Near-Optimal Distributed Minimax Optimization under the Second-Order Similarity

Qihao Zhou[1] Haishan Ye[2,3] Luo Luo[1,4]

[1]School of Data Science, Fudan University
[2]School of Management, Xi'an Jiaotong University
[3]SGIT AI Lab, State Grid Corporation of China
[4]Shanghai Key Laboratory for Contemporary Applied Mathematics

November 1, 2024

Problem Setup

We consider the distributed minimax optimization problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) := \frac{1}{n}\sum_{i=1}^{n} f_i(x,y),$$

where f_i is the differentiable local function associated with i-th node, and $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ are the constraint sets.

Centralized setting: one server node and $n-1$ client nodes.

Let $z = [x; y] \in \mathcal{Z}$ and $F(z) = [\nabla_x f; -\nabla_y f]$. We assume $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is closed and convex, each f_i is L-smooth and convex-concave, f is strongly-convex-strongly-concave with $\mu \geq 0$, and the similarity as below.

Assumption

*The local functions $f_1, \dots, f_n : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ are twice differentiable and hold the δ-**second-order similarity**, i.e., there exists $\delta > 0$ such that*

$$\left\|\nabla^2 f_i(x,y) - \nabla^2 f(x,y)\right\| \leq \delta$$

for all $i \in [n]$, $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$.

Related Work in Convex-Concave Case

We measure the sub-optimality by duality gap, that is

$$\mathrm{Gap}(z) := \max_{y' \in \mathcal{Y}} f(x, y') - \min_{x' \in \mathcal{X}} f(x', y).$$

Methods	CR	CC	LGC
EG [4]	$\mathcal{O}(\frac{LD^2}{\varepsilon})$	$\mathcal{O}(\frac{nLD^2}{\varepsilon})$	$\mathcal{O}(\frac{nLD^2}{\varepsilon})$
SMMDS [2]	$\mathcal{O}(\frac{\delta D^2}{\varepsilon})$	$\mathcal{O}(\frac{n\delta D^2}{\varepsilon})$	$\tilde{\mathcal{O}}(\frac{(n\delta+L)D^2}{\varepsilon}\log\frac{1}{\varepsilon})$
EGS [5]	$\mathcal{O}(\frac{\delta D^2}{\varepsilon})$	$\mathcal{O}(\frac{n\delta D^2}{\varepsilon})$	$\mathcal{O}(\frac{(n\delta+L)D^2}{\varepsilon})$
SVOGS	$\mathcal{O}(\frac{\delta D^2}{\varepsilon})$	$\mathcal{O}(n + \frac{\sqrt{n}\delta D^2}{\varepsilon})$	$\tilde{\mathcal{O}}(n + \frac{(\sqrt{n}\delta+L)D^2}{\varepsilon}\log\frac{1}{\varepsilon})$
Lower Bounds	$\Omega(\frac{\delta D^2}{\varepsilon})$	$\Omega(n + \frac{\sqrt{n}\delta D^2}{\varepsilon})$	$\Omega(n + \frac{(\sqrt{n}\delta+L)D^2}{\varepsilon})$

Abbr.: CR=Communication Rounds, CC=Communication Complexity, LGC=Local Gradient Calls.

Related Work in Strongly-Convex-Strongly-Concave Case

We measure the sub-optimality by $\mathbb{E}[\|z - z^*\|^2]$.

Methods	CR	CC	LGC
EG [4]	$\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon})$	$\mathcal{O}(\frac{nL}{\mu}\log\frac{1}{\varepsilon})$	$\mathcal{O}(\frac{nL}{\mu}\log\frac{1}{\varepsilon})$
SMMDS [2]	$\mathcal{O}(\frac{\delta}{\mu}\log\frac{1}{\varepsilon})$	$\mathcal{O}(\frac{n\delta}{\mu}\log\frac{1}{\varepsilon})$	$\tilde{\mathcal{O}}(\frac{n\delta+L}{\mu}\log\frac{1}{\varepsilon})$
EGS [5]	$\mathcal{O}(\frac{\delta}{\mu}\log\frac{1}{\varepsilon})$	$\mathcal{O}(\frac{n\delta}{\mu}\log\frac{1}{\varepsilon})$	$\mathcal{O}(\frac{n\delta+L}{\mu}\log\frac{1}{\varepsilon})$
OMASHA [1]†	$\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon})$	$\mathcal{O}((n+\frac{\sqrt{n}\delta+L}{\mu})\log\frac{1}{\varepsilon})$	$\mathcal{O}(\frac{nL}{\mu}\log\frac{1}{\varepsilon})$
TPA [3]†	$\mathcal{O}((n+\frac{\sqrt{n}\delta}{\mu})\log\frac{1}{\varepsilon})$	$\mathcal{O}((n+\frac{\sqrt{n}\delta}{\mu})\log\frac{1}{\varepsilon})$	$\mathcal{O}((n+\frac{\sqrt{n}L}{\delta}+\frac{L}{\mu})\log\frac{1}{\varepsilon})$
TPAPP (a) [3]♯	$\mathcal{O}((n+\frac{\sqrt{n}\delta}{\mu})\log\frac{1}{\varepsilon})$	$\mathcal{O}((n+\frac{\sqrt{n}\delta}{\mu})\log\frac{1}{\varepsilon})$	$\mathcal{O}((n+\frac{\sqrt{n}L}{\delta}+\frac{L}{\mu})\log\frac{1}{\varepsilon})$
TPAPP (b) [3]♯	$\mathcal{O}((n+\frac{\sqrt{n}\delta+L}{\mu})\log\frac{1}{\varepsilon})$	$\mathcal{O}((n+\frac{\sqrt{n}\delta+L}{\mu})\log\frac{1}{\varepsilon})$	$\tilde{\mathcal{O}}((n+\frac{\sqrt{n}\delta+L}{\mu})\log\frac{1}{\varepsilon})$
SVOGS	$\mathcal{O}(\frac{\delta}{\mu}\log\frac{1}{\varepsilon})$	$\mathcal{O}((n+\frac{\sqrt{n}\delta}{\mu})\log\frac{1}{\varepsilon})$	$\tilde{\mathcal{O}}((n+\frac{\sqrt{n}\delta+L}{\mu})\log\frac{1}{\varepsilon})$
Lower Bounds	$\Omega(\frac{\delta}{\mu}\log\frac{1}{\varepsilon})$	$\Omega((n+\frac{\sqrt{n}\delta}{\mu})\log\frac{1}{\varepsilon})$	$\Omega((n+\frac{\sqrt{n}\delta+L}{\mu})\log\frac{1}{\varepsilon})$

Abbr.: CR=Communication Rounds, CC=Communication Complexity, LGC=Local Gradient Calls.

†:Compressors used. ♯Different inner steps. $H_a = \lceil L/(\sqrt{n}\delta)\rceil$ and $H_b = \lceil 8\log(40nL/\mu)\rceil$.

Motivation of SVOGS

Gradient Sliding: $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y) := \underbrace{\frac{1}{n}\sum_{i=1}^{n}(f_i(x,y)-f_1(x,y))}_{g(x,y):=f(x,y)-f_1(x,y)} + f_1(x,y).$

OGDA: $z^{k+1} = \mathcal{P}_{\mathcal{Z}}\left(z^k - \eta\underbrace{(F(z^k)+F(z^k)-F(z^{k-1}))}_{\text{optimistic gradient}}\right).$

Approximation of $g(x,y)$:

$$\hat{g}(x,y) = g(x^k,y^k) + \langle \underbrace{\nabla_x g(x^k,y^k) + \nabla_x g(x^k,y^k) - \nabla_x g(x^{k-1},y^{k-1})}_{\text{optimistic gradient with respect to } x}, x - x^k\rangle + \frac{1}{2\eta}\left\|x - x^k\right\|^2$$
$$+ \langle \underbrace{\nabla_y g(x^k,y^k) + \nabla_y g(x^k,y^k) - \nabla_y g(x^{k-1},y^{k-1})}_{\text{optimistic gradient with respect to } y}, y - y^k\rangle - \frac{1}{2\eta}\left\|y - y^k\right\|^2.$$

Update: $(x^{k+1},y^{k+1}) \approx \arg\min_{\hat{x}\in\mathcal{X}}\max_{\hat{y}\in\mathcal{Y}} \hat{g}(\hat{x},\hat{y}) + f_1(\hat{x},\hat{y}).$

Mini-Batch (snapshot point w update with probability $\Theta(1/\sqrt{n})$):

$$G(z^k) + G(z^k) - G(z^{k-1}) \approx \frac{1}{|\mathcal{S}^k|}\sum_{j\in\mathcal{S}^k}\left(G(w^{k-1}) + G_j(z^k) - G_j(w^{k-1}) + \underbrace{\alpha(G_j(z^k) - G_j(z^{k-1}))}_{\text{momentum term}}\right).$$

Lyapunov Function

We analyze the convergence of SVOGS by establishing the Lyapunov function ($\mu = 0$ in convex-concave case) as

$$\Phi^k := \left(\frac{1}{\eta} + \mu\right) \|z^k - z^*\|^2 + 2\langle F(z^{k-1}) - F_1(z^{k-1}) - F(z^k) + F_1(z^k), z^k - z^*\rangle$$
$$+ \frac{1}{64\eta}\|z^k - z^{k-1}\|^2 + \frac{\gamma}{4\eta}\|w^{k-1} - z^k\|^2 + \frac{(2\gamma + \eta\mu)}{2p\eta}\|w^k - z^*\|^2.$$

Choosing $\eta \le 1/(32\delta)$ leads to the non-negativity of Lyapunov function.

Lemma

Suppose assumptions hold with $0 \le \mu \le \delta \le L$, running SVOGS with well chosen parameters, then we have

$$\mathbb{E}[\Phi^{k+1}] \le \max\left\{1 - \frac{\eta\mu}{6}, 1 - \frac{p\eta\mu}{2\gamma + \eta\mu}\right\} \mathbb{E}[\Phi^k]$$
$$- \frac{1}{16\eta}\mathbb{E}\left[\|z^k - \hat{u}^k\|^2\right] - \frac{\gamma}{2\eta}\mathbb{E}\left[\|w^k - \hat{u}^k\|^2\right].$$

Convergence: General Convex Concave Case

Theorem

Suppose assumptions hold with $0 = \mu < \delta \le L$ and $D > 0$, running SVOGS with well chosen parameters, then we have

$$\mathbb{E}\left[\max_{z \in \mathcal{Z}} \frac{1}{K} \sum_{k=0}^{K-1} \langle F(u^k), u^k - z \rangle\right] \le \frac{10D^2}{\eta K} + \frac{\varepsilon}{2}, \quad \textit{where} \;\; u_{\text{avg}}^K = \frac{1}{K} \sum_{k=0}^{K-1} u^k.$$

Corollary

Following the theorem, we can achieve $\mathbb{E}[\mathrm{Gap}(u_{\text{avg}}^K)] \le \varepsilon$ within communication rounds of $\mathcal{O}(\delta D^2/\varepsilon)$, communication complexity of $\mathcal{O}(n + \sqrt{n}\delta D^2/\varepsilon)$, and local gradient complexity of $\tilde{\mathcal{O}}(n + (\sqrt{n}\delta + L)D^2/\varepsilon \log(1/\varepsilon))$, where $u_{\text{avg}}^K = \frac{1}{K}\sum_{k=0}^{K-1} u^k$.

Convergence: Strongly Convex Strongly Concave Case

Theorem

Suppose assumptions hold with $0 < \mu \le \delta \le L$ and $D > 0$, running SVOGS with well chosen parameters, then we have

$$\mathbb{E}[\Phi^K] \le \max\left\{1 - \frac{\eta\mu}{6}, 1 - \frac{p\eta\mu}{2\gamma + \eta\mu}\right\}^K \Phi^0.$$

Corollary

Following the theorem, we can achieve $\mathbb{E}\left[\|z^K - z^*\|^2\right] \le \varepsilon$ within communication rounds of $\mathcal{O}(\delta/\mu \log(1/\varepsilon))$, communication complexity of $\mathcal{O}((n + \sqrt{n}\delta/\mu)\log(1/\varepsilon))$, and local gradient complexity of $\tilde{\mathcal{O}}((n + (\sqrt{n}\delta + L)/\mu)\log(1/\varepsilon))$.

Make the Gradient Small

Other than duality gap, we define gradient mapping $\mathcal{F}_\tau(z) := (z - \mathcal{P}_{\mathcal{Z}}(z - \tau F(z)))/\tau$ and measure the sub-optimality by $\mathbb{E}[\|\mathcal{F}_\tau(z)\|^2]$.

For smooth convex-concave f, we consider the problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \hat{f}(x,y) := f(x,y) + \frac{\lambda}{2}\left\|x - x^0\right\|^2 - \frac{\lambda}{2}\left\|y - y^0\right\|^2,$$

where $\hat{f}$ is strongly-convex-strongly-concave. Take $\lambda = \mathcal{O}(\sqrt{\varepsilon}/D)$, we have the results.

Methods	CR	CC	LGC
TPAPP [3][§]	$\mathcal{O}(\frac{n\delta^2D^2}{\varepsilon})$	$\mathcal{O}(\frac{n\delta^2D^2}{\varepsilon})$	$\mathcal{O}(\frac{n^2\delta^4L^2D^6}{\varepsilon^3})$
SVOGS	$\tilde{\mathcal{O}}(\frac{\delta D}{\sqrt{\varepsilon}}\log\frac{1}{\varepsilon})$	$\tilde{\mathcal{O}}((n + \frac{\sqrt{n}\delta D}{\sqrt{\varepsilon}})\log\frac{1}{\varepsilon})$	$\tilde{\mathcal{O}}((n + \frac{(\sqrt{n}\delta + L)D}{\sqrt{\varepsilon}})\log\frac{1}{\varepsilon})$

Abbr.: CR=Communication Rounds, CC=Communication Complexity, LGC=Local Gradient Calls.

[§] Additionally assume $\mathcal{Z} = \mathbb{R}^d$ and the sequence generated is bounded by $D > 0$.

Lower Bounds

Convex-concave case (to obtain $\mathbb{E}[\mathrm{Gap}(z)] < \varepsilon$):

Methods	CR	CC	LGC
SVOGS	$\mathcal{O}(\frac{\delta D^2}{\varepsilon})$	$\mathcal{O}(n + \frac{\sqrt{n}\delta D^2}{\varepsilon})$	$\tilde{\mathcal{O}}(n + \frac{(\sqrt{n}\delta + L)D^2}{\varepsilon} \log \frac{1}{\varepsilon})$
Lower Bounds	$\Omega(\frac{\delta D^2}{\varepsilon})$	$\Omega(n + \frac{\sqrt{n}\delta D^2}{\varepsilon})$	$\Omega(n + \frac{(\sqrt{n}\delta + L)D^2}{\varepsilon})$

Strongly-convex-strongly-concave case (to obtain $\mathbb{E}[\|z - z^*\|^2] < \varepsilon$):

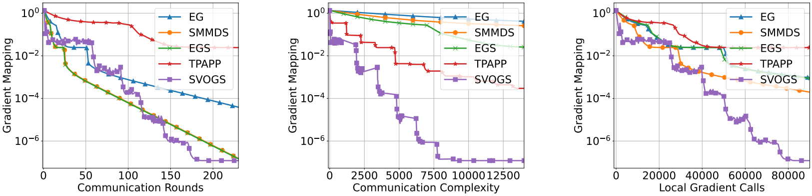
Methods	CR	CC	LGC
SVOGS	$\mathcal{O}(\frac{\delta}{\mu} \log \frac{1}{\varepsilon})$	$\mathcal{O}((n + \frac{\sqrt{n}\delta}{\mu}) \log \frac{1}{\varepsilon})$	$\tilde{\mathcal{O}}((n + \frac{\sqrt{n}\delta + L}{\mu}) \log \frac{1}{\varepsilon})$
Lower Bounds	$\Omega(\frac{\delta}{\mu} \log \frac{1}{\varepsilon})^\flat$	$\Omega((n + \frac{\sqrt{n}\delta}{\mu}) \log \frac{1}{\varepsilon})^\natural$	$\Omega((n + \frac{\sqrt{n}\delta + L}{\mu}) \log \frac{1}{\varepsilon})$

Abbr.: CR=Communication Rounds, CC=Communication Complexity, LGC=Local Gradient Calls.

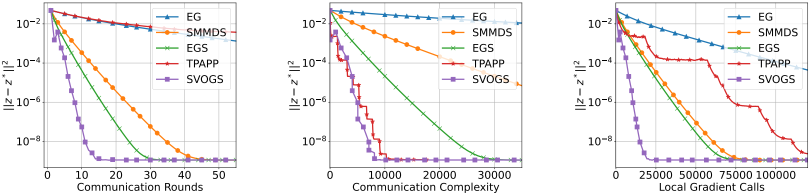
$^\flat$Given by Beznosikov et al. [2]. $^\natural$Given by Beznosikov et al. [3].

Experiments: Robust Linear Regression

$$\min_{\|x\|_1 \le R_x} \max_{\|y\|_2 \le R_y} \frac{1}{2N} \sum_{i=1}^{N} \left(x^\top (a_i + y) - b_i \right)^2 .$$



$$\min_{x \in \mathbb{R}^{d'}} \max_{y \in \mathbb{R}^{d'}} \frac{1}{2N} \sum_{i=1}^{N} \left(x^\top (a_i + y) - b_i \right)^2 + \frac{\lambda}{2}\|x\|^2 - \frac{\beta}{2}\|y\|^2 .$$



Summary

SVOGS compared to former methods

- A novel method combining OGDA, variance reduction and mini-batch
- Effective in three different complexity measures
- All the lower bounds (nearly) matched at the same time

Future work

- Non-centralized distributed minimax optimization
- Mini-batch for non-convex minimization

Thanks!

References

[1] Aleksandr Beznosikov and Alexander Gasnikov. Compression and data similarity: Combination of two techniques for communication-efficient solving of distributed variational inequalities. In *International Conference on Optimization and Applications*, 2022.

[2] Aleksandr Beznosikov, Gesualdo Scutari, Alexander Rogozin, and Alexander Gasnikov. Distributed saddle-point problems under data similarity. *Advances in Neural Information Processing Systems*, 2021.

[3] Aleksandr Beznosikov, Martin Takác, and Alexander Gasnikov. Similarity, compression and local steps: three pillars of efficient communications for distributed variational inequalities. *Advances in Neural Information Processing Systems*, 2023.

[4] Galina M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[5] Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. *Advances in Neural Information Processing Systems*, 2022.