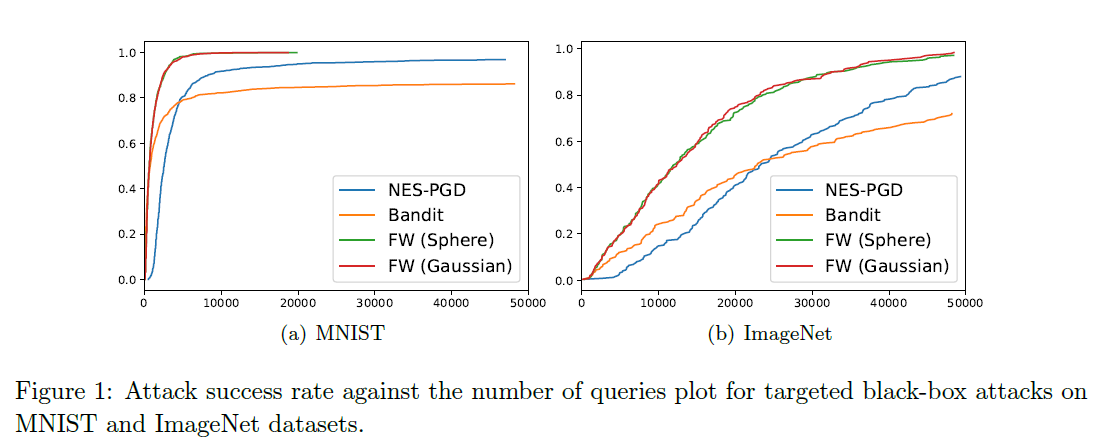
****

可以看到两个数据集的实验，FGSM每次攻击只需要1 gradient update，在MINIST和ImageNet的攻击成功率只有21.5%和1.2%，很低，其他方法都实现了100%成功率。但相比于PGD和MI-FGSM，本文方法FW-white的迭代次数更少，失真度更低。

**实验2 Black-box Attack Experiments**



可以看到，NES-PGD在MNIST数据集有很高的攻击成功率96.8%，但每次攻击需要问询5000多次，在ImageNet更是需要2万多次问询；Bandit方法improve了成功攻击样本的问询复杂度，但却降低了攻击成功率。而本文方法FW在欧几里得和高斯选项上，都显著提高了攻击成功率，大大降低了问询次数。



上图可以直观看出，红线标出的本文方法在较低的问询次数优先达到高攻击成功率，并趋于平稳。

**Future Work**

本文显著提高了在白盒攻击和黑盒攻击的攻击效率，在保证成功率的前提下，大大降低了训练时的问询次数。未来可以结合梯度/数据的思想，进一步提高该算法。

注：需要去补一下Euclidean unit sphere和standard multivariate Gaussian distribution，再看一下收敛证明。