Constructing Adversarial Examples for Vertical Federated Learning: Optimal Client Corruption through Multi-Armed Bandit

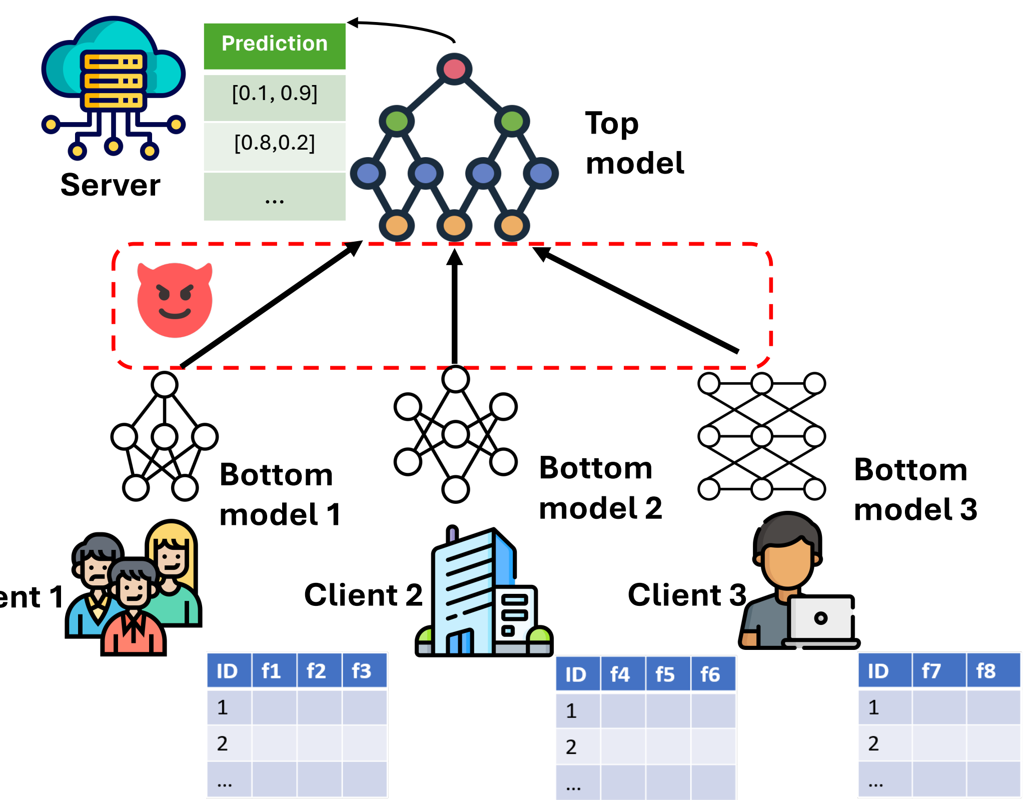
Duanyi Yao[1], Songze Li[2], Ye Xue[3], Jin Liu[4]

[1]HKUST, [2]Southeast University, [3]Shenzhen Research Institute of Big Data, CUHK(SZ), [4]HKUST(GZ)

Vertical federated learning (VFL)

VFL inference process:

1. Client $m \in [M]$ computes embedding h_m and sends to the server;
2. The server receives and aggregates all embeddings as $[h_1, h_2, \dots, h_M]$;
3. The server forward propagates the aggregated embedding and derives the prediction vectors, which are sent to all clients.

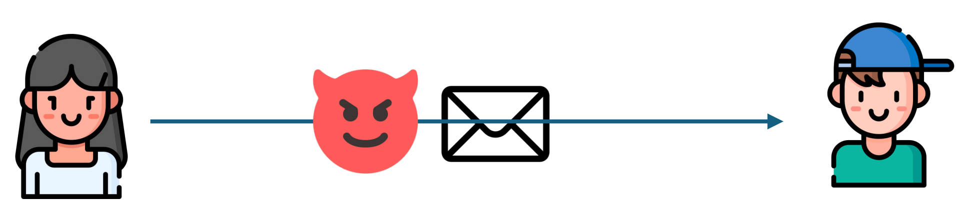


VFL inference process and adversary's capability

Adversary

Goal: (Target label y_v, Prediction $\hat{y}$)

1. Targeted attack: $\hat{y} = y_v$.
2. Untargeted attack: $\hat{y} \neq y_v$.



Metric: Attack success rate (ASR)

Capability:

1. The adversary can access, replay, and manipulate messages on the communication channel between two endpoints (i.e., the channel between client x and the server).
2. The adversary can corrupt at most $C \le M$ clients and perturb their embeddings $h_{i,a}$ to $\tilde{h}_{i,a}$ such that $\|\tilde{\boldsymbol{h}}_{i,a} - \boldsymbol{h}_{i,a}\|_\infty \le \beta(ub_i - lb_i)$.

Adaptive corruption: The adversary adaptively adjusting their corruption patterns C^t.

Problem definition

The adversary aims to find the **optimal set of corruption patterns** $\{\mathcal{C}^t\}_{t=1}^T$, and the **optimal set of perturbations** $\{\{\boldsymbol{\eta}_i^t\}_{i=1}^{B^t}\}_{t=1}^T$ for each sample $i \in [B^t]$ in attack round $t \in [T]$, maximizing the expected cumulative ASR over T attack rounds.

Formulate this attack as an online optimization problem

$$\max_{\{\mathcal{C}^t\}_{t=1}^T} \frac{\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_t\left[\max_{\{\boldsymbol{\eta}_i^t\}_{i=1}^{B^t}} A(\{\boldsymbol{\eta}_i^t\}_{i=1}^{B^t}, \mathcal{C}^t; B^t)\right]\right]}{\sum_{t=1}^T B^t}$$
$$\text{s.t. } |\mathcal{C}^t| = C,\ \|\boldsymbol{\eta}_i^t\|_\infty \le \beta(ub_i - lb_i),\ \forall t \in [T].$$

$\mathbb{E}_t$ is taken over the randomness with the t-th attack round

$\mathbb{E}$ is taking over the randomness of all T rounds

Methodology

Decompose into an inner problem **of adversarial example generation (AEG)** and an outer problem of **corruption pattern selection (CPS)**

$$\text{Inner problem (AEG):} \quad \min_{\boldsymbol{\eta}_i^t} L(\boldsymbol{\eta}_i^t; \mathcal{C}^t), \quad \text{s.t. } \|\boldsymbol{\eta}_i^t\|_\infty \le \beta(ub_i - lb_i), \forall i \in [B^t]. \tag{1}$$

$$\text{Outer problem (CPS):} \quad \min_{\{\mathcal{C}^t\}_{t=1}^T} \frac{\mathbb{E}\left[\sum_{t=1}^T (\alpha^* - \mathbb{E}_t\left[A^*(\mathcal{C}^t; B^t)\right])\right]}{\sum_{t=1}^T B^t} \quad \text{s.t. } |\mathcal{C}^t| = C,\ \forall t \in [T], \tag{2}$$

AEG solution: Use natural evolution strategy (NES) combined with projected gradient decent method to solve (1). NES is a type of zero-order gradient method, which employ gaussian noise to query model for estimating gradients. The estimation is given by:

$$\nabla_{\boldsymbol{\eta}_i^t} L(\boldsymbol{\eta}_i^t; \mathcal{C}^t) \approx \frac{1}{\sigma n} \sum_{j=1}^n \boldsymbol{\delta}_j L\left(\boldsymbol{\eta}_i^t + \sigma \boldsymbol{\delta}_j; \mathcal{C}^t\right). \tag{3}$$

CPS solution:
We transform CPS to an **MAB problem.**

Picking a corruption pattern C^t	→	A selected arm $k(t)$ in a round t
The expected reward $\mathbb{E}_t[A^*(C^t, B^t)]$	→	Mean $\mu_{k(t)}$
Best corruption pattern's mean	→	μ_1
The attack ASR $A(C^t, B^t)$	→	Reward $r_{k(t)}(t)$
CPS problem in (2)	→	$\min_{\{k(t)\}_{t=1}^T} \mathbb{E}\left[\sum_{t=1}^T (\mu_1 - \mu_{k(t)})\right]$

For solving this MAB problem, we propose a novel method named **Thompson sampling with empirical maximum reward (E-TS)** (Algorithm 1), enabling the adversary to efficiently identify the optimal corruption pattern.

Algorithm 1 E-TS for CPS

Initialization: $\forall k \in [N], \hat{\mu}_k = 0, \hat{\sigma}_k = 1, n_k = 0, r_k^{\max} = 0, \hat{\varphi}_k = 0.$
for $t = 1, 2, \dots, T$ **do**
  if $t > t_0$ **then**
    Select fully explored arms to construct the set $\mathcal{S}_t = \{k \in [N] : n_k \ge \frac{(t-1)}{N}\}$.
    Select the empirical best arm $k^{emp}(t) = \max_{k \in \mathcal{S}_t} \hat{\mu}_k$.
    Initialize $\mathcal{E}^t = \emptyset$, add arms $k \in [N]$ which satisfy $\hat{\varphi}_k \ge \hat{\mu}_{k^{emp}(t)}$ to $\mathcal{E}^t$.
  else
    Initialize set $\mathcal{E}^t = [N]$.
  end if
  $\forall k \in \mathcal{E}^t$: Sample $\theta_k \sim \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k)$.
  Choose the arm $k(t) = \arg\max_k \theta_k$ and decide the corrpution pattern $\mathcal{C}^t = k(t)$.
  Sample batch data $[B^t]$, play the arm $k(t)$ as the corruption pattern in Algorithm 2 and observe the reward $r_{k(t)}(t)$ from the attack result for the corrupted embedding $\boldsymbol{h}_{i,a}^t = [\boldsymbol{h}_{i,a_1}^t, \dots, \boldsymbol{h}_{i,a_C}^t], \forall i \in [B^t]$.
  Update $n_{k(t)} = n_{k(t)} + 1$, $\hat{\mu}_{k(t)} = \frac{\hat{\mu}_{k(t)}(n_{k(t)}-1) + r_{k(t)}(t)}{n_{k(t)}}$, $\hat{\sigma}_{k(t)} = \frac{1}{n_{k(t)}+1}$, $r_{k(t)}^{\max} = \max\{r_{k(t)}^{\max}, r_{k(t)}(t)\}$, $\hat{\varphi}_{k(t)} = \frac{\hat{\varphi}_{k(t)}(n_{k(t)}-1) + r_{k(t)}^{\max}}{n_{k(t)}}$.
end for
Output $\{k(1), \dots, k(T)\}$

t_0: warm-up round. $\hat{\varphi}_{k(t)}$: empirical maximum reward of $k(t)$. $\mathcal{E}_t$: competitive set at t round.

The key idea of E-TS is to **limit the exploration within the competitive set**, which is defined using the expected maximum reward of each arm

Regret Analysis

Lemma 1 (Expected pulling times of a non-competitive arm). *Under the above assumption, for a non-competitive arm $k^{nc} \neq 1$ with $\tilde{\Delta}_{k^{nc},1} < 0$, the expected number of pulling times in T rounds, i.e., $\mathbb{E}[n_{k^{nc}}(T)]$, is bounded by $\mathbb{E}[n_{k^{nc}}(T)] \le \mathcal{O}(1)$.*

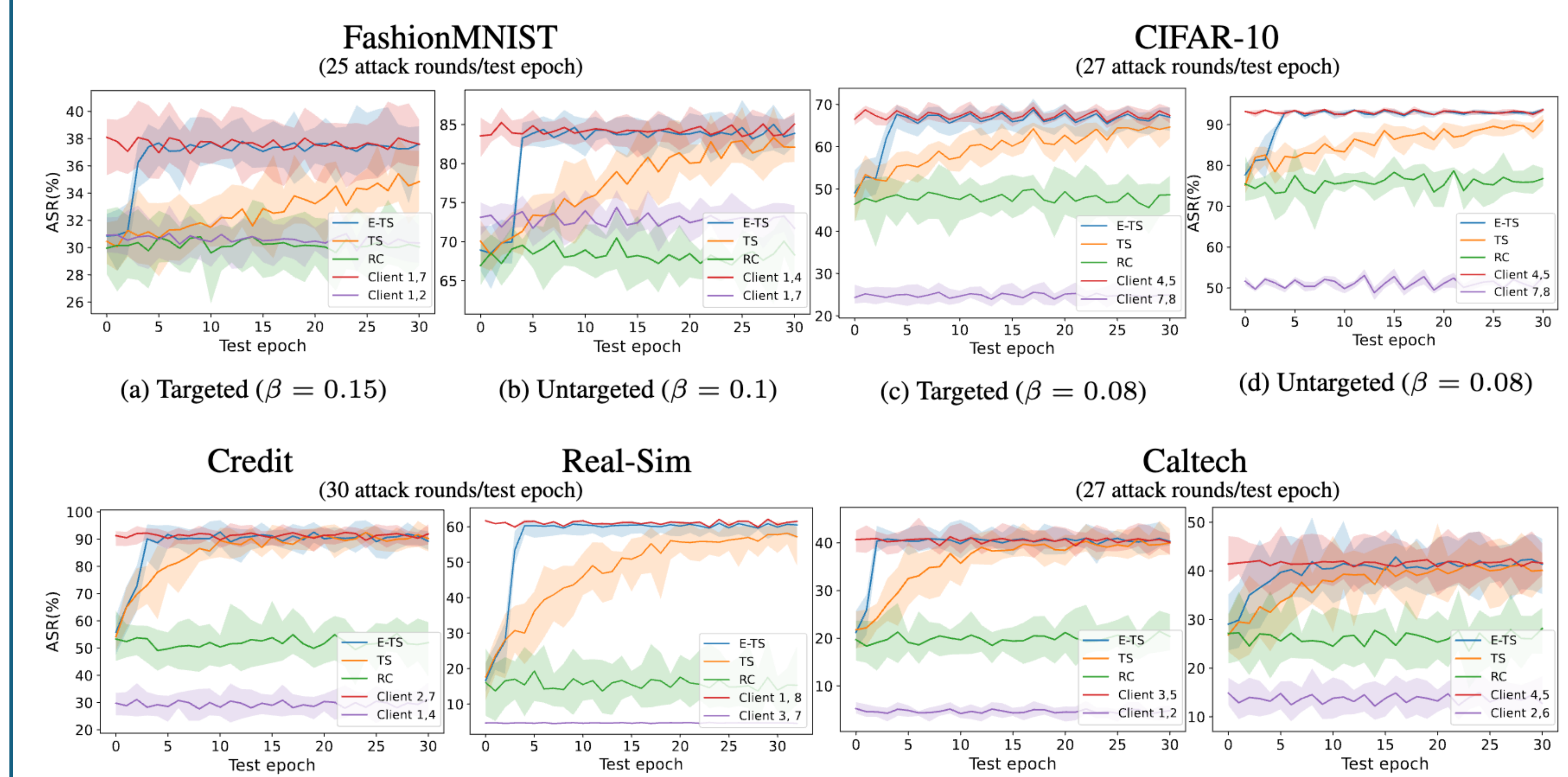
Lemma 2 (Expected pulling times of a competitive but sub-optimal arm). *Under the above assumption, the expected number of times pulling a competitive but sub-optimal arm k^{sub} with $\tilde{\Delta}_{k^{sub},1} \ge 0$ in T rounds is bounded as follows,*

$$\mathbb{E}[n_{k^{sub}}(T)] = \sum_{t=1}^T \Pr(k(t) = k^{sub}, n_1(t) \ge \frac{t}{N}) \le \mathcal{O}(\log(T)).$$

Theorem 1 (Upper bound on expected regret of E-TS). *Let $D \le N$ denote the number of competitive arms. Under the above assumption, the expected regret of the E-TS algorithm is upper bounded by $D\mathcal{O}(\log(T)) + (N - D)\mathcal{O}(1)$.*

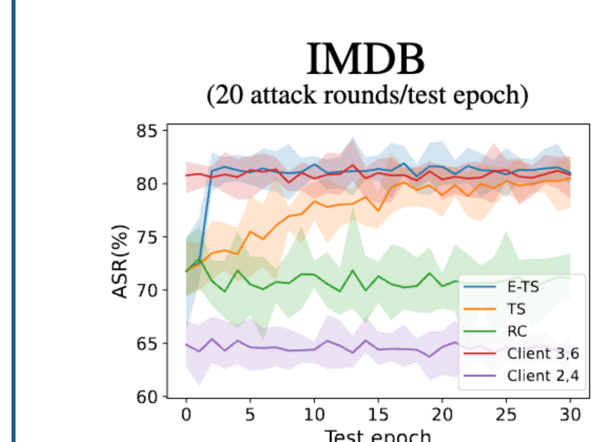
Note that the regret of traditional TS is bounded by $N\mathcal{O}(\log(T))$.

Experimental result



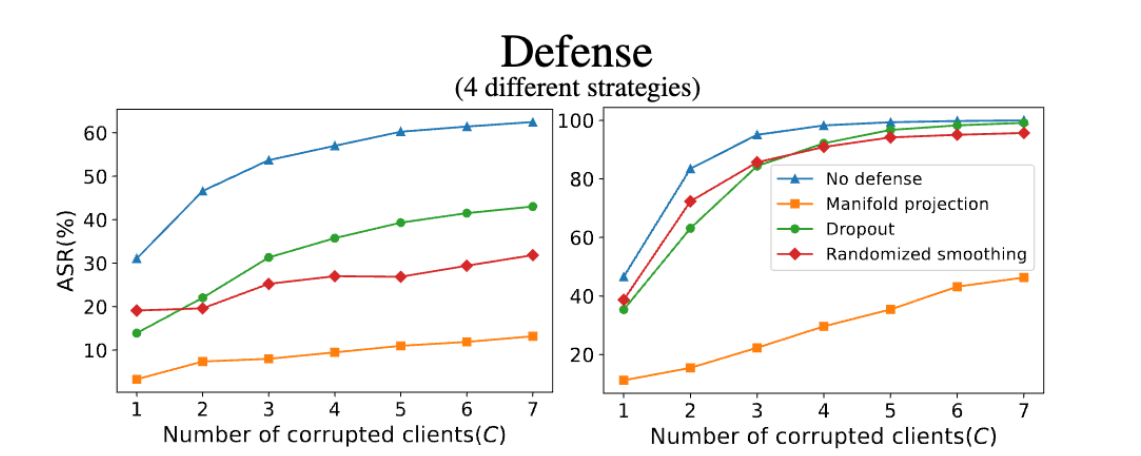
(a) Targeted ($\beta = 0.15$) (b) Untargeted ($\beta = 0.1$) (c) Targeted ($\beta = 0.08$) (d) Untargeted ($\beta = 0.08$)

(e) $\beta = 0.3$ (f) $\beta = 0.3$ (g) Targeted ($\beta = 0.3$) (h) Untargeted ($\beta = 0.3$)



(i) $\beta = 0.05$

Figure 1: Attack performance on six datasets of distinct VFL tasks.



(a) Targeted ($\beta = 0.3$) (b) Untargeted ($\beta = 0.1$)

Figure 2: Attack performance on FashionMNIST under different defense strategies.

The Twelfth International Conference on Learning Representations （ICLR 2024）