《A little is enough: Circumventing defenses for distributed learning》

文章的主要思想是：

1. 现有防御算法通过对客户端提交的参数进行分析，从而区别恶意客户端和良性客户端。这些防御算法（methods – Krum、Trimmed Mean）的假设前提是：各个客户端所提交的更新参数是服从正态分布的。
2. 基于 1 中的假设，本文的思想是：
1）攻击者可以利用所控制的客户端来掌控更新参数值的中位数，从而使得攻击能够绕过防御算法。
2）利用正态分布的特点，通过确定最大的 z 值，来保证攻击者所提交的更新在保证隐蔽性的前提下，最大化攻击效果。

文章的优势：
可以使用同一组参数设定对多种防御算法进行规避。

攻击目标与手段：
1. 使模型无法收敛：

Algorithm 3 Preventing Convergence Attack

Input: $\{p_i : i \in CorruptedWorkers\}, n, m$
Set the number of required workers for a majority by:
$$s = \lfloor \tfrac{n}{2} + 1 \rfloor - m$$
Set (using *z-table*):
$$z^{max} = \max_z \left(\phi(z) < \frac{n-s}{n} \right)$$
for $j \in [d]$ **do**
  calculate mean (μ_j) and standard deviation (σ_j).
  $(p_{mal})_j \leftarrow \mu_j + z^{max} \cdot \sigma_j$
for $i \in$ CorruptedWorkers **do**
  $p_i \leftarrow p_{mal}$

攻击者最大化 z 值，然后对更新的每个维度的值进行“$+z^{\max}\sigma$”的偏移。
举例：
客户端总数量：50
恶意客户端数量：24
需要“seduce”的客户端数量：$\left\lfloor \frac{50}{2} + 1 \right\rfloor - 24 = 2$
最大化 z 值：查找 z-table，取最大的 z，使得$\phi(z) < \frac{50-2}{50} = 0.96$，查得$z^{\max} = 1.75$。
攻击操作：对每个恶意客户端，对每一个服从$N(\mu_j, \sigma_j)$的参数，设置其为$v = \mu + 1.75 \cdot \sigma$。
2. 后门攻击：

Algorithm 4 Backdoor Attack

1: **Input:** $\{p_i : i \in CorruptedWorkers\}, n, m$
2: Calculate z^{max}, μ_j and σ_j as in Algo 3, lines 2-5.
3: Train the model with the backdoor, with initial parameters $\{\mu_j : j \in [d]\}$ and loss function described in equations 3 and 4.
4: $\mathcal{V} \leftarrow$ final model parameters.
5: **for** $j \in [d]$ **do**
6: Clamp $v_j \in \mathcal{V}$ to the range $\mu_j \pm z_j^{max}\sigma_j$ using: $(p_{mal})_j = max(\mu_j - z_j^{max}\sigma_j, min(v_j, \mu_j + z_j^{max}\sigma_j))$
7: **for** $i \in$ CorruptedWorkers **do**
8: $p_i \leftarrow p_{mal}$

基础算法是参考 How to backdoor federal leaning 中的基于 scale 的攻击方法，利用公式（3）中的 Loss 对本地恶意模型进行训练。在保证后门能够成功注入的情况下，尽可能“小”地设置α以避免攻击者提交的参数过于异常。

$$Loss = \alpha\ell_{\text{backdoor}} + (1-\alpha)\ell_{\Delta} \qquad (3)$$

本文贡献在于，攻击者可以利用每个参数的σ_j知识，而不是直接使用任何L^p距离来表示ℓ_{Δ}.

$$\ell_{\Delta} = \sum_{j=1}^{d}\left(\frac{NewParam_j - OldParam_j}{max(z^{max}\sigma_j, 1e-5)}\right)^2 \qquad (4)$$

防御实验：

本文提出的方法对 Krum、TrimmedMean 和 Bulyan 这三种防御方法都进行了有效的攻击，相较于这三种防御方法，针对本文攻击最有效的方法是 No Defense——不采取任何防御措施。在 No Defense 下，本文的方法对模型总体只产生了微弱的偏移影响。

No Defense 是不现实的——对于 No Defense 的情况，只需在本文的基础上单独设置一个针对 No Defense 的攻击客户端即可。

文章特点：

1. 基于 scale 的攻击。
2. 利用防御算法对参数的假定——服从正态分布，进而保证攻击者大概率掌控每轮全部更新的中位数以逃避防御。
3. 不基于欧氏距离，而是基于概率分布设置了ℓ_{Δ}以保证新参数更接近原始参数。