

# On Distributed Saddle-Point Problems (based on work [1])

Aleksandr Beznosikov

MIPT, HSE

02 May 2021

- Distributed saddle-point problem:

$$\min_{x \in X} \max_{y \in Y} f(x, y) := \frac{1}{M} \sum_{m=1}^M f_m(x, y).$$

- Relevance: GANs [3], Reinforcement Learning [4], SVM, Distributed and Federated Learning [5].

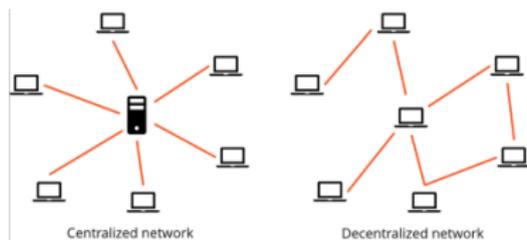


Figure: Centralized and Decentralized Learning

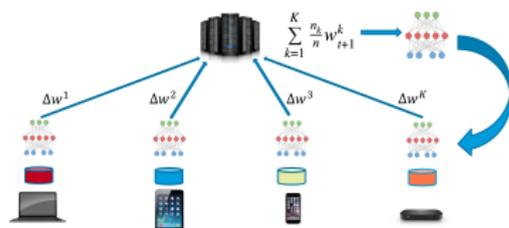


Figure: Centralized Federated Learning

# Assumptions

- Sets  $\mathcal{X} \subseteq \mathbb{R}^{n_x}$  and  $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$  are convex compact sets. For simplicity, we introduce the set  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ,  $z = (x, y)$  and the operator  $F$ :

$$F_m(z) = F_m(x, y) = \begin{pmatrix} \nabla_x f_m(x, y) \\ -\nabla_y f_m(x, y) \end{pmatrix}.$$

- We do not have access to the oracles for  $F_m(z)$ , only to some stochastic realisation  $F_m(z, \xi)$ .
- $f_m$  is stored locally on its own device. All devices are connected in a network (undirected graph  $G(\mathcal{V}, \mathcal{E})$  with diameter  $\Delta$  and condition number  $\chi$  of Laplace matrix).

# Assumptions

- **Assumption 1.**  $f(x, y)$  is Lipschitz continuous with constant  $L$ , i.e. for all  $z_1, z_2 \in \mathcal{Z}$

$$\|F_m(z_1) - F_m(z_2)\| \leq L\|z_1 - z_2\|.$$

- **Assumption 2.**  $f(x, y)$  is strongly-convex-strongly-concave with constant  $\mu$ , i.e. for all  $z_1, z_2 \in \mathcal{Z}$

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu\|z_1 - z_2\|^2.$$

- **Assumption 3.**  $F_m(z, \xi)$  is unbiased and has bounded variance, i.e. for all  $z \in \mathcal{Z}$

$$\mathbb{E}[F_m(z, \xi)] = F_m(z), \quad \mathbb{E}[\|F_m(z, \xi) - F_m(z)\|^2] \leq \sigma^2.$$

- **Assumption 4.**  $\mathcal{Z}$  – compact bounded, i.e. for all  $z, z' \in \mathcal{Z}$

$$\|z - z'\| \leq \Omega_z.$$

# Lower bounds

Lower bounds for distributed algorithms with  $K$  communications and  $T$  local iterations. Achieved on a bilinear problem  $\min_x \max_y x^T A y$  with "bad" matrix.

centralized	
sc	$\Omega \left( R_0^2 \exp \left( -\frac{32\mu K}{L\Delta} \right) + \frac{\sigma^2}{\mu^2 MT} \right)$
c	$\Omega \left( \frac{L\Omega_z^2 \Delta}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}} \right)$
decentralized	
sc	$\Omega \left( R_0^2 \exp \left( -\frac{128\mu K}{L\sqrt{\chi}} \right) + \frac{\sigma^2}{\mu^2 MT} \right)$
c	$\Omega \left( \frac{L\Omega_z^2 \sqrt{\chi}}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}} \right)$

**Table:** Lower bounds for distributed smooth stochastic strongly-convex–strongly-concave (sc) or convex-concave (c) saddle-point problems in centralized and decentralized cases.

# Centralized Extra Step Method

---

**Algorithm 1** Centralized Extra Step Method

---

**Parameters:** Step size  $\gamma \leq \frac{1}{4L}$ ; Communication rounds  $K$ , number of local steps  $T$ .

**Initialization:** Choose  $(x^0, y^0) = z^0 \in \mathcal{Z}$ ,  $k = \lfloor \frac{K}{r} \rfloor$  and batch size  $b = \lfloor \frac{T}{2k} \rfloor$ .

**for**  $t = 0, 1, 2, \dots, k$  **do**

**for** each machine  $m$  **do**

$$g_m^t = \frac{1}{b} \sum_{i=1}^b F_m(z^t, \xi_m^{t,i}), \text{ send } g_m^t,$$

**on server:**

$$z^{t+1/2} = \text{proj}_{\mathcal{Z}}(z^t - \frac{\gamma}{M} \sum_{m=1}^M g_m^t), \text{ send } z^{t+1/2},$$

**for** each machine  $m$  **do**

$$g_m^{t+1/2} = \frac{1}{b} \sum_{i=1}^b F_m(z^{t+1/2}, \xi_m^{t+1/2,i}), \text{ send } g_m^{t+1/2},$$

**on server:**

$$z^{t+1} = \text{proj}_{\mathcal{Z}}(z^t - \frac{\gamma}{M} \sum_{m=1}^M g_m^{t+1/2}), \text{ send } z^{t+1},$$

**end for**

**Output:**  $z^{k+1}$  or  $z_{avg}^{k+1}$ .

---

## Theorem

Let  $\{z^k\}_{k \geq 0}$  denote the iterates of Algorithm 1. Let Assumptions 1, 3 be satisfied. Then, if  $\gamma \leq \frac{1}{4L}$ , we have the following estimates for the distance to the solution  $z^*$  in

-  $\mu$ -strongly-convex-strongly-concave case (Assumption 2):

$$\mathbb{E}[\|z^{k+1} - z^*\|^2] = \tilde{\mathcal{O}} \left( \|z^0 - z^*\|^2 \exp\left(-\frac{\mu K}{4L\Delta}\right) + \frac{\sigma^2}{\mu^2 MT} \right),$$

- convex-concave case (Assumption 2 with  $\mu = 0$  and 4):

$$\mathbb{E}[\text{gap}(z_{\text{avg}}^{k+1})] = \mathcal{O} \left( \frac{L\Omega_z^2 \Delta}{K} + \frac{\sigma \Omega_z}{\sqrt{MT}} \right),$$

where  $z_{\text{avg}}^{k+1} = \frac{1}{k+1} \sum_{t=0}^k z^{t+1/2}$  and  $\text{gap}(z) = \max_{y'} f(x, y') - \min_{x'} f(x', y)$ .

# Decentralized Extra Step Method

Decentralized algorithms use mixing procedures [2], in this case FastMix [6].

---

**Algorithm 2** Decentralized Extra Step Method

---

**Parameters:** Step size  $\gamma \leq \frac{1}{4L}$ ; Communication rounds  $K$ , number of local calls  $T$ .

**Initialization:** Choose  $(x^0, y^0) = z^0 \in \mathcal{Z}$ ,  $z_m^0 = z^0$ ,  $k = \lfloor \frac{K}{H} \rfloor$  and batch size  $b = \lfloor \frac{T}{2k} \rfloor$ .

**for**  $t = 0, 1, 2, \dots, k$  **do**

**for** each machine  $m$  **do**

$$g_m^t = \frac{1}{b} \sum_{i=1}^b F_m(z_m^t, \xi_m^{t,i}), \quad \hat{z}_m^{t+1/2} = z_m^t - \gamma g_m^t,$$

**communication**

$$\hat{z}_1^{t+1/2}, \dots, \hat{z}_M^{t+1/2} = \text{FastMix}(\hat{z}_1^{t+1/2}, \dots, \hat{z}_M^{t+1/2}, H),$$

**for** each machine  $m$  **do**

$$z_m^{t+1/2} = \text{proj}_{\mathcal{Z}}(\hat{z}_m^{t+1/2}),$$

$$g_m^{t+1/2} = \frac{1}{b} \sum_{i=1}^b F_m(z_m^{t+1/2}, \xi_m^{t+1/2,i}),$$

$$\hat{z}_m^{t+1} = z_m^{t+1/2} - \gamma g_m^{t+1/2},$$

**communication**

$$\hat{z}_1^{t+1}, \dots, \hat{z}_M^{t+1} = \text{FastMix}(\hat{z}_1^{t+1}, \dots, \hat{z}_M^{t+1}, H),$$

**for** each machine  $m$  **do**

$$z_m^{t+1} = \text{proj}_{\mathcal{Z}}(\hat{z}_m^{t+1}),$$

**end for**

**Output:**  $\bar{z}^{k+1}$  or  $\bar{z}_{avg}^{k+1}$ .

---

## Theorem

Let  $\{z_m^k\}_{k \geq 0}$  denote the iterates of Algorithm 2. Let Assumptions 1, 3 be satisfied. Then, if  $\gamma \leq \frac{1}{4L}$ , we have the estimates for the distance to the solution  $z^*$  in

-  $\mu$ -strongly-convex-strongly-concave case (Assumption 2):

$$\mathbb{E}[\|\bar{z}^{k+1} - z^*\|^2] = \tilde{O} \left( \|z^0 - z^*\|^2 \exp \left( -\frac{\mu K}{8L\sqrt{\chi}} \right) + \frac{\sigma^2}{\mu^2 MT} \right),$$

where  $\bar{z}^{k+1} = \frac{1}{M} \sum_{m=1}^M z_m^{k+1}$ ,

- convex-concave case (Assumption 2 with  $\mu = 0$  and 4):

$$\mathbb{E}[\text{gap}(\bar{z}_{\text{avg}}^{k+1})] = \tilde{O} \left( \frac{L\Omega_z^2\sqrt{\chi}}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}} \right), \quad \bar{z}_{\text{avg}}^{k+1} = \frac{1}{M(k+1)} \sum_{t=0}^k \sum_{m=1}^M z_m^{t+1/2}.$$

**Assumption 5.** The values of the local operator are considered sufficiently close to the value of the mean operator, i.e. for all  $z \in \mathcal{Z}$

$$\|F_m(z) - F(z)\| \leq D.$$

---

**Algorithm 3** Extra Step Local SGD

---

**Parameters:** stepsize  $\gamma \leq \frac{1}{6HL_{\max}}$ ; number of local steps  $T$ , sets  $I$  of communications steps for  $x$  and  $y$  ( $|I| = K$ ).

**Initialization:** Choose  $(x^0, y^0) = z^0 \in \mathcal{Z}$ , for all  $m$   $z_m^0 = z^0$  and  $\bar{z} = z^0$ .

**for**  $k = 0, 1, 2, \dots, T$  **do**

**for** each machine  $m$  **do**

$$z_m^{k+1/2} = \text{proj}_{\mathcal{Z}}(z_m^k - \gamma F_m(z_m^k, \xi_m^k)),$$

$$z_m^{k+1} = \text{proj}_{\mathcal{Z}}(z_m^k - \gamma F_m(z_m^{k+1/2}, \xi_m^{k+1/2})),$$

  if  $k \in I$ , send  $z_m^{k+1}$  on server,

**on server:**

$$\text{if } k \in I \text{ compute } \bar{z} = \frac{1}{M} \sum_{m=1}^M z_m^{k+1}, \text{ send } \bar{z}.$$

**for** each machine  $m$  **do**

    if  $k \in I$ , get  $\bar{z}$  and set  $z_m^{k+1} = \bar{z}$ ,

**end for**

**Output:**  $\bar{z}$ .

---

## Theorem

Let  $\{z_m^k\}_{k \geq 0}$  denote the iterates of Algorithm 3 and  $\bar{z} = \bar{z}^{T+1}$  is an output. Let Assumptions 1(l), 2, 3 and 5 be satisfied. Also let  $H = \max_p |k_{p+1} - k_p|$  – maximum distance between moments of communication ( $k_p \in I$ ). Then, if  $\gamma \leq \frac{1}{6HL_{\max}}$ , we have the following estimate for the distance to the solution  $z^*$ :

$$\mathbb{E}[\|\bar{z}^{T+1} - z^*\|^2] \leq \left(1 - \frac{\mu\gamma}{2}\right)^T \|\bar{z}^0 - z^*\|^2 + \frac{20\gamma\sigma^2}{\mu M} + \frac{250\gamma^2 H^3 L_{\max}^2 (2\sigma^2 + D^2)}{\mu^2}.$$

## Corollary

Let  $\alpha = \frac{12HL_{\max}}{\mu}$ ,  $\gamma = \frac{2}{\mu\alpha} \leq \frac{1}{6HL_{\max}}$  and  $T = \alpha \log \alpha^2$ , then we get:

$$\mathbb{E}[\|\bar{z}^{T+1} - z^*\|^2] \leq \frac{\|\bar{z}^0 - z^*\|^2 \log^2 \alpha^2}{T^2} + \frac{20\sigma^2 \log \alpha^2}{\mu^2 MT} + \frac{250H^3 L_{\max}^2 \log^2 \alpha^2 (2\sigma^2 + D^2)}{\mu^4 T^2}.$$

It can be seen that if we take  $H = \mathcal{O}(T^{1/3}/M^{1/3})$ , we have a convergence rate of about  $\mathcal{O}(1/MT)$ . The estimate for the number of communication rounds is  $C = T/H = \Omega(M^{1/3} T^{2/3})$ .

# When Local method better than Optimal?

Bilinear problem:

$$\min_{x,y \in [-1;1]^n} \max \frac{1}{M} \sum_{m=1}^M \left( x^T A_m y + b_m^T x + c_m^T y \right),$$

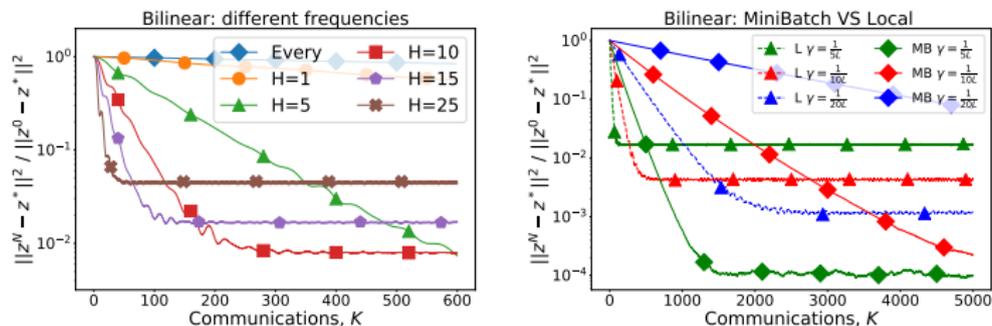


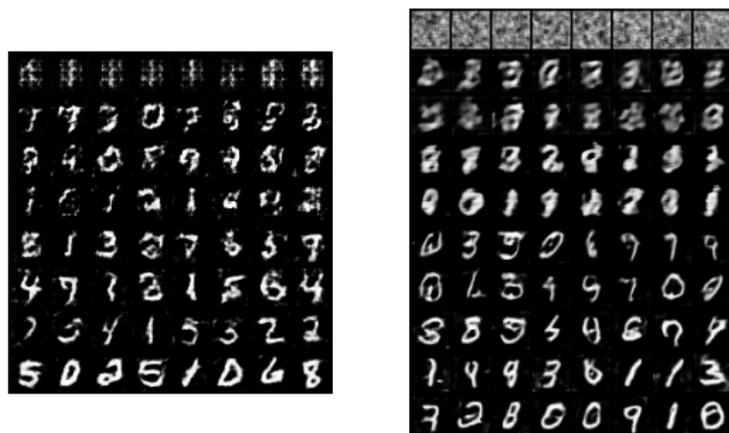
Figure: Left: comparison of Algorithm 3 with different communication frequencies  $H$ , as well as Algorithm 1 with batch size 1 (blue line – "Every"). Right: comparison of Algorithm 3 (L) with communication frequencies  $H = 3$  and Algorithm 1 (MB) with batch size 6.

The experiment simulates a classic federated learning setting:

- Each of the nodes has highly heterogeneous data. In the case of the MNIST dataset, each node is given unique digits.
- Devices rarely communicate with server - once every 20 epochs.
- Privacy - devices do not send local data, but only model parameters.
- In spite of federated restrictions, a global models (generator and discriminator) are trained with taking into account all local data.

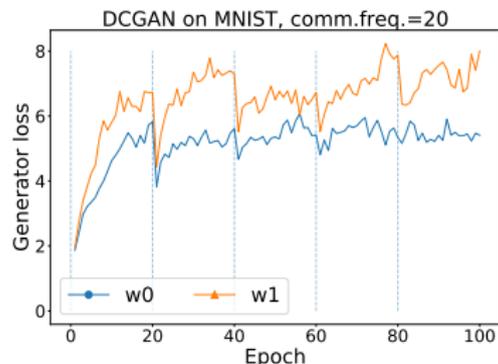
# Federated GANs

Results of Federated GANs training:

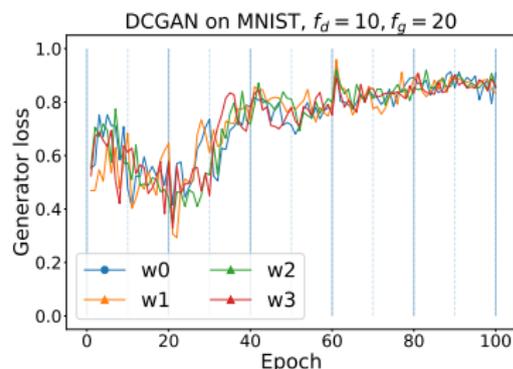


**Figure:** Heterogeneous case (each device has its our unique digits). Digits generated by global generator during training. 2 nodes, Local SGD (left) and 4 nodes, Local Adam (right).

# Federated GANs

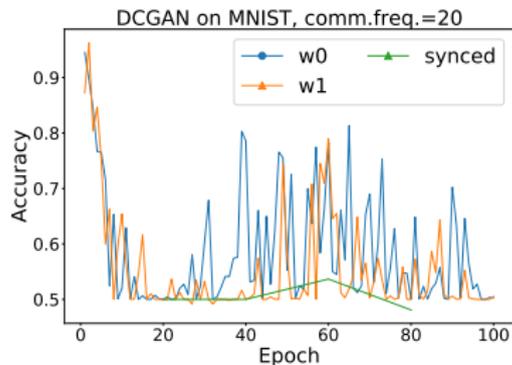


**Figure:** Generator empirical loss in experiment with 2 nodes, Local SGD,  $H_g = H_d = 20$

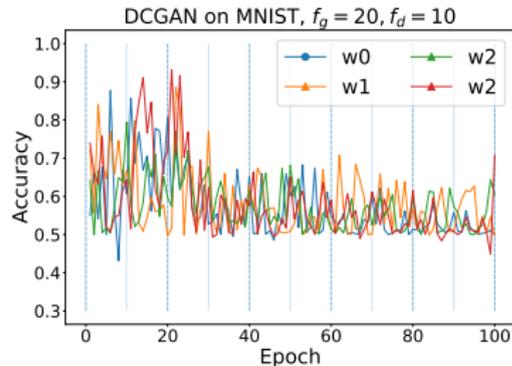


**Figure:** Generator empirical loss in experiment with 4 nodes, Local Adam  $H_g = H_d = 20$

# Federated GANs



**Figure:** Discriminator accuracy in experiment with 2 nodes, Local SGD,  $H_g = H_d = 20$ .



**Figure:** Discriminator accuracy in experiment with 4 nodes, Local Adam,  $H_g = H_d = 20$

-  Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov.  
Distributed saddle-point problems: Lower bounds, optimal algorithms and federated gans.  
2020.
-  Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah.  
Randomized gossip algorithms.  
*IEEE transactions on information theory*, 52(6):2508–2530, 2006.
-  Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.  
Generative adversarial networks, 2014.
-  Yujia Jin and Aaron Sidford.  
Efficiently solving MDPs with stochastic mirror descent.  
In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4890–4900. PMLR, 13–18 Jul 2020.



Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al.

Advances and open problems in federated learning.

*arXiv preprint arXiv:1912.04977*, 2019.



Ji Liu and A Stephen Morse.

Accelerated linear iterations for distributed averaging.

*Annual Reviews in Control*, 35(2):160–165, 2011.