Bregman Proximal Method for Efficient Communications under Similarity

Aleksandr Beznosikov [1] Darina Dvinskikh [3] Dmitry Bylinkin [1] Andrei Semenov [1] Alexander Gasnikov [1,2,4]

[1] Moscow Institute of Physics and Technology [2] Skoltech [3] HSE University [4] Ivannikov Institute for System Programming RAS

Distributed VIs

We study the regularized variational inequality (VI) problem formulated as finding $z^* \in \mathbb{Z}$ such that

$$\langle F(z^*), z - z^* \rangle + g(z) - g(z^*) \geq 0,$$

$\forall z \in \mathbb{Z}$, where $\mathbb{Z} \subseteq \mathbb{R}^d$ is a closed convex set, and $g : \mathbb{Z} \to \mathbb{R}$ is a proper convex lower semicontinuous function.

Modern applications often require working with an operator of the form

$$F(z) = \frac{1}{m}\sum_{i=1}^{m} F_i(z),$$

where $\{F_i\}_{i=1}^m$ are distributed across m nodes. One of the approaches to overcome the communication bottleneck is to exploit the similarity of local data.

Similarity

The essence of similarity approaches is to move most of the computation to the server, offloading the other nodes. If local datasets are i.i.d. samples from the same distribution, local operators F_i are statistically similar to their average F. In the case of convex optimization problems, this condition has the form

$$\|\nabla^2 f(z) - \nabla^2 f_i(z)\| \leq \delta.$$

In the case of VIs, the Hessian similarity can be generalized and written as

$$\|(F_i - F)(z_1) - (F_i - F)(z_2)\| \leq \delta\|z_1 - z_2\|.$$

This is the most natural measure of similarity because generally $\delta \sim 1/\sqrt{N}$.

Definitions

- The operator $F(\cdot)$ is called μ*−strongly monotone* with respect to distance generating function $w(\cdot)$, if

$$\langle F(u) - F(v), u - v \rangle \geq \mu \left(V(u, v) + V(v, u) \right),$$

for all $u, v \in \mathbb{Z}$, where $V(\cdot,\cdot)$ is the Bregman divergence corresponding to $w(\cdot)$.

- The operator $F(\cdot)$ is called L-Lipschitz, if

$$\|F(u) - F(v)\| \leq L\|u - v\|,$$

for all $u, v \in \mathbb{Z}$.

- We call the stochastic operator $F(\cdot,\xi)$ to be unbiased with bounded variance, if

$$\mathbb{E}_\xi[F(z,\xi)] = F(z),$$
$$\mathbb{E}_\xi[\|F(z^*,\xi) - F(z^*)\|^2] \leq \sigma_z^2,$$

for every $z \in \mathbb{Z}$.

Main Algorithm

Algorithm PAUS

1: **for** $k = 0, 1, 2, \ldots, K-1$ **do**
2: Sample random variable ξ^k on server
3: Collect $F(z^k, \xi^k) = \frac{1}{m}\sum_{i=1}^m F_i(z^k, \xi_i^k)$ on server
4: Find u^k as a solution to

$$\gamma\langle F_1(u^k) + F(z^k,\xi^k) - F_1(z^k), z - u^k\rangle + \langle \nabla w(u^k) - \nabla w(z^k), z - u^k\rangle + \gamma(g(z) - g(u^k)) \geq 0$$

for all $z \in \mathbb{Z}$ by SCMP procedure on server
5: Collect $F(u^k, \xi^k) = \frac{1}{m}\sum_{i=1}^m F_i(u^k, \xi_i^k)$ on server
6: Find z^{k+1} as a solution to

$$\langle \gamma(F(u^k,\xi^k) - F_1(u^k) - F(z^k,\xi^k) + F_1(z^k)) + (1+\alpha)(\nabla w(z^{k+1}) - \nabla w(u^k)), z - z^{k+1}\rangle \geq 0$$

for all $z \in \mathbb{Z}$ on server
7: **end for**
8: **return** $\tilde{u}^K = \frac{1}{K}\sum_{k=0}^{K-1} u^k$ for monotone VIs and z^K for strongly monotone ones

Convergence

Theorem

Consider the monotone operator $F(\cdot)$. Let the stochastic oracle $F(\cdot,\xi)$ be monotone, unbiased and have uniformly bounded variance. Suppose $F(\cdot,\xi) - F_1(\cdot)$ is δ-smooth. Let $\tilde{u}^K$ be the output of PAUS, run with appropriate parameters and starting points $z^0, u^0 \in \mathbb{Z}$ in

$$\mathcal{O}\left(\frac{D\delta}{\varepsilon} + \frac{D\sigma^2}{\varepsilon^2}\right)$$

communication rounds. Then it achieves $\mathrm{Gap}(\tilde{u}^K) \leq \varepsilon$.

Theorem

Consider the strongly monotone operator $F(\cdot)$. Let the stochastic oracle $F(\cdot,\xi)$ be strongly monotone, unbiased and have variance bounded at the solution. Suppose $F(\cdot,\xi) - F_1(\cdot)$ is δ-smooth. Let z^K be the output of PAUS, run with an appropriate parameters and a starting point $z^0 \in \mathbb{Z}$, in

$$\mathcal{O}\left(\frac{8\delta}{\mu}\log\frac{1}{\varepsilon} + \frac{8\sigma_*^2}{3\mu\varepsilon}\right)$$

communication rounds. Then it achieves $V(z^*, z^K) \leq \varepsilon$.

Approach to the Subproblem

For simplicity we introduce the function

$$H(v,\xi) = \gamma\left(F_1(v,\xi) + F(z^k,\xi^k) - F_1(z^k)\right).$$

Algorithm SCMP

1: Choose starting point $v^0 \in \mathbb{Z}$
2: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
3: Sample random variable ξ^t on server
4: Find $v^{t+\frac{1}{2}}$ as a solution to

$$\langle \eta H(v^t,\xi^t) + \eta(\nabla w(v^{t+\frac{1}{2}}) - \nabla w(z^k)) + \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^t), v - v^{t+\frac{1}{2}}\rangle + \gamma(g(v) - g(v^{t+\frac{1}{2}})) \geq 0$$

5: Find v^{t+1} as a solution to

$$\langle \eta H(v^{t+\frac{1}{2}},\xi^t) + \eta(\nabla w(v^{t+1}) - \nabla w(z^k)) + \nabla w(v^{t+1}) - \nabla w(v^t), v - v^{t+1}\rangle + \gamma(g(v) - g(v^{t+1})) \geq 0.$$

6: **end for**
7: **return** v^T

Theorem

Consider the monotone operator $F_1(\cdot)$. Let the stochastic oracle $F_1(\cdot,\xi)$ be Lipschitz, monotone, unbiased and have variance bounded at the solution of the subproblem. Suppose $F(\cdot,\xi) - F_1(\cdot)$ is δ-smooth. Consider stepsize $\gamma = 1/2\delta$ and starting point v^0. Then SCMP with appropriate choice of η needs

$$\mathcal{O}\left(\frac{L_{F_1}}{\delta}\log\frac{V(v^*,v^0)}{\varepsilon} + \frac{\sigma_{1,*}^2}{\varepsilon}\right) \text{ iterations}$$

to achieve $V(v^*, v^T) \leq \varepsilon$.

Experiments

We carry out numerical experiments for a stochastic matrix game

$$\min_{x\in\Delta}\max_{y\in\Delta}\left[x^\top \mathbb{E}[A_\xi] y\right],$$

where x, y are the mixed strategies of two players, Δ is the probability simplex, and A_ξ is a stochastic payoff matrix.

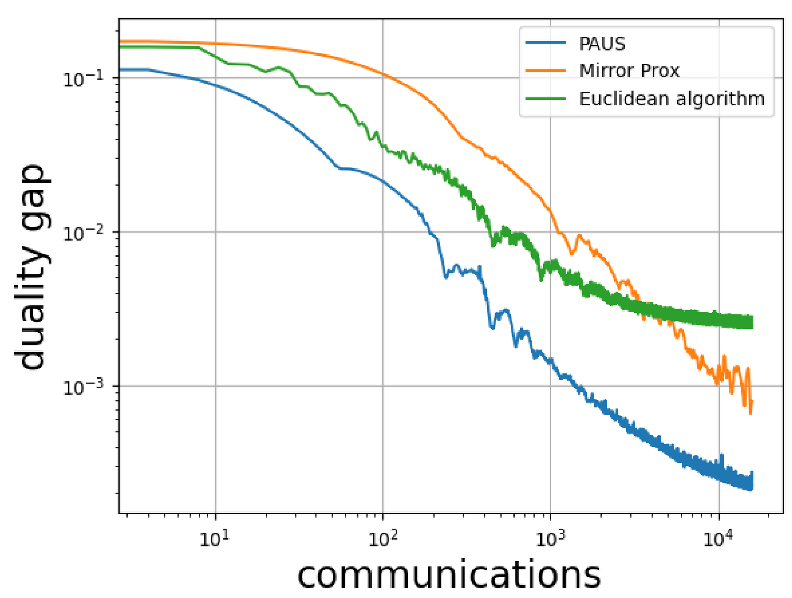

Figure: Comparison of state-of-the-art methods