A USEFUL FACTS

Fact 1. *For* $\forall \alpha > 0$, $\forall a, b \in R^d$, $\|a+b\|^2 \leq (1+\alpha)\|a\|^2 + (1-\alpha^{-1})\|b\|^2$. *And,* $\forall A, B \in R^{n\times m}$, $\|A+B\|^2 \leq (1+\alpha)\|A\|^2 + (1-\alpha^{-1})\|B\|^2$.

Fact 2. *For* $A \in R^{d\times n}$, $B \in R^{n\times n}$,

$$\|AB\|_F \leq \|A\|_F \|B\|_2 .$$

Fact 3. *Let* $B^{(t)} = [b_1^(t), \cdots, b_n^(t)] \in R^{n\times m}$, $\bar{B}^{(t)} = [\bar{b}^(t), \cdots, \bar{b}^(t)] \in R^{n\times m}$, $\bar{b}^(t) = \frac{1}{n}\sum_{i=1}^{n} b_i^{(t)}$, *A is doubly stochastic. Then*

$$\bar{B}^{(t)} = B^{(t)}\frac{1}{m}1_m 1_m^T,$$
$$\bar{B}^{(t)}A = \bar{B}^{(t)}.$$

Fact 4. *If A is the gossip matrix with second largest eigenvalue* $1-\delta = |\lambda_2| < 1$, *then*

$$\left\|A^k - \frac{1}{m}1_m 1_m^T\right\|_2 \leq (1-\delta)^k.$$

Fact 5. *Let* U_i *be a collection of subsets of* R^m. *Then for every* $u \in conv(\sum_{i=1}^{n} u_i)$, *there is a subset* $\Upsilon(u) \subseteq [n]$ *of size at most m such that*

$$u \in [\sum_{i\notin\Upsilon(u)} U_i + \sum_{i\in\Upsilon(u)} conv(U_i)]. \tag{23}$$

B PROOF OF THEOREM 1

Before starting the proof, describe the variables in Adaptive-Compressed-Gossip in matrix form. Define:

$$\begin{aligned}
&W_d^{(t)} = [w_{d_1}^{(t)}, \cdots, w_{d_m}^{(t)}] \in R^{d\times m},\\
&\hat{W}_d^{(t)} = [\hat{w}_{d_1}^{(t)}, \cdots, \hat{w}_{d_m}^{(t)}] \in R^{d\times m},\\
&G^{(t)} = [g_1, \cdots, g_m] \in R^m,\\
&R^{(t)} = [r_1, \cdots, t_m] \in R^m.
\end{aligned} \tag{24}$$

Then, in each iteration, the variables in Adaptive-Compressed-Gossip are updated as follows:

$$\begin{aligned}
&G^{(t)} = C(W_d^{(t)} - \hat{W}_d^{(t)}, R^{(t)}),\\
&\hat{W}_d^{(t+1)} = \hat{W}_d^{(t)} + G^{(t)},\\
&W_d^{(t+1)} = W_d^{(t)} + \gamma\hat{W}_d^{(t+1)}(A-I),
\end{aligned} \tag{25}$$

where A is the gossip matrix in Assumption 1, and I is the identity matrix.

Lemma 1. *Let* $\bar{W}_d = [\bar{w}_d, \cdots, \bar{w}_d]$, *where* $\bar{w}_d = \frac{1}{m}W_d^{(t)}1_m$, $\bar{w}_d \in R^d$. *Then, for* $\forall \alpha_1 > 0$,

$$\left\|W_d^{(t+1)} - \bar{W}_d\right\|_F^2 \leq (1-\delta\gamma)^2\left\|W_d^{(t)} - \bar{W}_d\right\|_F^2 + \gamma^2(1+\alpha_1^{-1})\beta^2\left\|\hat{W}_d^{(t+1)} - W_d^{(t)}\right\|_F^2, \tag{26}$$

where δ *and* β *is same as defined in Equation* (12).

Proof.

$$\begin{aligned}
\left\|W_d^{(t+1)} - \bar{W}_d\right\|_F^2 &= \left\|W_d^{(t)} - \bar{W}_d + \gamma\hat{W}_d^{(t+1)}(A-I)\right\|_F^2\\
&= \left\|W_d^{(t)} - \bar{W}_d + \gamma(W_d^{(t)} - \bar{W}_d)(A-I) + \gamma(\hat{W}_d^{(t+1)} - W_d^{(t)})(A-I)\right\|_F^2\\
&= \left\|(W_d^{(t)} - \bar{W}_d)((1-\gamma)I + \gamma A) + \gamma(\hat{W}_d^{(t+1)} - W_d^{(t)})(A-I)\right\|_F^2 .
\end{aligned} \tag{27}$$

Then, according to Fact 1 and 2, there is

$$\begin{aligned}\left\|W_d^{(t+1)} - \bar{W}_d\right\|_F^2 \leq (1+\alpha_1)\left\|(W_d^{(t)} - \bar{W}_d)((1-\gamma)I + \gamma A)\right\|_F^2 \\ + (1+\alpha_1^{-1})\gamma^2 \left\|A - I\right\|_F^2 \left\|\hat{W}^{(t+1)} - W^{(t)}\right\|_F^2 .\end{aligned} \tag{28}$$

For the first term,

$$\begin{aligned}\left\|(W_d^{(t)} - \bar{W}_d)((1-\gamma)I + \gamma A)\right\|_F &\leq (1-\gamma)\left\|W_d^{(t)} - \bar{W}_d\right\|_F + \gamma\left\|(W_d^{(t)} - \bar{W}_d)A\right\|_F \\ &= (1-\gamma)\left\|W_d^{(t)} - \bar{W}_d\right\|_F + \gamma\left\|(W_d^{(t)} - \bar{W}_d)(A - \frac{1}{m}1_m 1_m^T)\right\|_F \\ &\leq (1-\gamma\delta)\left\|W_d^{(t)} - \bar{W}_d\right\|_F ,\end{aligned} \tag{29}$$

where the second line uses the conclusion of Fact 3, the third line uses conclusion of Fact 2 and 4. Then this lemma can be obtained by combining Equation (28) and (29). □

Lemma 2. *Let $\bar{W}_d$ be defined as in Lemma 1. Then, for $\forall \alpha_2$*

$$E\left\|W_d^{(t+1)} - \hat{W}_d^{(t+2)}\right\|_F^2 \leq \varepsilon p(1+\gamma\beta)^2(1+\alpha_2)\left\|W_d^{(t)} - \hat{W}_d^{(t+1)}\right\|_F^2 + \varepsilon p\gamma^2\beta^2(1+\alpha_2^{-1})\left\|W_d^{(t)} - \bar{W}_d\right\|_F^2 \tag{30}$$

Proof. According to the definition of $W_d^{(t+1)}$ and $\hat{W}_d^{(t+2)}$, there is

$$\begin{aligned}E\left\|W_d^{(t+1)} - \hat{W}_d^{(t+2)}\right\|_F^2 &= E\left\|W_d^{(t+1)} - \hat{W}_d^{(t+1)} - C(W_d^{(t+1)} - \hat{W}_d^{(t+1)}, r)\right\|_F^2 \\ &\leq \varepsilon p\left\|W_d^{(t+1)} - \hat{W}_d^{(t+1)}\right\|_F^2 \\ &= \varepsilon p\left\|W_d^{(t)} + \gamma\hat{W}_d^{(t+1)}(A-I) - \hat{W}_d^{(t+1)}\right\|_F^2 \\ &= \varepsilon p\left\|(W_d^{(t)} - \hat{W}_d^{(t+1)})((1+\gamma)I - \gamma A) + \gamma(A-I)(W_d^{(t)} - \bar{W}_d)\right\|_F^2 \\ &\leq \varepsilon p(1+\alpha_2)\left\|(W_d^{(t)} - \hat{W}_d^{(t+1)})((1+\gamma)I - \gamma A)\right\|_F^2 \\ &+ \varepsilon p(1+\alpha_2^{-1})\left\|\gamma(A-I)(W_d^{(t)} - \bar{W}_d)\right\|_F^2 \\ &\leq \varepsilon p(1+\gamma\beta)^2(1+\alpha_2)\left\|W_d^{(t)} - \hat{W}_d^{(t+1)}\right\|_F^2 \\ &+ \varepsilon p\gamma^2\beta^2(1+\alpha_2^{-1})\left\|W_d^{(t)} - \bar{W}_d\right\|_F^2 ,\end{aligned} \tag{31}$$

since eigenvalues of $\gamma(I-A)$ are positive, here we used $\|I + \gamma(I-A)\|^2 = 1+\gamma\|I-A\|_F = 1+\gamma\beta$. □

With the conclusion of Lemma 1 and Lemma 2, we are now ready for the proof of Theorem 1. As shown in Fact 3, $\bar{W}_d = W_d^{(t)}\frac{1}{m}1_m 1_m^T$ for all $t \geq 0$, we have:

$$\begin{aligned}Ee^{(t+1)} &\leq \varphi_1(\gamma)\left\|W_d^{(t)} - \bar{W}_d\right\|_F^2 + \varphi_2(\gamma)\left\|\hat{W}_d^{(t+1)} - W_d^{(t)}\right\|_F^2 \\ &\leq e^{(t)}\max\{\varphi_1(\gamma), \varphi_2(\gamma)\},\end{aligned} \tag{32}$$

where

$$\begin{aligned}\varphi_1(\gamma) &= (1-\delta\gamma)^2(1+\alpha_1) + \varepsilon p\gamma^2\beta^2(1+\alpha_2^{-1}), \\ \varphi_2(\gamma) &= \gamma^2\beta^2(1+\alpha_1^{-1}) + \varepsilon p(1+\delta\gamma)^2(1+\alpha_2).\end{aligned} \tag{33}$$

Then, let

$$\begin{aligned} \alpha_1 &= \tfrac{\gamma\delta}{2}, \\ \alpha_2 &= \tfrac{1-\varepsilon p}{2}, \\ \gamma^* &= \tfrac{\delta(1-\varepsilon p)}{16\delta+\delta^2+4\beta^2+2\delta\beta^2-8\delta(1-\varepsilon p)}, \end{aligned} \tag{34}$$

it holds

$$\max\{\varphi_1(\gamma^*), \varphi_2(\gamma^*)\} \leq 1 - \frac{\delta^2(1-\varepsilon p)}{2(16\delta + \delta^2 + 4\beta^2 + 2\delta\beta^2 - 8\delta(1-\varepsilon p))}. \tag{35}$$

The claim of Theorem 1 then follows by observing

$$1 - \frac{\delta^2(1-\varepsilon p)}{2(16\delta + \delta^2 + 4\beta^2 + 2\delta\beta^2 - 8\delta(1-\varepsilon p))} \leq 1 - \frac{\delta^2(1-\varepsilon p)}{82}, \tag{36}$$

using the crude estimates $0 \leq \delta \leq 1$, $\beta \leq 2$, $\varepsilon \leq 1$, $p \leq 1$.

Then, we use the ε_{max} and p_{max} to represent the maximizes values of ε and p, the conclusion of Theorem 1 can be held.

C PROOF OF THEOREM 2

In this section, we use w_d to represent the consensus of D-Nodes. Before proving Theorem 2, we introduce the following lemmas.

Lemma 3. *Let $\phi(y) = \inf_{w_d} \sup_{w_{g_i}} \{\tilde{F}(w_g, w_d) - y^T w_d\}$, then*

$$\sup_{w_d} \inf_{w_{g_i}} \tilde{F}(w_g, w_d) = \hat{cl}(\phi(0)) \leq \phi(0) = \inf_{w_{g_i}} \sup_{w_d} \tilde{F}(w_g, w_d). \tag{37}$$

Proof. $\hat{cl}(\phi(0)) \leq \phi(0)$ because of the weak duality theorem. For $\sup_{w_d} \inf_{w_{g_i}} \tilde{F}(w_g, w_d) = \hat{cl}(\phi(0))$, we have $\phi(y) = \inf_{w_{g_j}} \hat{\phi}_{w_g}(y)$, where $\hat{\phi}_{w_g}(y) = \sup_{w_d}\{(-\tilde{F}(w_g, w_d))^*(-y)\}$, and then, by then definition of conjugate function, there is

$$\begin{aligned} \inf_y\{\hat{\phi}_{w_g}(y) + y^T\mu\} &= -\sup_y\{y^T(-\mu) - \phi_{w_g}(y)\} \\ &= -(\phi_{w_g})^*(-y) \\ &= -(-\tilde{F}(w_g, w_d))^{**}(\mu) \\ &= \tilde{F}(w_g, \mu). \end{aligned} \tag{38}$$

Therefore,

$$\begin{aligned} \hat{cl}(\phi(0)) &= \sup_\mu \inf_y \{\phi(y) + y^T\mu\} \\ &= \sup_\mu \inf_y \inf_{w_{g_i}} \{\hat{\phi}_{w_g}(y) + y^T\mu\} \\ &= \sup_\mu \inf_{w_{g_i}} \inf_y \{\hat{\phi}_{w_g}(y) + y^T\mu\} \\ &= \sup_\mu \inf_{w_{g_i}} \tilde{F}(w_g, \mu) \end{aligned} \tag{39}$$

□

Lemma 4. *Under the Assumptions 3, 4, and 5, we have $\phi(0) - \hat{cl}(\phi(0)) \leq (d+1)\mu_{g_i}^{(n)}$*

Proof.

$$
\begin{aligned}
\phi(y) &= \inf_{w_d} \sup_{w_{g_i}} \{\tilde{F}(w_g, w_d) - y^T w_d\} \\
&= \inf_{w_d} \sup_{w_{g_i}} \{-\sum_{i=1}^{n} \hat{cl}(-f(w_{g_i}), w_d) - y^T w_d\} \\
&= \inf_{w_d} \left(\sum_{i=1}^{n} \hat{cl}(-f(w_{g_i}), \cdot) \right)^* (-y) \\
&= \inf_{w_d} \inf_{y_1+\cdots+y_n=-y} \{\sum_{i=1}^{n} (\hat{cl}(-f(w_{g_i}, \cdot)))^*(y_i)\} \\
&= \inf_{w_d} \inf_{y_1+\cdots+y_n=-y} \{\sum_{i=1}^{n} (-f(w_{g_i}, \cdot))^*(y_i)\} \\
&= \inf_{y_1+\cdots+y_n=-y} \inf_{w_d} \{(-f(w_{g_1}, \cdot))^*(y_1) + \cdots + (-f(w_{g_n}, \cdot))^*(y_n)\} \\
&= \inf_{y_1+\cdots+y_n=-y} \{h_1(y_1) + \cdots + h_n(y_n)\},
\end{aligned}
\tag{40}
$$

where $y_1, \cdots, y_n, y \in R^d$. Then

$$
\phi(0) = \inf_{y_1+\cdots+y_n=-y} \sum_{i=1}^{n} h_i(y_i), \quad s.t. \quad \sum_{i=1}^{n} y_i = 0. \tag{41}
$$

Consider that the subset of R^{t+1}:

$$
U_i = \{u_i \in R^{t+1} : u_i = [y_i, h_i(y_i)]\}, \ i = 1, \cdots, n. \tag{42}
$$

Let $U = \sum_i^n U_i$, then U, $conv(U)$, U_i, and $conv(U_i)$ are compact sets. According to the standard duality argument, there is

$$
\phi(0) = \inf\{b : \ there\ exists\ (a, b) \in U\ such\ that\ a = 0\}, \tag{43}
$$

and

$$
\hat{cl}(\phi(0)) = \inf\{b : \ there\ exists\ (a, b) \in conv(U)\ such\ that\ a = 0\}, \tag{44}
$$

Let $(\bar{a}, \bar{b}) \in conv(U)$ be such that $\bar{a} = 0$ and $\bar{b} = \hat{cl}(\phi(0))$. According to Fact 5, we have $(\bar{a}_i, \bar{b}_i) \in conv(U_i)$, $i \in \Upsilon$ and $\bar{y}_i \in dom(h_i)$, $i \notin \Upsilon$, where Υ is a subset $\Upsilon \subseteq [n]$. Then

$$
\begin{aligned}
&\sum_{i \notin \Upsilon} \bar{y}_i + \sum_{i \in \Upsilon} \bar{a}_i = \bar{a} = 0, \\
&\sum_{i \notin \Upsilon} h_i(\bar{y}_i) + \sum_{i \in \Upsilon} \bar{b}_i = \hat{cl}(\phi(0)).
\end{aligned}
\tag{45}
$$

For each $i \in \Upsilon$, there are vectors $\{y_i^j\}_{j=1}^{d+2}$ and scalars $\{c_i^j\}_{j=1}^{d+2} \in R$ such that

$$
\begin{aligned}
&\sum_{j=1}^{d+2} c_i^j = 1, \ c_i^j \geq 0, \ j \in [d+2], \\
&\bar{a}_i = \sum_{j=1}^{d+2} c_i^j y_i^j = \bar{y}_i \in dom(h_i), \ \bar{b}_i = \sum_{j=1}^{d+2} c_i^j h_i(y_i^j).
\end{aligned}
\tag{46}
$$

For $i \in \Upsilon$,

$$\begin{aligned}\bar{b}_i &\geq \hat{cl}(h_i(\sum_{j=1}^{d+2} c_i^j y_i^j)) \\ &\geq h_i(\sum_{j=1}^{d+2} c_i^j y_i^j) - \mu_{g_i} \\ &= h_i(\bar{y}_i) - \mu_{g_i}.\end{aligned} \tag{47}$$

Then, we have

$$\sum_{i=1}^{n} \bar{y}_i = 0 \tag{48}$$

Therefore, there is

$$\begin{aligned}\phi(0) &= \sum_{i=1}^{n} h_i(\bar{y}_i) \\ &\leq \hat{cl}\phi(0) + \sum_{i \in \Upsilon} \mu_{g_i} \\ &\leq \hat{cl}\phi(0) + |\Upsilon|\, \mu_{g_i}^{(n)} \\ &= \hat{cl}\phi(0) + (d+1)\mu_{g_i}^{(n)}.\end{aligned} \tag{49}$$

□

Lemma 5. *Let*

$$v^* = \inf_{w_{g_i}} \sup_{w_d} F(w_g, w_d),$$

and

$$\hat{v}^* = \sup_{w_d} \inf_{w_{g_i}} F(w_i, w_d).$$

When the number of generators $n > \frac{d+1}{\iota}\mu_g^{(n)}$, there is

$$0 \leq v^* - \hat{v}^* \leq \mu_d + \hat{\mu}_d + \iota. \tag{50}$$

In addition, if $f(w_{g_i}, w_{d_j})$ is concave and closed, $v^* - \hat{v}^* \leq \iota$.

Proof. Combining Lemma 3 and Lemma 4 gives:

$$\inf_{w_{g_i}} \sup_{w_d} \tilde{F}(w_g, w_d) - \sup_{w_d} \inf_{w_{g_i}} \tilde{F}(w_g, w_d) \leq \iota, \tag{51}$$

note that

$$\begin{aligned}v^* - \hat{v}^* &= \inf_{w_{g_i}} \sup_{w_d} F(w_g, w_d) - \sup_{w_d} \inf_{w_{g_i}} F(w_g, w_d) \\ &= \inf_{w_{g_i}} \sup_{w_d} F(w_g, w_d) - \inf_{w_{g_i}} \sup_{w_d} \tilde{F}(w_g, w_d) \\ &+ \inf_{w_{g_i}} \sup_{w_d} \tilde{F}(w_g, w_d) - \sup_{w_d} \inf_{w_{g_i}} \tilde{F}(w_g, w_d) \\ &+ \sup_{w_d} \inf_{w_{g_i}} \tilde{F}(w_g, w_d) - \sup_{w_d} \inf_{w_{g_i}} F(w_g, w_d) \\ &= \mu_d + \hat{\mu}_d + \iota.\end{aligned} \tag{52}$$

□

Because $F(w_g, w_d)$ is concave, by Lemma 5, we have $v^* - \hat{v}^* \leq \iota$. Then, for a value $v = \frac{v^*+\hat{v}^*}{2}$, there is

$$\begin{aligned} v^* &\leq v + \iota, \\ \hat{v}^* &\geq v - \iota, \end{aligned} \tag{53}$$

Replace v^* with $F(w_g, w_d^*)$ and $\hat{v}^*$ with $F(w_g^*, w_d)$. The theorem is proved.

D MD-GAN AND FL-GAN

The MD-GAN algorithm is presented in Algorithm 3. In this algorithm, we assume that there are n nodes $N = \{n_1, \cdots, n_n\}$. $w_g^{(t)}$ represents the parameters of the generator at the t-th iteration, and $w_{d_i}^{(t)}$ represents the parameters of the discriminator at the t-th iteration in node n_i. And the definition of other symbols is the same as Algorithm 1.

Algorithm 3 MD-GAN

Input: number of training iterations T, number of local update iterations τ
Output: trained parameters $w_g^{(T)}$, $w_{d_i}^{(T)} \forall n_i \in N$

```
initialize parameters of sever and nodes w_g^(0), w_{d_i}^(0) ∀n_i ∈ N
for each round t = 0, 1, ..., T − 1 do
    generate a batch of data x_g, x̂_g, ← G(z, w_g^(t))
    send x_g, x̂_g to node n_i in N = {n_1, ··· , n_n}
    for n_i ∈ N do in parallel
        receive x_g, x̂_g
        x ← mixture of x_g and local data x_{r_i}
        update local discriminator w_{d_i}^(t+1) = Update(x, w_{d_i}^(t))
        get the validity v_i = D(x̂_g, w_{d_i}^(t+1))
        send validity v_i to the server
        if t mod τ = 0 then
            random send local parameters to neighbors and receive w_{d_j}^(t+1)
            aggregate received parameters w_{d_i}^(t+1) = w_{d_i}^(t) + Σ_j^N a_ij (w_{d_j}^(t+1) − w_{d_i}^(t+1))
        end if
    end for
    receive v_i
    aggregate validity v = (1/n) Σ_i^n v_i
    update generator w_g^(t+1) ← Update(v, w_g^(t))
end for
```

The FL-GAN algorithm is presented in Algorithm 4. In this algorithm, the number of training iterations T is divisible by the number of local update iterations τ. To simplify, in line , we use the aggregate $w^{(t+1)} = \frac{1}{n}\sum_i^n w_i^{(t+1)}$ to represent aggregate the discriminator parameters $w_{g_i}^{(t+1)}$ and the generator parameters $w_{d_i}^{(t+1)}$.

Algorithm 4 FL-GAN

Input: number of training iterations T, number of local update iterations τ

Output: trained parameters $w_g^{(T)}$, $w_d^{(T)}$

initialize parameters of sever $w_g^{(0)}$, $w_d^{(0)}$
for each round $t = 0, 1, \cdots, \frac{T}{\tau} - 1$ **do**
  send $w_g^{(t)}$, $w_d^{(t)}$ to all nodes
  for $n_i \in N$ **do in parallel**
    for $\hat{t} = 0, 1, \cdots, \tau - 1$ **do**
      receive $w_g^{(t)}$, $w_d^{(t)}$
      replace $w_{g_i}^{(t)} \leftarrow w_g^{(t)}$, $w_{d_i}^{(t)} \leftarrow w_d^{(t)}$
      generate data $x \leftarrow G(z, w_{g_i}^{(t)})$
      update local discriminator $w_{d_i}^{(t+1)} = Update(x, w_{d_i}^{(t)})$
      generate data again $x \leftarrow G(z, w_{g_i}^{(t)})$
      calculate the validity of fake data $v = D(x, w_{d_i}^{(t+1)})$
      update local generator $w_{g_i}^{(t+1)} \leftarrow Update(v, w_{g_i}^{(t)})$
    end for
    send local parameters $w_i^{(t+1)}$, $w_i^{(t+1)} = [w_{g_i}^{(t+1)} : w_{d_i}^{(t+1)}]$
  end for
  receive $w_i^{(t+1)}$
  aggregate parameters $w^{(t+1)} = \frac{1}{n} \sum_i^n w_i^{(t+1)}$
end for