Near-Optimal Decentralized Algorithms for Saddle Point Problems over Time-Varying Networks (based on work [1])

Aleksandr Beznosikov

MIPT, HSE, Yandex

27 September 2021

Statement

- 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 [2], Reinforcement Learning [3], SVM, Distributed and Federated Learning [4].

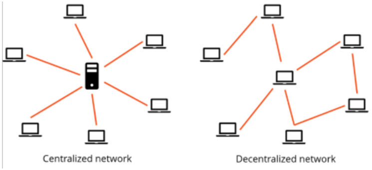

Figure: Centralized and Decentralized Learning

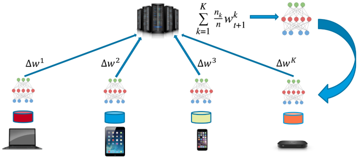

Figure: Centralized Federated Learning

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:

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

- f_m is stored locally on its own device. All devices are connected in a network (**time-varying** undirected graph $G(\mathcal{V}, \mathcal{E}(t))$ with condition number smaller than χ).

Assumptions

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

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

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

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

Class of algorithms

Definition

Each device m has its own local memories $\mathcal{M}_m^x$ and $\mathcal{M}_m^y$ for the x- and y-variables, respectively–with initialization $\mathcal{M}_m^x = \mathcal{M}_m^y = \{0\}$. $\mathcal{M}_m^x$ and $\mathcal{M}_m^x$ are updated as follows:

- **Local computation:** Each device m computes and adds to its $\mathcal{M}_m^x$ and $\mathcal{M}_m^y$ a finite number of points x, y, each satisfying

$$x \in \text{span}\left\{x' \ , \ \nabla_x f_m(x'', y'')\right\}, \quad y \in \text{span}\left\{y' \ , \ \nabla_y f_m(x'', y'')\right\},$$

for given $x', x'' \in \mathcal{M}_m^x$ and $y', y'' \in \mathcal{M}_m^y$.

Class of algorithms

Definition

• **Communication:** Based upon communication round among neighbouring nodes at the moment t, $\mathcal{M}_m^x$ and $\mathcal{M}_m^y$ are updated according to

$$\mathcal{M}_m^x := \mathrm{span}\left\{ \bigcup_{(i,m)\in\mathcal{E}(t)} \mathcal{M}_i^x \right\}, \quad \mathcal{M}_m^y := \mathrm{span}\left\{ \bigcup_{(i,m)\in\mathcal{E}(t)} \mathcal{M}_i^y \right\}.$$

• **Output:** The final global output at the current moment of time is calculated as:

$$x \in \mathrm{span}\left\{ \bigcup_{m=1}^{M} \mathcal{M}_m^x \right\}, \; y \in \mathrm{span}\left\{ \bigcup_{m=1}^{M} \mathcal{M}_m^y \right\}.$$

Lower bounds

Theorem

For any L and μ, there exists a SPP with $\mathcal{Z} = \mathcal{R}^{2d}$ (where d is sufficiently large) and non-zero solution y^*. All local functions f_m of this problem are L-smooth, μ-strongly-convex-strongly-concave. Then, for any $\chi \geq 1$, there exists a sequence of gossip matrices $W(t)$ over the connected (at each moment) graph $\mathcal{G}(t)$ with condition number χ, such that for any decentralized algorithm the number of communication rounds required to obtain a ε-solution is lower bounded by

$$\Omega\left(\chi\frac{L}{\mu}\cdot\log\left(\frac{\|y^*\|^2}{\varepsilon}\right)\right).$$

Additionally, we can get a lower bound for the number of local calculations on each of the devices:

$$\Omega\left(\frac{L}{\mu}\cdot\log\left(\frac{\|y^*\|^2}{\varepsilon}\right)\right).$$

Centralized Extra Step Method

Algorithm 1 Gossip Algorithm (`Gossip`)

Parameters: Vectors $z_1, ..., z_M$, communic. rounds H.
Initialization: Construct matrix $\mathbf{z}$ with rows $z_1^T, ..., z_M^T$.
Choose $\mathbf{z}^0 = \mathbf{z}$.
for $h = 0, 1, 2, \dots, H$ **do**
  $\mathbf{z}^{h+1} = \tilde{W}(h) \cdot \mathbf{z}^h$
end for
Output: rows $z_1, ..., z_M$ of $\mathbf{z}^{H+1}$.

Algorithm 2 Time-Varying Decentralized Extra Step Method (`TVDESM`)

Parameters: Stepsize $\gamma \leq \frac{1}{4L}$, number of `Gossip` steps H.
Initialization: Choose $(x^0, y^0) = z^0 \in \mathcal{Z}$, $z_m^0 = z^0$.
for $k = 0, 1, 2, \dots,$ **do**
  Each machine m computes $\hat{z}_m^{k+1/2} = z_m^k - \gamma \cdot F_m(z_m^k)$
  Communication: $\tilde{z}_1^{k+1/2}, \dots, \tilde{z}_M^{k+1/2} =$`Gossip`$(\hat{z}_1^{k+1/2}, \dots, \hat{z}_M^{k+1/2}, H)$
  Each machine m computes $z_m^{k+1/2} = \text{proj}_{\mathcal{Z}}(\tilde{z}_m^{k+1/2})$,
  Each machine m computes $\hat{z}_m^{k+1} = z_m^k - \gamma \cdot F_m(z_m^{k+1/2})$
  Communication: $\tilde{z}_1^{k+1}, \dots, \tilde{z}_M^{k+1} =$`Gossip`$(\hat{z}_1^{k+1}, \dots, \hat{z}_M^{k+1}, H)$
  Each machine m computes $z_m^{k+1} = \text{proj}_{\mathcal{Z}}(\tilde{z}_m^{k+1})$
end for

Convergence

Theorem

Let we use Algorithm 2 for solving distributed SPP. And let Assumptions 1 and 2 be satisfied for all f_m. Then, if $\gamma \leq \frac{1}{4L}$, the number of communication rounds required to obtain a ε-solution is upper bounded by

$$\tilde{\mathcal{O}}\left(\chi \frac{L}{\mu}\right).$$

Additionally, one can obtain upper bounds for the number of local calculations on each of the devices:

$$\mathcal{O}\left(\frac{L}{\mu} \cdot \log\left(\frac{\|z^0 - z^*\|^2}{\varepsilon}\right)\right).$$

Aleksandr Beznosikov, Alexander Rogozin, Dmitry Kovalev, and Alexander Gasnikov.
Optimal decentralized algorithms for saddle point problems over time-varying networks.
arXiv preprint arXiv:2107.05957, 2021.

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.