Bregman Proximal Method for Efficient Communications under Similarity

Aleksandr Beznosikov
MIPT, ISP RAS, Innopolis University

Darina Dvinskikh
HSE University

Dmitry Bylinkin
MIPT, ISP RAS

Andrei Semenov
MIPT

Alexander Gasnikov
Innopolis University, MIPT, ISP RAS

Abstract

We propose a novel stochastic distributed method for both monotone and strongly monotone variational inequalities with Lipschitz operator and proper convex regularizers arising in various applications from game theory to adversarial training. By exploiting *similarity*, our algorithm overcomes the communication bottleneck that is a major issue in distributed optimization. The proposed method enjoys optimal communication complexity. All the existing distributed algorithms achieving the lower bounds under similarity condition essentially utilize the Euclidean setup. In contrast to them, our method is built upon the Bregman proximal maps and it is compatible with an arbitrary problem geometry. Thereby the proposed method fills an existing gap in this area of research. Our theoretical results are confirmed by numerical experiments on a stochastic matrix game.

1 Introduction

Variational inequalities (VIs) with monotone operators provide a unified and natural framework to cover many optimization problems, including convex minimization and convex-concave saddle point problems (SPPs) (Harker and Pang, 1990). The interest in VIs is due to their wide applicability in economics, equilibrium theory, game theory, optimal control and differential equations, see (Facchinei and Pang, 2003; Bauschke and Combettes, 2011) for an introduction. They also play an important role in modern machine learning, see (Goodfellow et al., 2014; Omidshafiei et al., 2017; Madry et al., 2017; Daskalakis et al., 2017). We are interested in the regularized VI problem formulated as follows.

$$\text{Find } z^* \in \mathcal{Z}: \quad \langle F(z^*), z - z^* \rangle + g(z) - g(z^*) \geq 0, \quad \forall z \in \mathcal{Z}, \tag{1}$$

where $F : \mathcal{Z} \to \mathbb{R}^d$ is a monotone and Lipschitz operator, $\mathcal{Z} \subseteq \mathbb{R}^d$ is a closed convex set, and $g : \mathcal{Z} \to \mathbb{R}$ is a proper convex lower semicontinuous function. Solving modern problems inevitably involves the use of huge training datasets. This forces engineers to utilize distributed computational systems. The data is distributed over m nodes/clients/machines/devices coordinated by the server. Formally, this means working with an operator of the form $F(z) := \frac{1}{m}\sum_{i=1}^m F_i(z)$. However, such coordination can significantly slow down the learning process, especially for systems with large computational resources (Bekkerman et al., 2011). Therefore, one of our main challenges in building the algorithm is to overcome the communication bottleneck.

1.1 Similarity

To reduce the communication frequency, many techniques have been developed. Among them is *statistical preconditioning*, which reduces communication complexity by exploiting *similarity* (Shamir et al., 2014; Hendrikx et al., 2020), i.e., the information that operators F_i are similar to each other and to their average F. To account for similarity we will rewrite the VI problem (1) as a problem with two terms.

$$\text{Find } z^* \in \mathcal{Z}: \quad \langle Q(z^*) + F_1(z^*), z - z^* \rangle + g(z) - g(z^*) \geq 0, \quad \forall z \in \mathcal{Z}, \tag{2}$$

where $Q(z) := F(z) - F_1(z)$ and F_1 is accessed by the server, which is the most computationally powerful node. The essence of similarity approaches is to move most of the computation to the server, offloading the other nodes. This allows to significantly reduce the amount of communication. In the case of convex minimization, this has been more than extensively studied.

Convex minimization. The problem (1) captures optimality conditions for constrained convex optimization with operators $F_i(z) = \nabla f_i(z)$. In machine learning applications, similarity is quite common: local functions f_i are the average losses $\ell(\cdot, \cdot)$ on the training dataset $D_i = \{(x_1, y_1)^{(i)}, \dots, (x_N, y_N)^{(i)}\}$ stored at machine i:

$$f_i(z) := \frac{1}{N} \sum_{j=1}^{N} \ell(z, (x_j, y_j)^{(i)}), \quad i = 1 \dots, m. \tag{3}$$

Let us reformulate the minimization of the average risk on the functions from (3) as a stochastic optimization problem:

$$\min_{z \in \mathcal{Z}} \left[f(z) := \frac{1}{m} \sum_{i=1}^{m} f_i(z) \right]. \tag{4}$$

where $f_i(z) := f(z, D_i)$. D_i is the local dataset, which can be considered as a set of random variables. If D_i's on different machines are i.i.d. samples from the same distribution, the local empirical losses f_i are statistically similar to their average f. There are various ways to measure similarity, but the most natural and theoretically justified approach is the Hessian similarity: for all $i = 1, \dots, m$ with high probability (Hendrikx et al., 2020)

$$\|\nabla^2 f(z) - \nabla^2 f_i(z)\| \leq \delta. \tag{5}$$

This condition is referred to as δ-similarity, or δ-relatedness. It is shown that δ can be estimated (Hendrikx et al., 2020). For this purpose, the objective f is assumed to be L-smooth. Then, if the loss $f(z)$ is quadratic in z, then $\delta \sim {}^{L}/_{\sqrt{N}}$ (up to a log factor), for a non-quadratic loss $\delta \sim {}^{\sqrt{d}L}/_{\sqrt{N}}$ (up to a log factor) under the condition that (5) holds uniformly over a compact domain with high probability (Zhang and Xiao, 2018). A similar analysis can be done for saddle point problems.

SPPs. The problem (1) with operators $F_i(z) := [\nabla_x f_i(x, y), \ -\nabla_y f_i(x, y)]^\top$ also captures convex-concave SPPs of the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \left[f(x, y) := \frac{1}{m} \sum_{i=1}^{m} f_i(x, y) \right]. \tag{6}$$

In this case, we measure similarity in terms of second derivatives: for all $i = 1, \dots, m$ with high probability

$$\begin{aligned} \|\nabla^2_{xx} f(x, y) - \nabla^2_{xx} f_i(x, y)\| &\leq \delta, \\ \|\nabla^2_{xy} f(x, y) - \nabla^2_{xy} f_i(x, y)\| &\leq \delta, \\ \|\nabla^2_{yy} f(x, y) - \nabla^2_{yy} f_i(x, y)\| &\leq \delta. \end{aligned} \tag{7}$$

A significant surge of interest in SPPs is due to modern applications in GANs training (Gidel et al., 2018), reinforcement learning (Jin and Sidford, 2020; Omidshafiei et al., 2017; Wai et al., 2018), distributed control (Necoara et al., 2011), and optimal transport (Jambulapati et al., 2019).

1.2 Related works

VIs have been studied for more than half a century, and yet they remain an active area of research. Algorithms solving VIs with Lipschitz and monotone operators date back to Korpelevich (1976), who proposed the Extra Gradient method (an analog of the gradient method). A generalization of Extra Gradient is Mirror Prox proposed by Nemirovski (2004). The algorithm replaces the Euclidean projection with a more complex proximal Bregman step to fit the problem geometry.

Distributed methods. A distributed version of Extra Gradient was proposed by Beznosikov et al. (2020). For the distributed version of Mirror Prox, see (Rogozin et al., 2021). The authors also provided the lower bounds for the communication complexity: $^{L}/_{\varepsilon}$ for the monotone VIs, and $^{L}/_{\mu}$ for the strongly monotone VIs (up to a logarithmic factor). Here ε measures the non-optimality gap function, L is the parameter of Lipschitz continuity, μ is the parameter of strong monotonicity.

Distributed methods exploiting similarity. A first method exploiting Hessian similarity to reduce the communication complexity DANE was proposed by Shamir et al. (2014). It was designed for convex minimization problems. Later $\Omega\left(\sqrt{^{\delta}/_{\mu}}\log {^{1}/_{\varepsilon}}\right)$ was established as the lower boundary of communication rounds under Hessian similarity (Arjevani and Shamir, 2015). For quadratic problems, an optimal method has been developed using a modification of the DANE approach (Yuan and Li, 2020). For an optimal (up to the log factor) method in the general case, see (Tian et al., 2022). For SPPs, optimal (up to a log factor) algorithms have been proposed by Beznosikov et al. (2021) together with the lower bounds of communication complexity: $^{\delta}/_{\varepsilon}$ for the convex-concave SPPs, and $^{\delta}/_{\mu} \cdot \log {^{1}/_{\varepsilon}}$ for the strongly convex-strongly concave SPPs. Kovalev et al. (2022) improved the results of Tian et al. (2022) and Beznosikov et al. (2021) by proposing optimal methods for convex minimization and convex-concave SPPs with optimal computational complexity, i.e. the authors eliminated non-optimal logarithm. In addition, there are a number of state-of-the-art methods that work with similarity, the complexity of which is improved by using communication compression and client sampling (Beznosikov et al., 2024; Beznosikov and Gasnikov, 2022; Lin et al., 2024).

1.3 Contribution

All the existing distributed algorithms exploiting similarity for convex minimization, SPPs or VIs are substantially utilizes the Euclidean setup. However, a large number of problems have non-Euclidean geometry, e.g., minimization on the probability simplex arising in machine learning (Nemirko and Dulá, 2021), statistics, chemistry, portfolio management (Chang et al., 2000), optimal transport and Wasserstein barycenters (Agueh and Carlier, 2011). Indeed, the Euclidean distance is not well suited for probability measures. Generalizing Euclidean algorithms to non-Euclidean ones is non-trivial and usually requires the development of new methods. Motivated by this gap between the Euclidean and non-Euclidean algorithms, we aim to design a novel method to tackle the problem geometry. To the best of our knowledge, we present the first distributed method utilizing similarity with the non-Euclidean setup. Technically our approach is based on the Bregman proximal maps, which work particularly well in constrained optimization. Our contribution can be summarized as follows:

- We present a **P**roximal **A**lgorithm **u**nder **S**imilarity (PAUS) that utilizes similarity for VIs with a monotone Lipschitz operator and a convex composite g. It achieves optimal communication complexity of $^{\delta}/_{\varepsilon}$, where ε measures the non-optimality gap function, and δ is the parameter of similarity.
- Our analysis also shows a speedup of PAUS in the case of an operator, that is μ-strongly monotone with respect to the Bregman divergence. Our method achieves optimal communication complexity of $^{\delta}/_{\mu} \cdot \log {^{1}/_{\varepsilon}}$ up to a logarithmic factor. The analysis is compatible with an arbitrary Bregman divergence.
- In both cases, we generalize our analysis by adding stochasticity to the method. PAUS achieves $^{\delta}/_{\varepsilon} + {^{\sigma_*^2}/_{\varepsilon^2}}$ for monotone VIs and $^{\delta}/_{\mu} \cdot \log {^{1}/_{\varepsilon}} + {^{\sigma_*^2}/_{\mu\varepsilon}}$ for strongly

monotone ones. In contrast to classical works that use the SGD-like approach to construct stochastic methods, we require boundedness of the variance of the stochastic oracle σ_*^2 only at the solution.

- We confirm our theoretical results by numerical experiments.

Paper organization. The structure of the paper is as follows. Section 2 introduces the preliminaries: the necessary definitions and assumptions. Section 3 presents the stochastic algorithm for solving VIs with monotone and Lipschitz operators and convex composites. Section 4 extends the convergence theory to the case of strongly monotone VIs. Section 5 presents the numerical experiments demonstrating the superiority of our novel algorithm on the two-player stochastic matrix game.

2 Preliminaries

Notation. For vectors, $\|z\|$ is a general norm on space $\mathcal{Z}$, and $\|s\|_*$ is its dual norm on the dual space $\mathcal{Z}^*$: $\|s\|_* = \max_{z\in\mathcal{Z}}\{\langle s, x\rangle : \|z\| = 1\}$. For matrices, $\|A\|$ is the matrix norm induced by vector norm $\|z\|$: $\|A\| = \sup_{z\in\mathcal{Z}}\{\|Az\| : \|z\| = 1\}$.

When discussing the communication efficiency, it is important to choose the right definition. One possible quality metric is the number of communication rounds. In addition to this, each vector exchange between client and server or the time spent on communication can be considered. This paper assumes the definition of communication complexity in the first mentioned sense. We measure how often the server communicates with the nodes without considering the number of vector exchanges per round.

Methods that utilize data similarity move the most computational complexity to the server. Therefore, in addition to communication complexity, it is worth measuring the complexity of local computations performed on the server.

Next, we provide all the definitions and assumptions necessary to build the convergence theory.

Definition 2.1 *We say that $g : \mathcal{Z} \to \mathbb{R}$ is a μ-strongly convex function with respect to $\|\cdot\|$, if*

$$g(u) - g(v) \geq \langle h, u - v\rangle + \frac{\mu}{2}\|u - v\|^2, \quad \forall u, v \in Z, \quad h \in \partial g(v).$$

If $\mu = 0$, g is called a convex function.

Definition 2.2 *We say that $w : \mathcal{Z} \to \mathbb{R}$ is a distance generating function (DGF), if w is 1-strongly convex with respect to $\|\cdot\|$, i.e., for all $u, v \in \mathcal{Z}$: $w(v) \geq w(u) + \langle \nabla w(u), v - u\rangle + \frac{1}{2}\|v - u\|^2$. The corresponding Bregman divergence is*

$$V(u, v) = w(u) - w(v) - \langle \nabla w(v), u - v\rangle, \quad \forall u, v \in \mathcal{Z}.$$

The property of distance generating function ensures $V(u, u) = 0$ and $V(u, v) \geq \frac{1}{2}\|u - v\|^2$.

Since we aim to use stochastic methods to reduce computation time, we introduce $F(\cdot, \xi)$ and $F_1(\cdot, \xi)$. These are the stochastic oracles of the corresponding operators. Consider $F(z, \xi)$. The random variable ξ can be understood in different ways. For example, as a sample of the computational node to be communicated with: $F(z, \xi) = F_\xi(z)$. Another interesting case covered by our analysis is the imposition of additive noise on the operator belonging to each machine: $F(z, \xi) = \frac{1}{m}\sum_{i=1}^m (F_i(z) + \xi_i) = F(z) + \xi$. Informally, such a technique allows to maintain privacy protection of local data. This refers to Federated Learning, a very popular trend nowadays (Li et al., 2020). We take into account that the objective operator $F(\cdot)$ has the form of a finite sum, where each $F_i(\cdot)$ can also be a finite sum. Thus, it is possible to speed up the server by using a stochastic approach: $F_1(z, \xi) = F_{1,\xi}(z)$.

Assumption 2.3 (Stochastic oracle for $F_1(\cdot)$) *The stochastic oracle $F_1(\cdot, \xi)$ is unbiased and its variance is bounded at the solution:*

$$\mathbb{E}_\xi[F_1(z, \xi)] = F_1(z), \quad \mathbb{E}_\xi[\|F_1(z^*, \xi) - F_1(z^*)\|_*^2] \leq \sigma_{1,*}^2, \quad \forall z \in \mathcal{Z},$$

Assumption 2.4 (Stochastic oracle for $F(\cdot)$) *The stochastic oracle $F(\cdot, \xi)$ is unbiased and there are two options of variance boundedness:*

$$\mathbb{E}_\xi[F(z, \xi)] = F(z) \quad \forall x \in \mathbb{R}.$$

(a) *The variance of $F(\cdot, \xi)$ is uniformly bounded:*

$$\mathbb{E}_\xi[\|F(z, \xi) - F(z)\|_*^2] \leq \sigma^2, \quad \forall z \in \mathcal{Z}.$$

(b) *The variance of $F(\cdot, \xi)$ is bounded at the solution:*

$$\mathbb{E}_\xi[\|F(z^*, \xi) - F(z^*)\|_*^2] \leq \sigma_*^2, \quad \forall z \in \mathcal{Z}.$$

Here the uniform boundedness is only needed to show convergence by the dual gap function in the monotone case. For strongly monotone VIs, only boundedness at the solution is enough.

Assumption 2.5 (Monotonicity) *The operators $F(\cdot, \xi)$, $F_1(\cdot, \xi)$ are monotone, i.e. for all $u, v \in \mathcal{Z}$ and for every ξ:*

$$\langle F(u, \xi) - F(v, \xi), u - v\rangle \geq 0,$$
$$\langle F_1(u, \xi) - F_1(v, \xi), u - v\rangle \geq 0.$$

A special case of monotone VIs are convex optimization problems and convex-concave SPPs.

Assumption 2.6 (Lipschitzness) *The operator $F_1(\cdot, \xi)$ is L_{F_1}-Lipschitz continuous, i.e. for all $u, v \in \mathcal{Z}$ and for every ξ:*

$$\|F_1(u, \xi) - F_1(v, \xi)\|_* \leq L_{F_1}\|u - v\|.$$

For convex minimization, this means that server function $f_1(z, \xi)$ is L_{F_1}-smooth.

Assumption 2.7 (δ-similarity) *The operator $F_1(\cdot) - F(\cdot, \xi)$ is δ-Lipschitz continuous, i.e., for all $u, v \in \mathcal{Z}$:*

$$\|F_1(u) - F(u, \xi) - F_1(v) + F(v, \xi)\|_* \leq \delta\|u - v\|.$$

In the Euclidean case, this is a stochastic generalization of δ-similarity for convex minimization and convex-concave SPPs. This is evident by considering (5) and (7) and noting that uniform bounding the norm of the Hessian by some constant δ entails δ-smoothness of the function. Note that all assumptions are introduced on the stochastic oracles of the operators, which is a stronger case than if we had introduced them on the operators themselves. This is necessary in order to use a wider class of stochastic oracles whose variance is bounded only at the optimum. If the variance of the chosen stochastic oracle is uniformly bounded, then we can relax Assumptions 2.5 and 2.6, but not Assumption 2.7.

3 Monotone case

In this section, we present the algorithm which solves (2).

3.1 Main algorithm

Now we are in a position to provide PAUS for VIs, see Algorithm 1.

At each iteration $k = 0, 1, \ldots, K - 1$ of Algorithm 1, the server initiates two rounds of communication to compute $F(z^k, \xi^k)$ at Line 3 and $F(u^k, \xi^k)$ at Line 5. To solve the inner problem encountered in Line 4 we provide a procedure we call Stochastic Composite MP (SCMP) (see Algorithm 2). Importantly, the same random variable ξ^k is used in the formulation of both subproblems. This is inspired by the work of Mishchenko et al. (2020), where this technique is used to not require uniform boundedness of the stochastic oracle.

We further propose a descent lemma for PAUS.

Algorithm 1 PAUS

```
Input: parameter of similarity $\delta$, stepsize $\gamma \le 1/2\delta$, parameter $\alpha \ge 0$, number of iterations $K$, starting points $z^0 = u^0 \in \mathcal{Z}$
for $k = 0, 1, 2, \dots, K-1$ do
    Sample random variable $\xi^k$ on server
    Collect $F(z^k, \xi^k) = \frac{1}{m}\sum_{i=1}^m F_i(z^k, \xi_i^k)$ on server
    Find $u^k$ as a solution to
        $\langle \gamma(F_1(u^k) + F(z^k, \xi^k) - F_1(z^k)) + \nabla w(u^k) - \nabla w(z^k), z - u^k \rangle + \gamma\left(g(z) - g(u^k)\right) \ge 0$
    for all $z \in \mathcal{Z}$ by SCMP (Algorithm 2) procedure on server
    Collect $F(u^k, \xi^k) = \frac{1}{m}\sum_{i=1}^m F_i(u^k, \xi_i^k)$ on server
    Solve
        $z^{k+1} = \arg\min_{z \in \mathcal{Z}} \left\{ \gamma \langle F(u^k, \xi^k) - F_1(u^k) - F(z^k, \xi^k) + F_1(z^k), z \rangle + (1+\alpha) V(z, u^k) \right\}$
    on server
end for
return $\tilde{u}^K = \frac{1}{K}\sum_{k=0}^{K-1} u^k$ for monotone VIs and $z^K$ for strongly monotone ones
```

Lemma 3.1 *Consider Assumption 2.7. Then the inequality*

$$
\begin{aligned}
2\gamma\left[\langle F(u^k, \xi^k), u^k - z\rangle + g(u^k) - g(z)\right] \le & 2V(z, z^k) - 2V(u^k, z^k) - 2(1+\alpha)V(z, z^{k+1}) \\
& - 2V(z^{k+1}, u^k) + 2\alpha V(z, u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 \\
& + \|z^{k+1} - u^k\|^2.
\end{aligned}
$$

holds.

See the proof in Appendix A.1. The next theorem presents the convergence rate of PAUS by the following gap function:

$$
\mathrm{Gap}(u) = \max_{z \in \mathcal{Z}} \left\{ \langle F(z), u - z\rangle + g(u) - g(z) \right\}. \tag{8}
$$

This function is the standard criterion for VIs. It corresponds to the standard optimality criteria in convex minimization and SPPs (Nemirovski, 2004; Juditsky et al., 2011).

Theorem 3.2 *Consider assumptions of Lemma 3.1 with Assumptions 2.4(a) and 2.5. Then after K communication rounds,* PAUS *(Algorithm 1), run with $\alpha = 0$, stepsize $\gamma \le 1/2\delta$ and starting points $z^0, u^0 \in \mathcal{Z}$, outputs $\tilde{u}^K$ such that*

$$
\mathbb{E}\left[\mathrm{Gap}(\tilde{u}^K)\right] \le \frac{D}{K\gamma} + \frac{2\gamma}{3}\sigma^2,
$$

where $D = \sup_{z \in Z}\left\{V(z, z^0)\right\}$.

See the proof in Appendix A.2. The theorem guarantees convergence of the method to some neighborhood of the solution. The following corollary shows that if γ is chosen correctly, convergence with arbitrary accuracy can be guaranteed.

Corollary 3.3 *Consider assumptions of Theorem 3.2. Let $\tilde{u}^K$ be the output of* PAUS *(Algorithm 1), run with appropriate parameters and starting points $z^0, u^0 \in \mathcal{Z}$ in*

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

then $\mathrm{Gap}(\tilde{u}^K) \le \varepsilon$.

See proof in Appendix A.3.

3.2 Stochastic Composite MP

Let us rewrite the problem encountered in Line 4 of Algorithm 1 as finding v^* such that for all $z \in \mathcal{Z}$:

$$\langle \gamma(F_1(v^*) + F(z^k, \xi^k) - F_1(z^k)) + \nabla w(v^*) - \nabla w(z^k), z - v^* \rangle + \gamma(g(z) - g(v^*)) \geq 0. \quad (9)$$

Note that in Line 4 we have to solve the problem with a strongly monotone operator, i.e., the Bregman divergence can also be used as a convergence criterion.

Algorithm 2 to solve auxiliary problem in step 4 of Algorithm 1

```
procedure SCMP($\gamma, L_{F_1}, z^k$)
    Choose stepsize $\eta = \frac{1}{3\gamma L_{F_1}}$
    Choose starting point $v^0 \in \mathcal{Z}$
    for $t = 0, 1, 2, \dots, T-1$ do
        Sample random variable $\xi^t$ on server
        Solve $v^{t+\frac{1}{2}} = \arg\min_{v \in \mathcal{Z}} \{\gamma\eta\langle H(v^t, \xi^t), v\rangle + \eta V(v, z^k) + V(v, v^t) + \gamma g(v)\}$ on server
        Solve $v^{t+1} = \arg\min_{v \in \mathcal{Z}} \{\gamma\eta\langle H(v^{t+\frac{1}{2}}, \xi^t), v\rangle + \eta V(v, z^k) + V(v, v^t) + \gamma g(v)\}$ on server
    end for
    return $v^T$
end procedure
```

The next theorem presents the convergence guarantee for SCMP to solve the problem (9). For the sake of brevity of description, we denote $H(v, \xi) = \gamma(F_1(v, \xi) + F(z^k, \xi^k) - F_1(z^k))$.

Theorem 3.4 *Let Assumptions 2.3, 2.5, 2.6 and 2.7 hold. Then* SCMP *(Algorithm 2) with* $\gamma = \frac{1}{2\delta}$ *produce the sequence* $\{v^t\}$ *such that*

$$\mathbb{E}\left[V(v^*, v^{t+1})\right] \leq \left(1 - \frac{\eta}{2}\right)\mathbb{E}\left[V(v^*, v^t)\right] + 4\eta^2\sigma_{1,*}^2.$$

The proof of this theorem is given in Appendix A.4. As in the case of Theorem 3.2, fine tuning of the algorithm parameters is required to obtain convergence with arbitrary accuracy.

Corollary 3.5 *Consider assumptions of Theorem 3.4. Let* v^T *be the output of* SCMP *procedure. Consider stepsize* $\gamma = 1/2\delta$ *and starting point* v^0*. Then Algorithm 2 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 $V(v^*, v^T) \leq \varepsilon$*.*

See proof in Appendix A.5. The procedure supports the possibility of a stochastic solution. This can be useful when the server node has too much data and needs to speed up the computation. We use the "same sample" technique from (Mishchenko et al., 2020). To solve the subproblem by the proposed method, we only need boundedness of the variance of the stochastic oracle $F_1(\cdot, \xi)$ at the solution.

4 Extension to strongly monotone VIs

In this section, we obtain a linear convergence rate of PAUS by the Bregman divergence, which is an appropriate convergence criterion for the strongly monotone operator.

Assumption 4.1 *Operator* $F(\cdot, \xi)$ *is* μ*-strongly monotone with respect to* DGF $w(\cdot)$*, i.e. for all* $u, v \in Z$ *and for every random variable* ξ*:*

$$\langle F(u, \xi) - F(v, \xi), u - v\rangle \geq \frac{\mu}{2}\left(V(u, v) + V(v, u)\right).$$

We assume strong monotonicity with respect to the corresponding geometry. This definition is not used for the first time, it is often found in the literature (Lu et al., 2018; Ablaev et al., 2022; Stonyakin et al., 2021).

The main difficulty in constructing a convergence theory for strongly monotone VIs is that the Bregman divergence is in general non-symmetric and there is no triangle inequality for it. To solve this problem, we "inflate" the coefficient in the subproblem (6) by some value α and use the strong monotonicity of the operator to eliminate the extra summands.

Theorem 4.2 *Consider assumptions of Lemma 3.1 with Assumption 2.4(b) and Assumption 4.1. Consider $\alpha = {}^{\gamma\mu}/_{2}$, $\gamma \leq {}^{1}/_{2\delta}$ and a starting point $z^0 \in \mathcal{Z}$. Then the inequality*

$$\mathbb{E}\left[V(z^*, z^{k+1})\right] \leq \left(1 - \frac{\gamma\mu}{4}\right)\mathbb{E}\left[V(z^*, z^k)\right] + \frac{2\gamma^2}{3}\sigma_*^2$$

holds.

See the proof in Appendix A.6. Let us repeat the proof of Corollary 3.5 with other constants and obtain

Corollary 4.3 *Consider assumptions of Theorem 4.2. Let z^K be the output of* PAUS *(Algorithm 1), run with an appropriate parameters and a starting point $z^0 \in \mathcal{Z}$, in*

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

then $V(z^*, z^K) \leq \varepsilon$.

Note that PAUS performs two rounds of communication per iteration. Thus, asymptotically the communication complexity by rounds is equal to the complexity by iterations.

In PAUS (Algorithm 1), the main computational load is shifted to the server. At one iteration of the method, each device accesses the local oracle twice, while the server is forced to do it $\mathcal{O}\left({}^{L_{F_1}}/_{\delta} \cdot \log {}^{V(v^*, v^0)}/_{\varepsilon} + {}^{\sigma_{1,*}^2}/_{\varepsilon}\right)$ times. Nevertheless, the additional freedom to choose the Bregman divergence can make it easier to compute the proximal mapping or even produce a closed-form solution. See examples in Appendix B.

5 Numerical experiments

We evaluate the effectiveness of PAUS by comparing it with other distributed algorithms with and without exploiting data similarity. As a first algorithm we consider Decentralized Mirror Prox from (Rogozin et al., 2021). This algorithm is based on the Bregman proximal maps but does not exploit similarity. The second algorithm is the algorithm from (Kovalev et al., 2022) which we will refer to as Extra Gradient. The algorithm is designed under similarity but in the Euclidean setup.

Problem. 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. A solution of this problem can be approximated by a solution of the following empirical problem

$$\min_{x\in\Delta}\max_{y\in\Delta}\left[\frac{1}{m}\sum_{i=1}^m x^\top A_i y\right],$$

where $A_1, \ldots, A_m$ are i.i.d samples of stochastic matrix A_ξ. We study the convergence of algorithms in terms of the duality gap

$$\max_{y\in\Delta}\left[\frac{1}{m}\sum_{i=1}^m \left(\widetilde{x}^K\right)^\top A_i y\right] - \min_{x\in\Delta}\left[\frac{1}{m}\sum_{i=1}^m x^\top A_i \widetilde{y}^K\right] \leq \varepsilon,$$

where $\widetilde{u}^K = (\widetilde{x}^K, \widetilde{y}^K)$ is the output of algorithms. This duality gap is the upper bound for the gap function for VIs from (8).

Data. We model entries of stochastic matrix A_ξ of size $d \times d$ with $d = 25$ as $[A_\xi]_{ij} = (1 + \nu\xi)[C]_{ij}$, where $\nu \in (0, 1)$, ξ is a Rademacher random variable and C is a deterministic matrix given in Exercise 5.3.1 of (Ben-Tal and Nemirovski, 2001). ν is used to adjust the similarity parameter δ. From this matrix we sample $m = 10^4$ matrices $A_1, \dots, A_m$. These matrices are shared between 5 devices, each holds local datasets of size $n = 2 \cdot 10^3$.

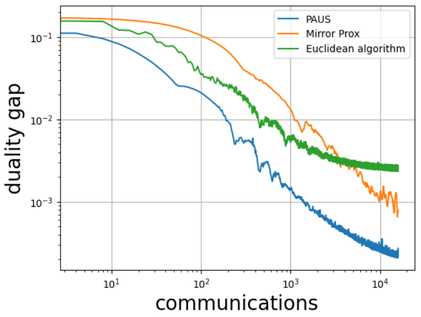

Figure 1: Comparison of state-of-the-art methods

Figure 1 demonstrates the superiority of PAUS in comparison with other distributed algorithms on the problem with $L \approx 10^{-1}$ and $\delta \approx 10^{-2}$, $\nu \approx 10^{-3}$. The parameters of algorithms were estimated theoretically (see Appendix C) and then tuned to get faster convergence. All algorithms have approximate slope ratio -1 according to their theoretical bounds ($K \sim \varepsilon^{-1}$). A faster convergence of our algorithm in comparison with the Euclidean algorithm (Kovalev et al., 2022) is achieved due to better utilizing the constrained set, namely the probability simplex, for which the Euclidean distance is worse than the ℓ_1-distance. Thus our method is able to better estimate the parameter δ in the proper norm and the distance to a solution (in terms of the Bregman divergence). A slow convergence of Mirror Prox (Rogozin et al., 2021) is explained by ignoring data similarity.

6 Conclusion

In this paper, we introduced the novel distributed stochastic proximal algorithm using similarity for both monotone and strongly monotone variational inequalities: PAUS. It achieves optimal communication complexity: ${}^{\delta}\!/_{\varepsilon}$ in the general monotone case and ${}^{\delta}\!/_{\mu} \cdot \log {}^{1}\!/_{\varepsilon}$ in the strongly monotone case. In contrast to existing communication-efficient algorithms that exploit similarity, PAUS is able to tackle non-Euclidean problems because it uses the Bregman setup.

References

Ablaev, S. S., Titov, A. A., Stonyakin, F. S., Alkousa, M. S., and Gasnikov, A. (2022). Some adaptive first-order methods for variational inequalities with relatively strongly monotone operators and generalized smoothness. In *International Conference on Optimization and Applications*, pages 135–150. Springer.

Agueh, M. and Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Arjevani, Y. and Shamir, O. (2015). Communication complexity of distributed convex learning and optimization. *Advances in neural information processing systems*, 28.

Bauschke, H. and Combettes, P. (2011). Convex analysis and monotone operator theory in hilbert spaces. *CMS books in mathematics). DOI*, 10:978–1.

Bekkerman, R., Bilenko, M., and Langford, J. (2011). *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press.

Ben-Tal, A. and Nemirovski, A. (2001). *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM.

Beznosikov, A. and Gasnikov, A. (2022). Compression and data similarity: Combination of two techniques for communication-efficient solving of distributed variational inequalities. In *International Conference on Optimization and Applications*, pages 151–162. Springer.

Beznosikov, A., Samokhin, V., and Gasnikov, A. (2020). Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms. *arXiv preprint arXiv:2010.13112*.

Beznosikov, A., Scutari, G., Rogozin, A., and Gasnikov, A. (2021). Distributed saddle-point problems under data similarity. *Advances in Neural Information Processing Systems*, 34:8172–8184.

Beznosikov, A., Takác, M., and Gasnikov, A. (2024). Similarity, compression and local steps: three pillars of efficient communications for distributed variational inequalities. *Advances in Neural Information Processing Systems*, 36.

Chang, T.-J., Meade, N., Beasley, J. E., and Sharaiha, Y. M. (2000). Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research*, 27(13):1271–1302.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2017). Training gans with optimism. *arXiv preprint arXiv:1711.00141*.

Facchinei, F. and Pang, J.-S. (2003). *Finite-dimensional variational inequalities and complementarity problems*. Springer.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2018). A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Harker, P. T. and Pang, J.-S. (1990). Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1):161–220.

Hendrikx, H., Xiao, L., Bubeck, S., Bach, F., and Massoulie, L. (2020). Statistically preconditioned accelerated gradient method for distributed optimization. In *International conference on machine learning*, pages 4203–4227. PMLR.

Jambulapati, A., Sidford, A., and Tian, K. (2019). A direct tilde {O}(1/epsilon) iteration parallel algorithm for optimal transport. *Advances in Neural Information Processing Systems*, 32.

Jin, Y. and Sidford, A. (2020). Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pages 4890–4900. PMLR.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58.

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

Kovalev, D., Beznosikov, A., Sadiev, A., Persiianov, M., Richtárik, P., and Gasnikov, A. (2022). Optimal algorithms for decentralized stochastic variational inequalities. *Advances in Neural Information Processing Systems*, 35:31073–31088.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60.

Lin, D., Han, Y., Ye, H., and Zhang, Z. (2024). Stochastic distributed optimization under average second-order similarity: Algorithms and analysis. *Advances in Neural Information Processing Systems*, 36.

Lu, H., Freund, R. M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. (2020). Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR.

Necoara, I., Nedelcu, V., and Dumitrache, I. (2011). Parallel and distributed optimization methods for estimation and control in networks. *Journal of Process Control*, 21(5):756–766.

Nemirko, A. and Dulá, J. (2021). Machine learning algorithm based on convex hull analysis. *Procedia Computer Science*, 186:381–386.

Nemirovski, A. (2004). Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., and Vian, J. (2017). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR.

Rogozin, A., Beznosikov, A., Dvinskikh, D., Kovalev, D., Dvurechensky, P., and Gasnikov, A. (2021). Decentralized distributed optimization for saddle point problems. *arXiv preprint arXiv:2102.07758*.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR.

Stich, S. U. (2019). Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*.

Stonyakin, F., Titov, A., Makarenko, D., and Alkousa, M. (2021). Some methods for relatively strongly monotone variational inequalities. *arXiv preprint arXiv:2109.03314*.

Tian, Y., Scutari, G., Cao, T., and Gasnikov, A. (2022). Acceleration in distributed optimization under similarity. In *International Conference on Artificial Intelligence and Statistics*, pages 5721–5756. PMLR.

Wai, H.-T., Yang, Z., Wang, Z., and Hong, M. (2018). Multi-agent reinforcement learning via double averaging primal-dual optimization. *Advances in Neural Information Processing Systems*, 31.

Yuan, X.-T. and Li, P. (2020). On convergence of distributed approximate newton methods: Globalization, sharper bounds and beyond. *The Journal of Machine Learning Research*, 21(1):8502–8552.

Zhang, Y. and Xiao, L. (2018). Communication-efficient distributed optimization of self-concordant empirical loss. *Large-Scale and Distributed Optimization*, pages 289–341.

A Appendix with missing proofs

A.1 Proof of Lemma 3.1

Lemma A.1 ***(Lemma 3.1)*** *Consider Assumption 2.7. Then the inequality*

$$\begin{aligned}2\gamma\left[\langle F(u^k,\xi^k), u^k - z\rangle + g(u^k) - g(z)\right] \leq & 2V(z,z^k) - 2V(u^k,z^k) - 2(1+\alpha)V(z,z^{k+1}) \\ & - 2V(z^{k+1},u^k) + 2\alpha V(z,u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 \\ & + \|z^{k+1} - u^k\|^2\end{aligned}$$

holds.

Proof: **Step 1.** We employ the following identity for the Bregman divergence:

$$V(z,z^k) - V(z,u^k) - V(u^k,z^k) = \langle \nabla w(u^k) - \nabla w(z^k), z - u^k)\rangle.$$

Using this identity for Line 4 of Algorithm 1, we get

$$\begin{aligned}\gamma(\langle F_1(u^k) + F(z^k,\xi^k) - F_1(z^k), z - u^k\rangle + g(z) - g(u^k)) & \\ + V(z,z^k) - V(z,u^k) - V(u^k,z^k) & \geq 0.\end{aligned}$$

We rewrite this and obtain

$$\gamma\langle F_1(u^k) + F(z^k,\xi^k) - F_1(z^k), u^k - z\rangle + \gamma(g(u^k) - g(z)) \leq V(z,z^k) - V(z,u^k) - V(u^k,z^k). \tag{10}$$

Step 2. Writing the optimality condition for Line 6 of Algorithm 1, we get for all $z \in \mathcal{Z}$

$$\langle \gamma(F(u^k,\xi^k) - F_1(u^k) - F(z^k,\xi^k) + F_1(z^k)) + (1+\alpha)(\nabla w(z^{k+1}) - \nabla w(u^k)), z - z^{k+1}\rangle \geq 0. \tag{11}$$

Then we utilize the following identity for the Bregman divergence

$$V(z,u^k) - V(z,z^{k+1}) - V(z^{k+1},u^k) = \langle \nabla w(z^{k+1}) - \nabla w(u^k), z - z^{k+1})\rangle.$$

Using this identity for (11) and denoting $H(u^k,z^k,\xi^k) = F(u^k,\xi^k) - F_1(u^k) - F(z^k,\xi^k) + F_1(z^k)$, we obtain

$$\gamma\langle H(u^k,z^k,\xi^k), z^{k+1} - z\rangle \leq (1+\alpha)V(z,u^k) - (1+\alpha)V(z,z^{k+1}) - V(z^{k+1},u^k).$$

Let us add and then subtract u^k in $\langle H(u^k,z^k,\xi^k), z^{k+1} - z\rangle$:

$$\begin{aligned}\gamma\langle H(u^k,z^k,\xi^k), z^{k+1} - u^k\rangle + \gamma\langle H(u^k,z^k,\xi^k), u^k - z\rangle \leq & (1+\alpha)V(z,u^k) \\ & - (1+\alpha)V(z,z^{k+1}) \\ & - V(z^{k+1},u^k).\end{aligned} \tag{12}$$

Step 3. Now we summarize (10) and (12) and get

$$\begin{aligned}\gamma\langle F(u^k,\xi^k), u^k - z\rangle + \gamma(g(u^k) - g(z)) \leq & V(z,z^k) - V(u^k,z^k) - (1+\alpha)V(z,z^{k+1}) \\ & - V(z^{k+1},u^k) + \alpha V(z,u^k) \\ & + \gamma\langle H(u^k,z^k,\xi^k), u^k - z^{k+1}\rangle.\end{aligned}$$

Next for this we use the Cauchy–Schwarz inequality

$$\begin{aligned}2\gamma\left[\langle F(u^k,\xi^k), u^k - z\rangle + g(u^k) - g(z)\right] \leq & 2V(z,z^k) - 2V(u^k,z^k) - 2(1+\alpha)V(z,z^{k+1}) \\ & - 2V(z^{k+1},u^k) + 2\alpha V(z,u^k) \\ & + \gamma^2\|H(u^k,z^k,\xi^k)\|_*^2 \\ & + \|z^{k+1} - u^k\|^2.\end{aligned} \tag{13}$$

Assumption 2.7 gives

$$\|H(u^k,z^k,\xi^k)\|_* = \|F(u^k,\xi^k) - F_1(u^k) - F(z^k,\xi^k) + F_1(z^k)\|_* \leq \delta\|u^k - z^k\|.$$

We use this for (13) and obtain

$$\begin{aligned}2\gamma\left[\langle F(u^k,\xi^k), u^k - z\rangle + g(u^k) - g(z)\right] \leq & 2V(z,z^k) - 2V(u^k,z^k) - 2(1+\alpha)V(z,z^{k+1}) \\ & - 2V(z^{k+1},u^k) + 2\alpha V(z,u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 \\ & + \|z^{k+1} - u^k\|^2.\end{aligned}$$

This transition completes the proof of the lemma. □

A.2 Proof of Theorem 3.2

Theorem A.2 ***(Theorem 3.2)*** *Consider assumptions of Lemma 3.1 with Assumptions 2.4(a) and 2.5. Then after K communication rounds,* PAUS *(Algorithm 1), run with $\alpha = 0$, stepsize $\gamma \leq 1/2\delta$ and a starting point $z^0 \in \mathcal{Z}$, outputs $\tilde{u}^K$ such that*

$$\mathbb{E}\left[\text{Gap}(\tilde{u}^K)\right] \leq \frac{D}{K\gamma} + \frac{2\gamma}{3}\sigma^2,$$

where $D = \sup_{z \in Z}\left\{V(z, z^0)\right\}$.

Proof: From Lemma 3.1 we have

$$\begin{aligned}2\gamma\left[\langle F(u^k,\xi^k), u^k - z\rangle + g(u^k) - g(z)\right] \leq & 2V(z,z^k) - 2V(u^k,z^k) - 2(1+\alpha)V(z,z^{k+1}) \\ & - 2V(z^{k+1},u^k) + 2\alpha V(z,u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 \\ & + \|z^{k+1} - u^k\|^2.\end{aligned}$$

Next we take $\alpha = 0$ and utilize the monotonicity of $F(\cdot,\xi)$ (Assumption 2.5):

$$\begin{aligned}2\gamma\left[\langle F(z,\xi^k), u^k - z\rangle + g(u^k) - g(z)\right] \leq & 2V(z,z^k) - 2V(u^k,z^k) - 2(1+\alpha)V(z,z^{k+1}) \\ & - 2V(z^{k+1},u^k) + 2\alpha V(z,u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 \\ & + \|z^{k+1} - u^k\|^2.\end{aligned}$$

We add and subtract $F(z)$ in the scalar product and obtain

$$\begin{aligned}2\gamma\left[\langle F(z), u^k - z\rangle + g(u^k) - g(z)\right] \leq & 2V(z,z^k) - 2V(u^k,z^k) - 2V(z,z^{k+1}) - 2V(z^{k+1},u^k) \\ & + \gamma^2\delta^2\|u^k - z^k\|^2 + \|z^{k+1} - u^k\|^2. \\ & + 2\gamma\langle F(z) - F(z,\xi^k), u^k - z\rangle.\end{aligned}$$

Note that:

$$\mathbb{E}_\xi\langle F(z) - F(z,\xi^k), u^k - z\rangle = \mathbb{E}_\xi\langle F(z) - F(z,\xi^k), u^k - z^k\rangle,$$

because $F(\cdot,\xi)$ is unbiased and z, z^k are independent on ξ^k. Thus, we calculate the expectation and get

$$\begin{aligned}2\gamma\mathbb{E}\langle F(z), u^k - z\rangle + g(u^k) - g(z) \leq & \mathbb{E}2V(z,z^k) - 2V(u^k,z^k) - 2V(z,z^{k+1}) \\ & - 2V(z^{k+1},u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 + \|z^{k+1} - u^k\|^2 \\ & + \frac{4\gamma^2}{3}\sigma_z^2 + \frac{3}{4}\|u^k - z^k\|^2.\end{aligned}$$

Here we also apply the Cauchy-Schwartz inequality to $2\gamma\langle F(z) - F(z,\xi^k), u^k - z^k\rangle$. For the Bregman divergence the inequality $V(x,y) \geq \frac{1}{2}\|x - y\|^2$ is satisfied for all x, y. Thus,

$$-2V(u^k,z^k) \leq -\|u^k - z^k\|^2,$$

$$-2V(z^{k+1},u^k) \leq -\|z^{k+1} - u^k\|^2.$$

Using this inequalities, we write

$$\begin{aligned}2\gamma\mathbb{E}\langle F(z), u^k - z\rangle + g(u^k) - g(z) \leq & \mathbb{E}2V(z,z^k) - 2V(z,z^{k+1}) + \left(\gamma^2\delta^2 + \frac{3}{4} - 1\right)\|u^k - z^k\|^2 \\ & + (1-1)\|z^{k+1} - u^k\|^2 + \frac{4\gamma^2}{3}\sigma^2.\end{aligned}$$

If we choose stepsize $\gamma \leq \frac{1}{2\delta}$, we obtain

$$2\gamma\mathbb{E}\langle F(z), u^k - z\rangle + g(u^k) - g(z) \leq \mathbb{E}2V(z,z^k) - 2V(z,z^{k+1}) + \frac{4\gamma^2}{3}\sigma^2.$$

We summarize this for $k = 0, 1, 2, \dots, K-1$

$$2\gamma \mathbb{E} \frac{1}{K} \sum_{k=0}^{K-1} \left[\langle F(z), u^k - z \rangle + g(u^k) - g(z) \right] \le \frac{2}{K} V(z, z^0) + \frac{4\gamma^2}{3} \sigma^2.$$

The statement of the theorem follows by taking the maximum:

$$\mathbb{E}[\mathrm{Gap}(\tilde{u}^k)] \le \frac{1}{K\gamma} \sup_{z \in Z} \left\{ V(z, z^0) \right\} + \frac{2\gamma}{3} \sigma^2.$$

□

A.3 PROOF OF COROLLARY 3.3

Corollary A.3 ***(Corollary 3.3)*** *Consider assumptions of Theorem 3.2. Let $\tilde{u}^K$ be the output of* PAUS *(Algorithm 1), run with an appropriate parameters and starting points $z^0, u^0 \in \mathcal{Z}$ in*

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

then $\mathrm{Gap}(\tilde{u}^K) \le \varepsilon$.

Proof: From Theorem 3.2 we have

$$\mathbb{E}\left[\mathrm{Gap}(\tilde{u}^K) \right] \le \frac{D}{K\gamma} + \frac{2\gamma}{3} \sigma^2,$$

where $D = \sup_{z \in Z} \left\{ V(z, z^0) \right\}$. Let us find γ, equating the summands in the right-hand side:

$$\gamma = \frac{1}{\sigma} \sqrt{\frac{3D}{2K}}.$$

- If $\frac{1}{\sigma}\sqrt{\frac{3D}{2K}} \le \frac{1}{2\delta}$, choose $\gamma = \frac{1}{\sigma}\sqrt{\frac{3D}{2K}}$. In this case we obtain

$$\mathbb{E}\left[\mathrm{Gap}(\tilde{u}^K) \right] \le 2\sigma \sqrt{\frac{2D}{3K}}.$$

- If $\frac{1}{\sigma}\sqrt{\frac{3D}{2K}} \ge \frac{1}{2\delta}$, choose $\gamma = \frac{1}{2\delta}$. In this case we have

$$\mathbb{E}\left[\mathrm{Gap}(\tilde{u}^K) \right] \le \frac{2\delta D}{K} + \frac{2\sigma^2}{3} \frac{2}{2\delta} \le \frac{2\delta D}{K} + \frac{2\sigma^2}{3} \frac{1}{\sigma} \sqrt{\frac{3D}{2K}}.$$

Getting rid of unnecessary constants, we get the communication complexity

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

□

A.4 PROOF OF THEOREM 3.4

Theorem A.4 ***(Theorem 3.4)*** *Let Assumptions 2.3, 2.5, 2.6 and 2.7 hold. Then* SCMP *(Algorithm 2) with $\gamma = \frac{1}{2\delta}$ produce the sequence $\{v^t\}$ such that*

$$\mathbb{E}\left[V(v^*, v^{t+1}) \right] \le \left(1 - \frac{\eta}{2} \right) \mathbb{E}\left[V(v^*, v^t) \right] + 4\eta^2 \sigma_{1,*}^2.$$

Proof: Let us introduce

$$H(x,\xi) := \gamma(F_1(x,\xi) + F(z^k,\xi^k) - F_1(z^k)).$$

Then we can rewrite the iterates of SCMP procedure:

$$v^{t+\frac{1}{2}} \leftarrow \arg\min_{v\in\mathcal{Z}}\{\eta\langle H(v^t,\xi^t),v\rangle + \eta V(v,z^k) + V(v,v^t) + \gamma g(v)\}. \tag{14}$$

$$v^{t+1} \leftarrow \arg\min_{v\in\mathcal{Z}}\{\eta\langle H(v^{t+\frac{1}{2}}.\xi^t),v\rangle + \eta V(v,z^k) + V(v,v^t) + \gamma g(v)\}. \tag{15}$$

From the optimality conditions for (14) and (15) we have

$$\begin{aligned}\langle \eta H(v^t,\xi^t) + &\eta(\nabla w(v^{t+\frac{1}{2}}) - \nabla w(z^k)) + \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^t), v^{t+\frac{1}{2}} - v\rangle \\ &\leq \gamma(g(v) - g(v^{t+\frac{1}{2}})).\end{aligned} \tag{16}$$

$$\begin{aligned}\langle \eta H(v^{t+\frac{1}{2}},\xi^t) + &\eta(\nabla w(v^{t+1}) - \nabla w(z^k)) + \nabla w(v^{t+1}) - \nabla w(v^t), v^{t+1} - v\rangle \\ &\leq \gamma(g(v) - g(v^{t+1})).\end{aligned} \tag{17}$$

Let v^* be an exact solution of problems in Line 4 of PAUS for which we employ the SCMP procedure. Plugging $v = v^{t+1}$ in (16) and $v = v^*$ in (17) and the summarizing these, we get

$$\begin{aligned}\langle \eta H(v^t,\xi^t) + &\eta(\nabla w(v^{t+\frac{1}{2}}) - \nabla w(z^k)) + \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^t), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &+\langle \eta H(v^{t+\frac{1}{2}},\xi^t) + \eta(\nabla w(v^{t+1}) - \nabla w(z^k)) + \nabla w(v^{t+1}) - \nabla w(v^t), v^{t+1} - v^*\rangle \\ &\leq \gamma(g(v^*) - g(v^{t+\frac{1}{2}})).\end{aligned}$$

Let us add and subtract $v^{t+\frac{1}{2}}$ into the second inner product and obtain

$$\begin{aligned}\langle \eta H(v^{t+\frac{1}{2}},\xi^t) + &\eta(\nabla w(v^{t+1}) - \nabla w(z^k)), v^{t+\frac{1}{2}} - v^*\rangle \\ &+\langle \eta(H(v^t,\xi^t) - H(v^{t+\frac{1}{2}},\xi^t)), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &+\eta\langle \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^{t+1}), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &+\langle \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^t), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &+\langle \nabla w(v^{t+1}) - \nabla w(v^t), v^{t+1} - v^*\rangle \\ &\leq \gamma(g(v^*) - g(v^{t+\frac{1}{2}})).\end{aligned}$$

After rearranging the terms we obtain

$$\begin{aligned}\langle \eta H(v^{t+\frac{1}{2}},\xi^t) + &\eta(\nabla w(v^{t+1}) - \nabla w(z^k)), v^{t+\frac{1}{2}} - v^*\rangle \\ &\leq \gamma(g(v^*) - g(v^{t+\frac{1}{2}})) \\ &\quad + \eta\langle H(v^{t+\frac{1}{2}},\xi^t) - H(v^t,\xi^t), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &\quad + \eta\langle \nabla w(v^{t+1}) - \nabla w(v^{t+\frac{1}{2}}), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &\quad + \langle \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^t), v^{t+1} - v^{t+\frac{1}{2}}\rangle \\ &\quad + \langle \nabla w(v^{t+1}) - \nabla w(v^t), v^* - v^{t+1}\rangle.\end{aligned}$$

Assumption 2.5 implies

$$\begin{aligned}\langle H(v^{t+\frac{1}{2}},\xi^t), v^{t+\frac{1}{2}} - v^*\rangle \geq &\langle H(v^*,\xi^t), v^{t+\frac{1}{2}} - v^*\rangle \\ = &\langle H(v^*), v^{t+\frac{1}{2}} - v^*\rangle + \langle F_1(v^*,\xi^t) - F_1(v^*), v^{t+\frac{1}{2}} - v^*\rangle.\end{aligned} \tag{18}$$

Moreover, we use the optimality condition for problem in Line 4 of PAUS, and we obtain the following for all $v \in \mathcal{Z}$

$$\langle H(v^*) + \nabla w(v^*) - \nabla w(z^k), v - v^*\rangle \geq \gamma(g(v^*) - g(v)).$$

Plugging $v = v^{t+\frac{1}{2}}$ in this, we have

$$-\langle H(v^*) + \nabla w(v^*), v^{t+\frac{1}{2}} - v^*\rangle + \langle \nabla w(z^k), v^{t+\frac{1}{2}} - v^*\rangle \leq \gamma(g(v^{t+\frac{1}{2}}) - g(v^*)). \tag{19}$$

Thus, summarizing (19) and the origin inequality, we get

$$\begin{aligned}\eta\langle F_1(v^*,\xi^t) - F_1(v^*), v^{t+\frac{1}{2}} - v^*\rangle + &\eta\langle \nabla w(v^{t+1}) - \nabla w(v^*), v^{t+\frac{1}{2}} - v^*\rangle \\ &\leq \eta\langle H(v^{t+\frac{1}{2}},\xi^t) - H(v^t,\xi^t), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &\quad + \eta\langle \nabla w(v^{t+1}) - \nabla w(v^{t+\frac{1}{2}}), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &\quad + \langle \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^t), v^{t+1} - v^{t+\frac{1}{2}}\rangle \\ &\quad + \langle \nabla w(v^{t+1}) - \nabla w(v^t), v^* - v^{t+1}\rangle. \end{aligned} \tag{20}$$

Note that $\mathbb{E}_{\xi^t}[\langle F_1(v^*,\xi^t) - F_1(v^*), v^{t+\frac{1}{2}} - v^*\rangle] \neq 0$, because $v^{t+\frac{1}{2}}$ depends on ξ^t. Assumption 2.3 allows to replace v^* by v^t in the first summand of the left-hand side while taking expectation:

$$\mathbb{E}_{\xi^t}\langle F_1(v^*,\xi^t) - F_1(v^*), v^{t+\frac{1}{2}} - v^*\rangle = \mathbb{E}_{\xi^t}\langle F_1(v^*,\xi^t) - F_1(v^*), (v^{t+\frac{1}{2}} - v^t) - (v^* - v^t)\rangle.$$

Both v^* and v^t are independent on ξ^t, Thus, we obtain

$$\mathbb{E}_{\xi^t}\langle F_1(v^*,\xi^t) - F_1(v^*), v^{t+\frac{1}{2}} - v^*\rangle = \mathbb{E}_{\xi^t}\langle F_1(v^*,\xi^t) - F_1(v^*), v^{t+\frac{1}{2}} - v^t\rangle.$$

Hence, using this, we rewrite (20) as follows

$$\begin{aligned}\eta\mathbb{E}\langle \nabla w(v^{t+1}) - \nabla w(v^*), v^{t+\frac{1}{2}} - v^*\rangle \leq &\mathbb{E}\eta\langle F_1(v^*,\xi^t) - F_1(v^*), v^t - v^{t+\frac{1}{2}}\rangle \\ &+ \eta\langle H(v^{t+\frac{1}{2}},\xi^t) - H(v^t,\xi^t), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &+ \eta\langle \nabla w(v^{t+1}) - \nabla w(v^{t+\frac{1}{2}}), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &+ \langle \nabla w(v^{t+\frac{1}{2}}) - \nabla w(v^t), v^{t+1} - v^{t+\frac{1}{2}}\rangle \\ &+ \langle \nabla w(v^{t+1}) - \nabla w(v^t), v^* - v^{t+1}\rangle. \end{aligned} \tag{21}$$

From the definition of the Bregman divergence, we have

$$-V(v^{t+1},v^{t+\frac{1}{2}}) - V(v^{t+\frac{1}{2}},v^{t+1}) = \langle \nabla w(v^{t+1}) - \nabla w(v^{t+\frac{1}{2}}), v^{t+\frac{1}{2}} - v^{t+1}\rangle. \tag{22}$$

$$V(v^*,v^{t+1}) + V(v^{t+1},v^t) - V(v^*,v^t) = \langle \nabla w(v^t) - \nabla w(v^{t+1}), v^* - v^{t+1}\rangle. \tag{23}$$

$$V(v^{t+1},v^{t+\frac{1}{2}}) + V(v^{t+\frac{1}{2}},v^t) - V(v^{t+1},v^t) = \langle \nabla w(v^t) - \nabla w(v^{t+\frac{1}{2}}), v^{t+1} - v^{t+\frac{1}{2}}\rangle. \tag{24}$$

$$V(v^{t+\frac{1}{2}},v^*) + V(v^*,v^{t+1}) - V(v^{t+\frac{1}{2}},v^{t+1}) = \langle \nabla w(v^{t+1}) - \nabla w(v^*), v^{t+\frac{1}{2}} - v^*\rangle. \tag{25}$$

Plugging (22), (23), (24) and (25) in (21), we obtain

$$\begin{aligned}\mathbb{E}\eta V(v^{t+\frac{1}{2}},v^*) + \eta V(v^*,v^{t+1}) - \eta V(v^{t+\frac{1}{2}},v^{t+1}) \leq &\mathbb{E}\eta\langle F_1(v^*,\xi^t) - F_1(v^*), v^t - v^{t+\frac{1}{2}}\rangle \\ &+ \eta\langle H(v^{t+\frac{1}{2}},\xi^t) - H(v^t,\xi^t), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &- \eta V(v^{t+1},v^{t+\frac{1}{2}}) - \eta V(v^{t+\frac{1}{2}},v^{t+1}) \\ &- V(v^{t+1},v^{t+\frac{1}{2}}) - V(v^{t+\frac{1}{2}},v^t) \\ &+ V(v^{t+1},v^t) - V(v^*,v^{t+1}) - V(v^{t+1},v^t) \\ &+ V(v^*,v^t). \end{aligned}$$

Rearranging the terms and using $V(v^{t+\frac{1}{2}},v^*) \geq 0$, we get

$$\begin{aligned}0 \leq &\mathbb{E}\eta\langle F_1(v^*,\xi^t) - F_1(v^*), v^t - v^{t+\frac{1}{2}}\rangle + \eta\langle H(v^{t+\frac{1}{2}},\xi^t) - H(v^t,\xi^t), v^{t+\frac{1}{2}} - v^{t+1}\rangle \\ &- (1+\eta)V(v^{t+1},v^{t+\frac{1}{2}}) - V(v^{t+\frac{1}{2}},v^t) - (1+\eta)V(v^*,v^{t+1}) + V(v^*,v^t). \end{aligned} \tag{26}$$

By using the Cauchy–Schwarz inequality with some constants C_1 and C_2, we obtain

$$\begin{aligned}(1+\eta)\mathbb{E}V(v^*,v^{t+1}) \leq &\mathbb{E}V(v^*,v^t) - (1+\eta)V(v^{t+1},v^{t+\frac{1}{2}}) - V(v^{t+\frac{1}{2}},v^t) \\ &+ C_1\sigma_*^2 + \frac{1}{C_2}\|v^{t+\frac{1}{2}} - v^{t+1}\|^2 + \left(\frac{\eta^2}{C_1} + C_2\eta^2\gamma^2 L_{F_1}^2\right)\|v^{t+\frac{1}{2}} - v^t\|^2. \end{aligned}$$

Next we use the fact that $V(x,y) \geq \frac{1}{2}\|x-y\|^2$ for all $x, y \in \mathbb{R}^d$. Let us choose

$$\gamma = \frac{1}{2\delta}, C_1 = 4\eta^2, C_2 = 2, \eta \leq \frac{1}{3\gamma L_{F_1}}. \tag{27}$$

$\frac{1}{1+\eta} \leq 1 - \frac{x}{2}$, since $\eta \leq \frac{1}{3\gamma L_{F_1}} = \frac{2\delta}{3L_{F_1}} < 1$. Thus, we have

$$\mathbb{E}[V(v^*, v^{t+1})] \leq \left(1 - \frac{\eta}{2}\right)\mathbb{E}[V(v^*, v^t)] + 4\eta^2\sigma_*^2.$$

A.5 Proof of Corollary 3.5

Corollary A.5 ***(Corollary 3.5)*** *Consider assumptions of Theorem 3.4. Let v^* be a solution of the subproblem in Line 4 of Algorithm 2 and let v^T be the output of* SCMP *procedure. Consider stepsize $\gamma = 1/2\delta$ and starting point v^0. Then Algorithm 2 with appropriate choice of η needs*

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

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

Proof: Denote $a = \frac{1}{2}$ and $c = 4\sigma_*^2$. Using (Stich, 2019), we obtain:

$$\mathbb{E}V(v^*, v^T) \leq \frac{2V(v^*, v^0)}{\eta}\exp\left\{-\frac{1}{2}\eta(T+1)\right\} + 8\eta\sigma_*^2.$$

- If $\frac{\delta}{3L_{F_1}} \geq {}^{\ln\left(\max\left\{2, \frac{V^0T^2}{16\sigma_*^2}\right\}\right)}/_{T}$, then choose $\eta = {}^{2\ln\left(\max\left\{2, \frac{V^0T^2}{16\sigma_*^2}\right\}\right)}/_{T}$ and obtain that the right side is $\mathcal{O}\left(\frac{8\sigma_*^2}{T}\right)$.
- Otherwise, choose $\eta = \frac{2\delta}{3L_{F_1}}$ and obtain $\mathcal{O}\left(\frac{3L_{F_1}V^0}{2\delta}\exp\{-\frac{\delta T}{3L_{F_1}}\} + \frac{8\sigma_*^2}{T}\right)$.

A.6 Proof of Theorem 4.2

Theorem A.6 ***(Theorem 4.2)*** *Consider assumptions of Lemma 3.1 with Assumption 2.4(b) and Assumption 4.1. Consider $\alpha = \gamma\mu/2$, $\gamma \leq 1/2\delta$ and a starting point $z^0 \in \mathcal{Z}$. Then the inequality*

$$\mathbb{E}\left[V(z^*, z^{k+1})\right] \leq \left(1 - \frac{\gamma\mu}{4}\right)\mathbb{E}\left[V(z^*, z^k)\right] + \frac{2\gamma^2}{3}\sigma_*^2$$

holds.

Proof: Let us start with Lemma 3.1:

$$\begin{aligned}2\gamma\left[\langle F(u^k, \xi^k), u^k - z\rangle + g(u^k) - g(z)\right] \leq& 2V(z, z^k) - 2V(u^k, z^k) - 2(1+\alpha)V(z, z^{k+1}) \\ &- 2V(z^{k+1}, u^k) + 2\alpha V(z, u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 \\ &+ \|z^{k+1} - u^k\|^2. \end{aligned} \tag{28}$$

Write down the optimality condition for problem (1):

$$\langle F(z^*), z - z^*\rangle \geq g(z^*) - g(z), \quad \forall z \in Z \tag{29}$$

Take $z = z^*$ in (28) and $z = u^k$ in (29). By summing these two expressions and then adding and subtracting $\langle F(z^*, \xi^k), u^k - z^*\rangle$, we obtain the following:

$$\begin{aligned}2\gamma\langle F(u^k, \xi^k) - F(z^*, \xi^k), u^k - z^*\rangle \leq& 2V(z^*, z^k) - 2V(u^k, z^k) - 2(1+\alpha)V(z^*, z^{k+1}) \\ &- 2V(z^{k+1}, u^k) + 2\alpha V(z^*, u^k) + \gamma^2\delta^2\|u^k - z^k\|^2 \\ &+ \|z^{k+1} - u^k\|^2 + 2\gamma\langle F(z^*) - F(z^*, \xi^k), u^k - z^*\rangle.\end{aligned}$$

Again we use the trick of replacing the point independent of ξ^k by an arbitrary point independent of ξ^k, under the expectation due to Assumption 2.4:

$$\mathbb{E}_{\xi^k}\langle F(z^*) - F(z^*,\xi^k), u^k - z^*\rangle = \mathbb{E}_{\xi^k}\langle F(z^*) - F(z^*,\xi^k), u^k - z^k\rangle.$$

Let us apply Young's inequality to $\mathbb{E}_{\xi^k}\langle F(z^*) - F(z^*,\xi^k), u^k - z^k\rangle$ and Assumption 4.1 to $\langle F(u^k,\xi^k) - F(z^*,\xi^k), u^k - z^*\rangle$:

$$\mathbb{E}\gamma\mu V(z^*,u^k) \le \mathbb{E}2V(z^*,z^k) + 2\alpha V(z^*,u^k) - 2(1+\alpha)V(z^*,z^{k+1}) + \frac{4\gamma^2}{3}\sigma_*^2.$$

Since $\alpha = {}^{\gamma\mu}/_{2}$, $V(z^*,u^k)$ is reduced. Note that ${}^{\gamma\mu}/_{2} \le {}^{\mu}/_{4\delta} < 1$. Thus, we obtain

$$\mathbb{E}\left[V(z^*,z^{k+1})\right] \le \left(1 - \frac{\gamma\mu}{4}\right)\mathbb{E}\left[V(z^*,z^k)\right] + \frac{2\gamma^2}{3}\sigma_*^2$$

□

B Closed forms for monotone VIs

For simplicity of presentation, we consider a non-stochastic version of PAUS.

Convex minimization. For convex minimization problem, operator $F(z) = \nabla f(z)$, $Q = \nabla f(z) - f_1(z) =: \nabla q(z)$. At each iteration of of PAUS server forms the gradient by averaging local gradients calculated by all machines and then computes the next iterates z^{k+1} and u^k as follows

$$u^k = \arg\min_{z\in\mathcal{Z}}\{\gamma f_1(z) + \gamma\langle\nabla q(z^k), z\rangle + V(z,z^k) + \gamma g(z)\}, \tag{30}$$

$$z^{k+1} = \arg\min_{z\in\mathcal{Z}}\{\gamma\langle\nabla q(u^k) - \nabla q(z^k), z\rangle + V(z,u^k)\}. \tag{31}$$

Then the server broadcasts z^{k+1} and u^k to all other devices.

In the entropy setup when $\mathcal{Z} \equiv \Delta$ the inner problem encountered in (31) has a closed-form solution known as entropic mirror descent (Nemirovski, 2004):

$$z^{k+1} = \frac{u^k \odot e^{-\gamma\left(\nabla q(u^k) - \nabla q(z^k)\right)}}{\mathbf{1}^\top\left(u^k \odot e^{-\gamma(\nabla q(u^k) - \nabla q(z^k))}\right)}, \tag{32}$$

where $\mathbf{1}$ is the vector of ones, exp is applied element-wise for vectors and symbols $\odot$ and / stand for the element-wise product and division respectively.

SPPs. For SPPs, operator $F(z) = [\nabla_x f(x,y),\ -\nabla_y f(x,y)]$ and $F_1(z) = [\nabla_x f_1(x,y),\ -\nabla_y f_1(x,y)]$, $G(z) := F(z) - F_1(z)$. with $z := (x,y) \in \mathcal{X}\times Y =: \mathcal{Z}$. Then the server computes

$$u^k = \arg\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\{\gamma f_1(x,y) + \gamma\left\langle Q(z^k), z\right\rangle + V(z,z^k) + \gamma g(z)\}, \tag{33}$$

$$z^{k+1} = \arg\min_{z\in\mathcal{Z}}\{\gamma\langle Q(u^k) - Q(z^k), z\rangle + V(z,u^k)\}. \tag{34}$$

Similarly to convex minimization problem, in the entropy setup ($\mathcal{X} \equiv \Delta$ and $\mathcal{Y} \equiv \Delta$) the inner problem from 34 has a closed-form solution.

Closed-form solutions for subproblems in Composite MP procedure. Next we comment on the existence of closed-form solutions for steps (6) and (7) of the SCMP procedure in the Entropy setup. Particularly for convex minimization problem (6) with $\mathcal{Z} \equiv \Delta$ and $g(v) \equiv 0$, SCMP can be rewritten as follows:

$$v^{t+\frac{1}{2}} = \frac{(z^k)^{\frac{\eta}{\eta+1}} \odot (v^t)^{\frac{1}{1+\eta}} \odot e^{-\frac{\gamma\eta}{1+\eta}h(v^t)}}{\mathbf{1}^\top\left((z^k)^{\frac{\eta}{\eta+1}} \odot (v^t)^{\frac{1}{1+\eta}} \odot e^{-\frac{\gamma\eta}{1+\eta}h(v^t)}\right)},$$

$$v^{t+1} = \frac{(z^k)^{\frac{\eta}{\eta+1}} \odot (v^t)^{\frac{1}{1+\eta}} \odot e^{-\frac{\gamma\eta}{1+\eta}h\left(v^{t+\frac{1}{2}}\right)}}{\mathbf{1}^\top\left((z^k)^{\frac{\eta}{\eta+1}} \odot (v^t)^{\frac{1}{1+\eta}} \odot e^{-\frac{\gamma\eta}{1+\eta}h\left(v^{t+\frac{1}{2}}\right)}\right)},$$

where $h(v) := \nabla f_1(v) + \nabla f(z^k) - \nabla f_1(z^k)$ and $\mathbf{1}$ is the vector of ones, exp is applied element-wise for vectors and symbols $\odot$ and $/$ stand for the element-wise product and division respectively. Similarly closed-form solutions can be obtained for SPP with $\mathcal{X} \equiv \Delta$ and $\mathcal{Y} \equiv \Delta$.

C Experiments details

We consider a two-player matrix game

$$\min_{x\in\Delta}\max_{y\in\Delta}\left[x^\top \bar{A} y := \frac{1}{m}\sum_{i=1}^m x^\top A_i y\right], \tag{35}$$

where $A_1, \ldots, A_m$ are i.i.d samples of stochastic matrix A_ξ of size $d \times d$, $m = 10^4$. Local datasets are of size $n = 2 \cdot 10^3$, the server holds also n matrices.

Next we comment on theoretical bounds for parameters L and δ for this problem since $1/L$ is the stepsize for Mirror Prox (Rogozin et al., 2021), and $1/2\delta$ is the stepsize for PAUS and the Eucidean algorithm (Kovalev et al., 2022).

Lipschitz constant L. For SPP (35), Assumption 2.6 is equivalent to the notion of smoothness.

Definition C.1 (L-smoothness) *$f(x,y)$ is $(L_{xx}, L_{xy}, L_{yx}, L_{yy})$-smooth if for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,*

$$\begin{aligned}
\|\nabla_x f(x,y) - \nabla_x f(x',y)\|_* &\le L_{xx}\|x - x'\|,\\
\|\nabla_x f(x,y) - \nabla_x f(x,y')\|_* &\le L_{xy}\|y - y'\|,\\
\|\nabla_y f(x,y) - \nabla_y f(x,y')\|_* &\le L_{yy}\|y - y'\|,\\
\|\nabla_y f(x,y) - \nabla_y f(x',y)\|_* &\le L_{yx}\|x - x'\|.
\end{aligned}$$

Then we define $L = \max\{L_{xx}, L_{xy}, L_{yx}, L_{yy}\}$ and seek to estimate L. We equip both $\mathcal{X} := \Delta$ and $\mathcal{Y} := \Delta$ with the ℓ_1-norm. The corresponding dual norm is the ℓ_∞-norm. By the Definition C.1

$$\|\bar{A}(y - y')\|_\infty \le \|\sum_{i=1}^d A^{(i)}(y_i - y_i')\|_\infty \le \|\bar{A}\|_{\max}\|y - y'\|_1, \tag{36}$$

where we used $A^{(i)}$ for the i-th column of A, and $\|A\|_{\max}$ for the maximal entry of A (in absolute value). Thus, $L_{xy} = L_{yx} = \|\bar{A}\|_{\max}$, and $L_{xx} = L_{yy} = 0$. Hence, $L = \|\bar{A}\|_{\max}$.

δ-similarity. Now we seek to estimate δ. We use Assumption 2.7 particularly for $F(z) = [\nabla_x f(x,y),\ -\nabla_y f(x,y)]$

$$\begin{aligned}
&\left\|\begin{bmatrix}\nabla_x f_1(x,y)\\ -\nabla_y f_1(x,y)\end{bmatrix} - \begin{bmatrix}\nabla_x f(x,y)\\ -\nabla_y f(x,y)\end{bmatrix} - \begin{bmatrix}\nabla_{x'} f_1(x',y')\\ -\nabla_{y'} f_1(x',y')\end{bmatrix} + \begin{bmatrix}\nabla_{x'} f(x',y')\\ -\nabla_{y'} f(x',y')\end{bmatrix}\right\|_\infty\\
&\le \delta\left\|\begin{bmatrix}x\\ y\end{bmatrix} - \begin{bmatrix}x'\\ y'\end{bmatrix}\right\|_1.
\end{aligned} \tag{37}$$

where global objective is

$$f(x,y) = x^\top \bar{A} y := \frac{1}{m}\sum_{i=1}^m x^\top A_i y,$$

and local (stored on the server) objective is

$$f_1(x,y) = x^\top \bar{A}^{N_1} y := \frac{1}{N_1}\sum_{\ell=1}^{N_1} x^\top A_\ell y.$$

Using this, we can rewrite (37) as follows

$$\left\| \begin{bmatrix} (\bar{A} - \bar{A}^{N_1})x \\ -(\bar{A} - \bar{A}^{N_1})^\top y \end{bmatrix} - \begin{bmatrix} (\bar{A} - \bar{A}^{N_1})x' \\ -(\bar{A} - \bar{A}^{N_1})^\top y' \end{bmatrix} \right\|_\infty \leq \delta \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\|_1 . \tag{38}$$

Thus,

$$\left\| \begin{bmatrix} (\bar{A} - \bar{A}^{N_1})(x - x') \\ -(\bar{A} - \bar{A}^{N_1})^\top (y - y') \end{bmatrix} \right\|_\infty \leq \delta \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\|_1 . \tag{39}$$

Let us define $z = (x, y)$ and $z' = (x', y')$. Then we rewrite (39) as follows

$$\|\mathbf{A}(z - z')\|_\infty \leq \delta \, \|z - z'\|_1 ,$$

where

$$\mathbf{A} := \begin{pmatrix} (\bar{A} - \bar{A}^{N_1}) & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & -(\bar{A} - \bar{A}^{N_1})^\top \end{pmatrix} .$$

Here $\mathbf{0}_{d\times d}$ is the zero matrix of size $d \times d$. By the same arguments as in (36) we conclude that $\delta = \|\mathbf{A}\|_{\max}$.