Compression and Data Similarity: Combination of Two Techniques for Communication-Efficient Solving of Distributed Variational Inequalities

Aleksandr Beznosikov

MIPT

26 September 2022

Distributed Variational Inequalities

Definition

Find $z^* \in \mathbb{R}^d$ such that $\langle F(z^*), z - z^* \rangle + g(z) - g(z^*) \geq 0, \;\forall z \in \mathbb{R}^d$,

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is an operator, and $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous convex function. We assume that the training data describing F is *distributed* across M workers/nodes/clients

$$F(z) \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} F_m(z),$$

where $F_m : \mathbb{R}^d \to \mathbb{R}^d$ for all $m \in \{1, 2, \ldots, M\}$.

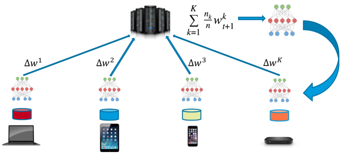


Figure: Centralized Distributed/Federated Learning

Distributed Variational Inequalities

- Minimization problem:

$$\min_{z \in \mathbb{R}^d} f(z) + g(z).$$

We can take $F(z) \stackrel{\text{def}}{=} \nabla f(z)$.

- Saddle point problem:

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} g_1(x) + f(x, y) - g_2(y).$$

Here $F(z) \stackrel{\text{def}}{=} F(x, y) = [\nabla_x f(x, y), -\nabla_y f(x, y)]$.
Examples: adversarial training/robust optimization, GANs, RL, image denoising, SVM, Lagrange multipliers.

- Fixed point problem:

$$\text{Find } z^* \in \mathbb{R}^d \text{ such that } T(z^*) = z^*,$$

where $T : \mathbb{R}^d \to \mathbb{R}^d$ is an operator. We can take $F(z) = z - T(z)$.

Assumptions

Definition (Lipschitzness)

The operator F is L-Lipschitz continuous, i.e. for all $z_1, z_2 \in \mathbb{R}^d$ we have $\|F(z_1) - F(z_2)\| \leq L\|z_1 - z_2\|$.

For saddle point problems, these properties are equivalent to smoothness.

Definition (Strong monotonicity)

The operator F is μ-strongly monotone, i.e. for all $z_1, z_2 \in \mathbb{R}^d$ we have $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu\|z_1 - z_2\|^2$.

For saddle point problems, these properties are equivalent to convexity.

Assumptions

Definition (δ-relatedness)

Each operator F_m is δ-related. It means that each operator $F_m - F$ is δ-Lipschitz continuous, i.e. for all $u, v \in \mathbb{R}^d$ we have $\|F_m(u) - F(u) - F_m(v) + F(v)\| \leq \delta \|u - v\|$.

For minimization problems:

$$\|\nabla^2 f(z) - \nabla^2 f_m(z)\| \leq \delta,$$

For saddle point problems:

$$\begin{aligned}
\|\nabla^2_{xx} f(x, y) - \nabla^2_{xx} f_m(x, y)\| &\leq \delta, \\
\|\nabla^2_{xy} f(x, y) - \nabla^2_{xy} f_m(x, y)\| &\leq \delta, \\
\|\nabla^2_{yy} f(x, y) - \nabla^2_{yy} f_m(x, y)\| &\leq \delta.
\end{aligned}$$

For uniform splitting of the data $\delta = \tilde{\mathcal{O}}\left(\frac{L}{\sqrt{b}}\right)$, where b is the number of local data points on each of the workers.

Optimistic MASHA

Algorithm 1 Optimistic MASHA

Parameters: Stepsize $\gamma > 0$, parameter τ, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$.
Server sends to devices $z^0 = w^0 = w^{-1}$ and devices compute $F_m(z^0)$ and send to server and get $F(z^0)$
for $k = 0, 1, 2, \ldots, K-1$ **do**
    for each device m in parallel **do**
        Compute $F_m(z^k)$
        $\delta_m^k = F_m(z^k) - F_m(w^{k-1}) + \alpha[F_m(z^k) - F_m(z^{k-1})]$
        Send $Q_m\left(\delta_m^k\right)$ to server
    end for
    for server **do**
        Compute $\frac{1}{M}\sum\limits_{m=1}^{M} Q_m(\delta_m^k)$ and send to devices
        Sends to devices b_k: 1 with probability γ, 0 with. probability $1-\gamma$
    end for
    for each device m in parallel **do**
        $\Delta^k = \frac{1}{M}\sum\limits_{m=1}^{M} Q_m^{\text{dev}}(\delta_m^k) + F(w^{k-1})$
        $z^{k+1} = \text{prox}_{\eta g}\left(z^k + \gamma(w^k - z^k) - \eta\Delta^k\right)$
        if $b_k = 1$ **then**
            $w^{k+1} = z^k$
            Compute $F_m(w^{k+1})$ and send it to server
            Get $F(w^{k+1})$ as a response from server
        else
            $w^{k+1} = w^k$
        end if
    end for
end for

Optimistic MASHA: Permutation compressors

Definition (Permutation compressors [5])

• **for $d \geq M$.** Assume that $d \geq M$ and $d = qM$, where $q \geq 1$ is an integer. Let $\pi = (\pi_1, \dots, \pi_d)$ be a random permutation of $\{1, \dots, d\}$. Then for all $u \in \mathbb{R}^d$ and each $m \in \{1, 2, \dots, M\}$ we define

$$Q_m(u) \overset{\text{def}}{=} M \cdot \sum_{i=q(m-1)+1}^{qm} u_{\pi_i} e_{\pi_i}.$$

• **for $d \leq M$.** Assume that $M \geq d$, $M > 1$ and $M = qd$, where $q \geq 1$ is an integer. Define the multiset $S \overset{\text{def}}{=} \{1, \dots, 1, 2, \dots, 2, \dots, d, \dots, d\}$, where each number occurs precisely q times. Let $\pi = (\pi_1, \dots, \pi_M)$ be a random permutation of S. Then for all $u \in \mathbb{R}^d$ and each $m \in \{1, 2, \dots, M\}$ we define

$$Q_m(u) \overset{\text{def}}{=} d u_{\pi_m} e_{\pi_m}.$$

Convergence of Optimistic MASHA

Theorem

Let Assumption on Lipschitzness, strong monotonicity and δ-relatedness are satisfied. Then for some step η and momentums α and γ the following estimates on Optimistic MASHA number of bits to achieve ε-solution holds

$$O\left(\left[\frac{L}{M\mu} + \frac{\delta}{\sqrt{M}\mu}\right]\log\frac{1}{\varepsilon}\right)$$

Convergence: comparison

Table: Summary of complexities on the number of transmitted information for different approaches to communication bottleneck.
Notation: μ = constant of strong monotonicity of the operator F, L = Lipschitz constant of the operator F, δ = relatedness constant, M = number of devices, b = local data size, ε = precision of the solution.

Method	**Reference**	**Technique**	**Amount of information**	**If $\delta \sim \frac{L}{\sqrt{b}}$**
`Extra Gradient`	[4, 2]		$O\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	$O\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$
`SMMDS`	[3]	similarity	$O\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$O\left(\frac{1}{\sqrt{b}}\cdot\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$
`MASHA`	[1]	compression	$O\left(\frac{L}{\sqrt{M}\mu}\log\frac{1}{\varepsilon}\right)$	$O\left(\frac{1}{\sqrt{M}}\cdot\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$
`Optimistic MASHA`	This work	compression similarity	$O\left(\left[\frac{L}{M\mu}+\frac{\delta}{\sqrt{M}\mu}\right]\log\frac{1}{\varepsilon}\right)$	$O\left(\left[\frac{1}{M}+\frac{1}{\sqrt{Mb}}\right]\cdot\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$

Experiments: Toy for Theory Verification

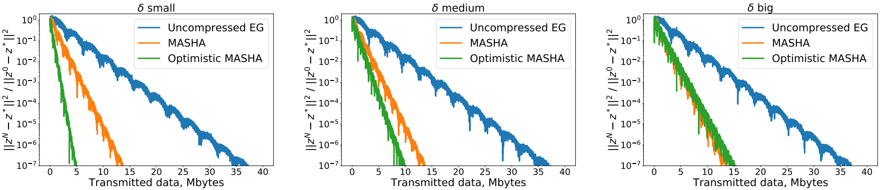
- Bilinear saddle point problem:

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} g(x,y) \stackrel{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} g_m(x,y) \quad \text{with}$$

$$g_m(x,y) \stackrel{\text{def}}{=} x^\top A_m y + a_m^\top x + b_m^\top y + \frac{\lambda}{2}\|x\|^2 - \frac{\lambda}{2}\|y\|^2,$$

  where $A_m \in \mathbb{R}^{d\times d}$, $a_m, b_m \in \mathbb{R}^d$. This problem is λ-strongly convex–strongly concave and, moreover, L-smooth with $L = \|A\|_2$ for $A = \frac{1}{M}\sum_{m=1}^{M} A_m$. We take $M = 10$, $d = 100$ and generate matrix A (with $\|A\|_2 \approx 100$) and vectors a_m, b_m randomly. We also generate matrices B_m such that all elements of these matrices are independent and have an unbiased normal distribution with variance σ^2. Using these matrices, we compute $A_m = A + B_m$. It can be considered that $\delta \sim \sigma$. In particular, we run three experiment setups: with small $\sigma \approx \frac{\|A\|_2}{100}$, medium $\sigma \approx \frac{\|A\|_2}{10}$ and big $\sigma \approx \|A\|_2$. λ is chosen as $\frac{\|A\|_2}{10^5}$.
- We use the new algorithm – `Optimistic MASHA`, the existing compression algorithm `MASHA` [1], and the classic uncompressed `Extra Gradient` [4, 2] as competitors. In `Optimistic MASHA` and `MASHA` we use the Permutation compressors.

Experiments: Bilinear Saddle Point Problem



(a) small δ (b) medium δ (c) big δ

Figure: Bilinear problem: Comparison of state-of-the-art methods with compression for variational inequalities for small, medium and big similarity parameters.

Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov.
Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees.
arXiv preprint arXiv:2110.03313, 2021.

Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov.
Local sgd for saddle-point problems.
arXiv preprint arXiv:2010.13112, 2020.

Aleksandr Beznosikov, Gesualdo Scutari, Alexander Rogozin, and Alexander Gasnikov.
Distributed saddle-point problems under data similarity.
Advances in Neural Information Processing Systems, 34, 2021.

Anatoli Juditsky, Arkadii S. Nemirovskii, and Claire Tauvel.
Solving variational inequalities with stochastic mirror-prox algorithm, 2008.

Rafał Szlendak, Alexander Tyurin, and Peter Richtárik.
Permutation compressors for provably faster distributed nonconvex optimization.
arXiv preprint arXiv:2110.03300, 2021.