A Stochastic Second-Order Proximal Method for Distributed Optimization

Chenyang Qiu✉, Shanying Zhu✉, Zichong Ou✉, and Jie Lu✉, Member, IEEE

Abstract—We propose a distributed stochastic second-order proximal (St-SoPro) method that enables agents in a network to cooperatively minimize the sum of their local loss functions without any centralized coordination. St-SoPro incorporates a decentralized second-order approximation into an augmented Lagrangian function, and randomly samples the local gradients and Hessian matrices to update, so that it is efficient in solving large-scale problems. We show that for restricted strongly convex and smooth problems, the agents linearly converge in expectation to a neighborhood of the optimum, and the neighborhood can be arbitrarily small under proper parameter settings. Simulations over real machine learning datasets demonstrate that St-SoPro outperforms several state-of-the-art methods in terms of convergence speed as well as computation and communication costs.

Index Terms—Distributed optimization, second-order method, stochastic optimization.

I. INTRODUCTION

STOCHASTIC optimization algorithms have been flourishing recently due to their appealing efficiency in machine learning [1], [2], and parallel stochastic algorithms are often used to process large datasets [3], [4]. However, such methods need a central node to ensure the consistency of the variables from all the nodes, so that the communication burden of the central node becomes the bottleneck that restricts the algorithm performance. On the other hand, a collection of distributed optimization algorithms have been proposed over the past decade in order to tackle various network control and resource allocation problems, where agents in a network only communicate with their neighbors and do not rely on any centralized coordination, eliminating potential communication bottlenecks in the computing infrastructure [5].

Manuscript received 16 November 2022; revised 7 January 2023; accepted 7 February 2023. Date of publication 14 February 2023; date of current version 29 February 2023. This work was supported by the National Natural Science Foundation of China under Grant 61603254. Recommended by Senior Editor F. Dabbene. (Corresponding author: Jie Lu.)

Chenyang Qiu and Zichong Ou are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: qichy@shanghaitech.edu.cn; ouzch@shanghaitech.edu.cn).

Shanying Zhu is with the Department of Automation and the Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shyzhu@sjtu.edu.cn).

Jie Lu is with the School of Information Science and Technology and the Shanghai Engineering Research Center of Energy Efficient and Custom AI IC, ShanghaiTech University, Shanghai 201210, China (e-mail: lujie@shanghaitech.edu.cn).

Digital Object Identifier 10.1109/LCSYS.2023.3244740

Inheriting the merits of the above two algorithm types, distributed stochastic optimization algorithms have been attracting a lot of recent interest. For example, [6], [7], [8], [9] consider smooth, strongly convex optimization problems. The distributed stochastic gradient descent (DSGD) method in [6] and the exact diffusion method with adaptive step-sizes (EDAS) in [7] are shown to attain the optimal sublinear convergence rate for a centralized stochastic gradient descent (SGD) method. The distributed stochastic gradient tracking (DSGT) method in [8] is guaranteed to linearly converge to a neighborhood of the optimal solution. In [9], an inexact accelerated-decentralized augmented Lagrangian (IDEAL) framework is proposed, in which SGD can be employed as an inner solver, leading to a sublinear rate of convergence to the optimum. For smooth, nonconvex problems, [10] designs a distributed primal-dual SGD algorithm with adaptive stepsizes (DPD-SGD-T), which achieves sublinear convergence to the optimum under the Polyak-Łojasiewicz (PL) condition. Furthermore, by utilizing proximal operators, [11], [12] are capable of solving nonsmooth optimization problems. In particular, the distributed stochastic proximal-gradient algorithm in [11] achieves linear convergence for strongly convex but nonsmooth problems, which, however, needs to regularly compute the exact gradients. The stochastic proximal primal-dual algorithm with momentum [12] converges to a neighborhood of some stationary point at a sublinear rate for a class of nonsmooth, nonconvex optimization problems. All the aforementioned convergence results are established in expectation.

The aforementioned distributed stochastic algorithms are all first-order by nature. As second-order information often leads to more accurate approximation and accelerates problem solving, we endeavor to develop a second-order distributed stochastic optimization algorithm. To this end, we choose SoPro [13], a deterministic distributed second-order proximal algorithm, as the cornerstone. SoPro is developed by introducing a novel decentralized second-order approximation of an augmented Lagrangian function to the method of multipliers [14], and its convergence performance outperforms that of many existing methods in the deterministic setting.

In this letter, instead of letting each agent compute the exact local gradient and Hessian matrix determined by all its local data like SoPro, we allow each agent to update using stochastic gradient and Hessian, which come from a uniformly and randomly chosen batch of samples from its local loss. Such a stochastic variant of SoPro can significantly enhance the computational and memory efficiency. We call this algorithm a stochastic second-order proximal algorithm, referred to as St-SoPro. Under the assumptions that the local loss functions are smooth and convex and that their sum (i.e., the global loss...
function) is restricted strongly convex [13], we show that our proposed St-SoPro algorithm linearly converges to a neighborhood of the optimal solution in expectation over undirected networks. In particular, we provide an explicit upper bound on its ultimate suboptimality, and illustrate that this upper bound can be made arbitrarily small as long as the parameters are properly set. From the theoretical perspective, we consider a weaker convexity condition than [6], [7], [8], [9] and achieve a better convergence rate than [6], [7], [9], [10], [12]. Finally, we validate the superior performance of St-SoPro in comparison with several related methods over real machine learning datasets in terms of convergence speed, communication load, computational efficiency, and classification accuracy.

This letter is organized as follows. Section II formulates the problem, and Section III develops St-SoPro. Section IV provides the convergence analysis, Section V presents the numerical results, and Section VI concludes this letter.

Notation: For any differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $\nabla f(x)$ denotes its gradient at $x \in \mathbb{R}^d$, and if $f$ is twice-differentiable, $\nabla^2 f(x)$ denotes its Hessian matrix at $x$. For any set $S$, $|S|$ represents the cardinality of $S$. In addition, $\otimes$ is the Kronecker product, $(\cdot, \cdot)$ is the Euclidean inner product, and $\| \cdot \|$ is the $\ell_2$ norm. We use $\mathbf{0}_d$, $\mathbf{I}_d$, $\mathbf{O}_d$, $\mathbf{I}$ to denote the $d$-dimensional all-zero vector, all-one vector, zero matrix, and identity matrix, respectively. Also, diag($A_1, \ldots, A_n$) represents the block diagonal matrix whose diagonal blocks are sequentially $A_1, \ldots, A_n$. Given $A \in \mathbb{R}^{d \times d}$, $A^2$ is the square root of $A$ (i.e., $A^2 A^2 = A$), and we write $A \succeq A_0$ if it is positive semidefinite and $A > A_0$ if it is positive definite. For any $A \succeq O_d$ and $x \in \mathbb{R}^d$, $\|x\|^2_A = x^T A x$, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the largest and smallest real eigenvalues of $A$, respectively, and $A^T$ is $A$’s pseudoinverse.

II. PROBLEM FORMULATION

Consider a set $\mathcal{V} = \{1, \ldots, N\}$ of agents, where the agents are connected through the link set $\mathcal{E} \subseteq \{(i, j) \subseteq \mathcal{V} \times \mathcal{V} | i \neq j\}$. We model such a network as a connected undirected graph $(\mathcal{V}, \mathcal{E})$, and denote the set of each agent $i$’s neighbors by $\mathcal{N}_i = \{j \in \mathcal{V} | (i, j) \in \mathcal{E}\}$. Suppose each agent $i$ has $\mathcal{C}_i$ local samples $\xi_{i,n}$, $n = 1, \ldots, \mathcal{C}_i$, and each sample $\xi_{i,n}$ corresponds to a sample loss function $l_i(x, \xi_{i,n})$. Then, the local loss function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ of each agent $i$ is given by

$$f_i(x) = \frac{1}{\mathcal{C}_i} \sum_{n=1}^{\mathcal{C}_i} l_i(x, \xi_{i,n}).$$

Suppose all the agents attempt to minimize the total loss throughout the network by solving

$$\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{V}} f_i(x).$$

(1)

Below, we impose the assumptions on problem (1).

Assumption 1: Problem (1) satisfies the following:

a) There exists an optimal solution $x^* \in \mathbb{R}^d$ to problem (1), and $\sum_{i \in \mathcal{V}} f_i(x)$ is restricted strongly convex with respect to $x^*$ with convexity parameter $m_j > 0$, i.e., $\left\langle \sum_{i \in \mathcal{V}} \nabla f_i(x) - \sum_{i \in \mathcal{V}} \nabla f_i(x^*), x - x^* \right\rangle \geq m_j \|x - x^*\|^2 \forall x \in \mathbb{R}^d$.

b) For any given $\xi_{i,n}$, $l_i(x, \xi_{i,n})$ is twice continuously differentiable and convex, and there exists $M_i > 0$ such that $l_i(x, \xi_{i,n})$ is $M_i$-smooth.

Assumption 1b) suggests that there is $m_i \in [0, M_i]$ such that for any given $x, y \in \mathbb{R}^d$ and $\xi_{i,n}$, we have $m_i \|x - y\|^2 \leq \langle \nabla l_i(x, \xi_{i,n}) - \nabla l_i(y, \xi_{i,n}), x - y \rangle \leq M_i \|x - y\|^2$ and $m_i d \leq \nabla^2 l_i(x, \xi_{i,n}) \preceq M_i d$. Also, the restricted strong convexity in Assumption 1a) guarantees the uniqueness of the optimum $x^*$, and is weaker than the commonly adopted strong convexity condition [6], [7], [8], [9].

Problem (1) requires that the agents reach a consensus while minimizing all the sample losses throughout the network. Indeed, a wide range of real-world problems can be cast into the form of problem (1), such as distributed model predictive control [15], distributed spectrum sensing [16], and logistic regression [17]. Under many circumstances, these engineering problems involve huge datasets. Thus, we focus on solving problem (1) in a fully decentralized and stochastic fashion. Specifically, we only allow each agent to communicate with its neighbors and use a randomly chosen subset of its local samples to compute.

III. STOCHASTIC SECOND-ORDER PROXIMAL METHOD

In this section, we develop a distributed stochastic algorithm for solving problem (1) over undirected networks.

To this end, we first provide a brief review of the distributed (deterministic) second-order proximal algorithm (SoPro) in [13]. Note that problem (1) is equivalent to

$$\begin{align*}
\min_{x \in \mathbb{R}^{Nd}} & f(x) := \sum_{i \in \mathcal{V}} f_i(x_i) \\
\text{subject to} & \ W^2 x = 0_{Nd},
\end{align*}$$

(3)

where $x = (x_1^T, \ldots, x_N^T)^T$, $W = P \otimes I_d \succeq O_{Nd}$, and $\left[P\right]_{ij} = \begin{cases} 
\sum_{i \in \mathcal{N}_j} \beta \ \text{otherwise,}
\end{cases}$

$$\begin{align*}
\left[P\right]_{ij} &= \begin{cases} 
\sum_{i \in \mathcal{N}_j} P_{ij} \beta, & i = j, \\
-p_{ij} & j \in \mathcal{N}_i, \\
0, & \text{otherwise,}
\end{cases}
\end{align*}$$

where $p_{ij} = p_{ji} > 0$ is the weight of the edge $(i, j) \in \mathcal{E}$. Note that the null space of $P$ is span$\{1_N\}$. Also, the unique optimal solution of problem (3) is $x^* = ((x_1^*)^T, \ldots, (x_N^*)^T)^T$.

The application of the method of multipliers [14] to solve (3) gives the following: Starting from any $x^0 \in \mathbb{R}^{Nd}$,

$$\begin{align*}
x^{k+1} &= \arg \min_{x \in \mathbb{R}^{Nd}} L_\beta(x, v^k), \\
v^{k+1} &= v^k + \beta W^2 x^{k+1},
\end{align*}$$

(4)

(5)

where $x^k = ((x_1^k)^T, \ldots, (x_N^k)^T)^T$ and $v^k$ are the primal and dual variables, respectively, and $L_\beta(x, v) : \mathbb{R}^{Nd} \times \mathbb{R}^{Nd} \rightarrow \mathbb{R}$ is an augmented Lagrangian function given by $L_\beta(x, v) = f(x) + v^T W^2 x + \frac{\beta}{2} \|W^2 x\|^2$, $\beta > 0$.

Since (4)--(5) cannot be applied in a distributed way, the SoPro algorithm in [13] introduces a decentralized second-order proximal approximation of $L_\beta(x, v^k)$ in (4) and applies a variable change to (5). Specifically, it replaces $L_\beta(x, v^k)$ with its second-order Taylor’s expansion at $x^k$. Then, it replaces the remaining coupling term $\frac{1}{2} (x - x^k)^T D(x - x^k)$ in the primal update with $\frac{1}{2} (x - x^k)^T D(x - x^k)$, where $D = \text{diag}(D_1, \ldots, D_N)$ is a symmetric block diagonal matrix satisfying $\nabla^2 f_i(x) + D_i \succeq O_d \forall x \in \mathbb{R}^d \forall i \in \mathcal{V}$. Furthermore, we define $q^k = ((q_1^k)^T, \ldots, (q_N^k)^T)^T = W^2 v^k$ as a substitute for $v^k$, and $q^k$ can be ensured to identically stay in the range of
\(W^k\) by letting \(\sum_{i \in V} q_i^0 = 0\). To summarize, SoPro takes the following form: Starting from \(q_0^i\) satisfying \(\sum_{i \in V} q_i^0 = 0\),
\[
    x^{k+1} = x^k - (\nabla f(x^k) + D)^{-1}(\nabla f(x^k) + \beta Wx^k + q^k) ,
    \]
\[
    q^{k+1} = q^k + \beta Wx^{k+1}, \quad \forall k \geq 0,
\]
where \(\nabla f(x^k) = (\nabla f_1(x_1^k), \ldots, \nabla f_N(x_N^k))^T\) and \(\nabla^2 f(x^k) = \text{diag}(\nabla^2 f_1(x_1^k), \ldots, \nabla^2 f_N(x_N^k))\) satisfying \(\nabla^2 f(x^k) + D > O_{Nd}\).

The primal update of SoPro (6) requires that each agent uses all its local samples. However, the agents may only be able to access or process a portion of their local samples at one time, especially in the big data scenario. Motivated by this, we consider approximating the gradient \(\nabla f(x^k)\) and the Hessian \(\nabla^2 f(x^k)\) in (6) via a stochastic gradient \(g(x^k)\) and a stochastic Hessian \(h(x^k)\) given by
\[
g(x^k) = (g_1(x_1^k)^T, \ldots, g_N(x_N^k)^T)^T,
\]
where each \(g_i(x_i^k) = \frac{1}{\vert G^i \vert} \sum_{n \in G^i} \nabla l_i(x_i^k, \xi_{i,n})\),
\[
h(x^k) = \text{diag}(h_1(x_1^k), \ldots, h_N(x_N^k)) ,
\]
where each \(h_i(x_i^k) = \frac{1}{\vert S^i \vert} \sum_{n \in S^i} \nabla^2 l_i(x_i^k, \xi_{i,n})\).

Here, for each agent \(i \in V\), \(G^i\) and \(S^i\) are two independent random sample sets uniformly chosen from \(\{1, \ldots, C_i\}\) without replacement, so that \(g(x^k)\) and \(h(x^k)\) are unbiased, i.e.,
\[
E_{G^i}[g(x^k)] = \nabla f_i(x^k), \quad E_{S^i}[h(x^k)] = \nabla^2 f_i(x^k)
\]
for all \(x^k\). Due to (2), \(m_i \delta \leq h_i(x_i^k) \leq M \delta \delta \forall i \in V\) and \(\Lambda_m \leq h(x^k) \leq \Lambda_M\),
\[
\Lambda_m = \text{diag}(m_1, \ldots, m_N) \otimes I_d \succ O_{Nd} \quad \text{and} \quad \Lambda_M = \text{diag}(M_1, \ldots, M_N) \otimes I_d \succ O_{Nd}.
\]

Using the above randomly sampled gradient and Hessian, we obtain the following stochastic variant of SoPro: Starting from any \(q^0\) such that \(\sum_{i \in V} q_i^0 = 0\),
\[
x^{k+1} = x^k - (h(x^k) + D)^{-1}(g(x^k) + \beta Wx^k + q^k) ,
    \]
\[
q^{k+1} = q^k + \beta Wx^{k+1}, \quad \forall k \geq 0,
\]
where each \(D_i\) satisfies \(h_i(x_i^k) + D_i > O_d \forall x \in R^d\), i.e.,
\(h(x) + D > O_{Nd} \forall x \in R^{Nd}\), so that (10) is well-posed. The above initialization and updates compose our proposed stochastic second-order proximal (St-SoPro) method. The distributed implementation of St-SoPro over the undirected network \((V, E)\) is described in Algorithm 1, in which \(\beta_i \forall i \in V\) are auxiliary variables for better presentation.

IV. CONVERGENCE ANALYSIS

This section provides the convergence analysis of St-SoPro.

We first impose an assumption to bound the deviation of each sample loss \(l_i(x_i^k, \xi_{i,n})\) from each entire local loss \(f_i(x_i^k)\) in terms of their gradients.

**Assumption 2:** There exists \(\sigma > 0\) such that for any \(i \in V\) and \(n = 1, \ldots, C_i\), \(\|l_i(x_i^k, \xi_{i,n}) - \nabla f_i(x_i^k)\| \leq \sigma \forall x_i \in R^d\).

Assumption 2 is satisfied in many real-world scenarios such as logistic regression [17] and spectrum sensing with LASSO [16]. To simplify the notation, below we let \(C_i = C \forall i \in V\) and \(\vert G^i \vert = G \forall i \in V \forall k \geq 0\). We also abbreviate
\[
\sum_{i \in V} q_i^0 = 0.
\]

**Algorithm 1 St-SoPro**

1. **Initialization:**
   1. Each agent \(i \in V\) sets \(q_i^0 \in R^d\) such that \(\sum_{i \in V} q_i^0 = 0\) (or simply sets \(q_i^0 = 0\)).
   2. Each agent \(i \in V\) arbitrarily sets \(x_i^0 \in R^d\), and sends \(x_i^0\) to each neighbor \(j \in N_i\). After receiving \(x_j^0 \forall j \in N_i\), each agent \(i \in V\) sets \(q_i^0 = \sum_{j \in N_i} p_{ij}(x_j^0 - x_i^0)\).
   3. For \(k = 0, 1, 2, \ldots\)
      4. **end for**

\(E_{G^i}[\cdot]\) and \(E_{S^i}[\cdot]\) to \(E[\cdot]\). From [18, Ch. 2], Assumption 2 yields
\[
E \left[ \|g(x^k) - \nabla f(x^k)\|^2 \right] = \sum_{i \in V} E \left[ \|g_i(x_i^k) - \nabla f_i(x_i^k)\|^2 \right]
\]
\[
\leq N\tau^2\sigma^2, \quad \tau := \frac{C - G}{CG}.
\]
This is consistent with the fact that when computing the stochastic gradient \(g(x^k)\), if we reduce the number of \(G\) of randomly selected samples, then the discrepancy between \(g(x^k)\) and \(\nabla f(x^k)\) becomes larger in expectation.

Next, for the sake of presenting the convergence result, we introduce the following notations. According to [13], any \(v \in R^{Nd}\) satisfying \(\nabla f(x) = -W^T v\) is a dual optimum of problem (3). Thus, we define
\[
v^* = -\left(W^T\right)^{1/2} \nabla f(x^*)
\]
as a particular dual optimum of (3). Also, throughout this section, we let \(v^k = (W^T)^{1/2} v^k\), \(z^k = (z^k)^T, (z^k)^T\), and \(z^* = (z^*)^T, (z^*)^T\). In addition, although the restricted strong convexity of \(\sum_{i \in V} f_i(x)\) cannot guarantee the same property for \(f(x)\), it guarantees \(f_{\beta}(x) := f(x) + \frac{\beta}{2} \|x\|^2\) is restricted strongly convex with respect to \(x^*\) [13, Lemma 1], i.e., for any \(x \in R^{Nd}\), there exists \(m_{\beta} > 0\) such that
\[
\|f_{\beta}(x) - f_{\beta}(x^*) - (x - x^*)\| \geq m_{\beta} \|x - x^*\|^2.
\]
Such a convexity parameter \(m_{\beta}\) is used in our parameter condition, which says that
\[
D > \frac{\Lambda_M}{2(1 - \eta_s)} + \frac{(\Lambda_M - \Lambda_m)^2}{8 \eta_s m_{\beta}} + \frac{\Lambda_M - 3\Lambda_m}{2} + \beta \left(\frac{I_{Nd}}{2} + W\right)
\]
for some \(\eta_s \in (0, 1)\). Moreover, (14) ensures the positive definiteness of \(R := \Lambda_m + \frac{\Lambda_M}{\Lambda_m} + D \succ 0\) and \(Q := \text{diag}(\beta R, I_{Nd})\). Furthermore, it follows from (9) and (14) that the condition \(h(x) + D > O_{Nd}\) required in Section III holds.

With the above notations, we provide our main result below.

**Theorem 1:** Suppose Assumptions 1 and 2 hold. If (14) holds for some \(\eta_s \in (0, 1)\), then \(x^k\) converges linearly to a
neighborhood of \( z^* \) in expectation, i.e., there exist \( \delta_x \in (0, 1) \) and \( \Gamma > 0 \) such that for each \( k \geq 0 \),
\[
E\left[\|z^{k+1} - z^*\|_Q^2\right] \leq (1 - \delta_x)E\left[\|z^k - z^*\|_Q^2\right] + \Gamma N^2 \sigma^2,
\]
(15)
\[
\limsup_{k \to \infty} E\left[\|z^k - z^*\|_Q^2\right] \leq \frac{\Gamma N^2 \sigma^2}{\delta_x}.
\]
(16)

In particular, given any \( c_1 > 0 \), \( \beta = \frac{2(1+c_1)\delta_x}{\kappa W} + 2 \) and
\[
\delta_x \equiv \min_{c_2 > 0} \left\{ \frac{\beta \lambda W \kappa_{\theta_0, \eta_1}}{2(1+c_1)(\Lambda_M + D)^2}, \frac{1-\eta_1}{(1+c_1)(1+c_2)} \right\}.
\]
(17)
where \( \lambda_W \) is the smallest non-zero eigenvalue of \( W \) and \( c_0 \in (0, 2\eta, m_\beta) \) is such that \( \kappa_{\theta_0, \eta_1} \equiv \lambda_W \min(R - \frac{\Lambda_M - \lambda_W}{\lambda_W}, \frac{1}{4c_0}) \).

**Proof:** See the Appendix.

The parameter condition (14) can always be satisfied by proper parameter selections, which is exemplified as follows. For simplicity, we first arbitrarily fix \( \beta > 0, \eta_1 \in (0, 1), \) and \( W \) defined in Section III and then select \( D \) subject to (14). It remains to evaluate the convexity parameter \( m_\beta \). Similar to the analysis in [13], we define
\[
\zeta(\gamma, \beta) \equiv \min_{\gamma} \left\{ \frac{m_\beta}{\sqrt{2}} - 2\gamma W, \frac{\beta \lambda W}{\sqrt{2}(1+c_1)} \right\}
\]
with \( m_\beta \) given in Assumption 1a) and \( M = \max_{i\in\mathcal{V}} M_i > 0 \). It can be shown that \( \zeta(\gamma) > 0 \) if and only if \( \gamma \in (0, \gamma(\beta, W)) \), and \( m_\beta \) can be any positive value of \( \zeta(\gamma) \). Moreover, the maximum value of \( \gamma \) is attained at the unique positive root of
\[
4MN\gamma^2 + (\beta N\lambda_W - 2m_\beta)\gamma^2 + 4MN\gamma - 2m_\beta = 0.
\]

Although the entire parameter selection procedure involves a few global quantities, these quantities can indeed be estimated in a distributed way using similar ideas as [13].

Like its deterministic counterpart SoPro, St-SoPro achieves a faster convergence rate with larger \( \min_{i\in\mathcal{V}} m_i \), smaller \( \max_{i\in\mathcal{V}} M_i \), or smaller \( \lambda_W \). In addition, (16) implies that the ultimate bound drops with the decrease of \( \Gamma \). Hence, essentially, larger random sample sets for computing the stochastic gradients lead to smaller optimality error.

In fact, the expected distance between \( x^k \) and \( x^* \) can eventually be arbitrarily small with proper parameters. To see this, for simplicity, let \( m_i = m > 0 \) and \( M_i = M \geq m \) for all \( i \in \mathcal{V} \), and pick any \( \eta_i \in (0, 1) \), \( c_0 \in (0, 2\eta, m_\beta) \), and \( c_1 > 0 \). We choose, for example, \( D = a4I_{\mathcal{Bd}} \) with \( \alpha = \left( \frac{1}{2} + \frac{1}{\max(W)} \right) \beta + \mu, \beta > 0, \) and \( \mu > \frac{M - 2m}{2} + \frac{M}{2(1-\eta_1)} + \frac{(M-m)^2}{4c_0} \), so that (14) holds. This also gives \( \kappa_{\theta_0, \eta_1} = \mu = \frac{M - 2m}{2} - \frac{M}{2(1-\eta_1)} - \frac{(M-m)^2}{4c_0} > 0 \). From (16),
\[
\limsup_{k \to \infty} E[\|x^k - x^*\|^2] \leq \frac{\Gamma N^2 \sigma^2}{\delta_x (\beta + \sqrt{\beta + \max(W)}) + \mu}.
\]
It can thus be shown that as \( \beta \to \infty \), such an upper bound on \( \limsup_{k \to \infty} E[\|x^k - x^*\|^2] \) goes to zero. Since this bound is continuous at \( \beta \), given any \( \epsilon > 0 \), the above parameter setting with a sufficiently large \( \beta \) guarantees \( \limsup_{k \to \infty} E[\|x^k - x^*\|^2] < \epsilon \).

**VI. NUMERICAL EXPERIMENT**

This section compares the practical convergence performance of St-SoPro with several state-of-the-art distributed stochastic optimization algorithms.

| Algorithm | d | N | da | C1 | G1 | S1 | λ |
|-----------|---|---|----|----|----|----|---|
| a4a       | 123 | 20 | 5  | 259| 80 | 10 | 10^{-2} |
| mushrooms | 112 | 10 | 3  | 600| 80 | 25 | 10^{-2} |

In the numerical experiment, we intend to learn linear classifiers by solving \( h_2 \)-regularized logistic regression of the following form over a randomly generated, undirected, and connected network:
\[
\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{V}} 1 - \frac{C_i}{2} \|x^i\|^2 + \ln(1 + e^{-a_i^Tx^i b_i})
\]
where \( \lambda > 0 \) is the regularization parameter and \( [a_i^i, b_i] \) are the data samples. Our experiment is conducted on two standard real datasets \( a4a \) and \( mushrooms \) from the LIBSVM library. Table I lists the values of the problem and network parameters corresponding to these two datasets, including the problem dimension \( d \), the number \( N \) of agents, the network’s average degree \( \bar{d}_a = \sum_{i \in \mathcal{V}} \bar{\xi}_i \), the total number \( C_i \) of samples that we assign to each agent \( i \), the sizes \( |G_i| \) and \( |S_i| \) of the random sample sets that each agent \( i \) chooses per iteration, as well as the regularization parameter \( \lambda \).

The simulations include DSGD [6], EDAS [7], DSGT [8], DPD-SGD-T [10] and IDEAL-SGD [9], which are all first-order methods, for comparison with our proposed St-SoPro. We fine-tune all the algorithm parameters so that the algorithms reach a given accuracy \((2 \times 10^{-4} \text{ for } a4a \text{ and } 10^{-1} \text{ for } mushrooms)\) within fewest possible iterations.

Figures 1(a)–(c) and 2(a)–(c) plot the evolutions of the optimality error \( \frac{1}{2} \sum_{i \in \mathcal{V}} \|x^i - x^*\|^2 \) generated by the aforementioned algorithms over \( a4a \) and \( mushrooms \) with respect to the number of iterations, the number of communication bits (set as 32 times the number of transmitted real scalars according to [19]), and computation time. Observe that St-SoPro converges faster than the other algorithms to reach the given accuracy, validating its computational and communication efficiency. It is worth mentioning that although St-SoPro is a second-order method, its computational cost can be comparable with the first-order methods when addressing those common machine learning problems. Figures 1(d) and 2(d) present the correct rates of classification on the test sets upon completing each iteration of these algorithms as training, whereby St-SoPro outperforms the others in the training effect.

**VI. CONCLUSION**

We have developed St-SoPro, a distributed stochastic second-order proximal method, for addressing smooth, restricted strongly convex optimization over undirected networks. Different from the existing first-order distributed stochastic algorithms, St-SoPro incorporates a second-order approximation of an augmented Lagrangian function and randomly samples each local gradient and Hessian. We show that St-SoPro linearly converges to a neighborhood of the optimum in expectation, and the neighborhood can be arbitrarily small. Simulations over two real datasets demonstrate that St-SoPro is both computationally and communication-wise efficient.
The following lemma intends to bound the difference between $E[\|z^k - z^*\|^2_Q]$ and $E[\|z^{k+1} - z^*\|^2_Q]$.

**Lemma 1:** For each $k \geq 0$, we have

$$E\left[\|z^k - z^*\|^2_Q\right] - E\left[\|z^{k+1} - z^*\|^2_Q\right] \leq \beta (2\eta_0 \mu_0 - c_0) E\left[\|x^k - x^*\|^2\right] + \beta^2 (1 - \eta_0) E\left[\|x^k\|^2\right]$$

$$- \beta \left(\frac{x^k}{\mu_0} + \beta W - R\right) - 2\eta_0 \alpha^2,$$

where $c_0 > 0$ is arbitrary and $\mu_0 \eta_0 = \Lambda_M / (2(1 - \eta_0)) + (\Lambda_M - \Lambda_m)^2 / (4c_0) + \Lambda_M - \Lambda_m + \frac{\beta_0M_d}{4}$. 

**Proof:** We first equivalently expand the left-hand side of (18). Similar to [13, eq. (27)], we derive

$$E\left[\|z^k - z^*\|^2_Q\right] - E\left[\|z^{k+1} - z^*\|^2_Q\right] = E\left[\|z^k - z^{k+1}\|^2_Q\right] + 2\beta \left(\frac{x^k - x^*}{\lambda_0\eta_0} + \beta W - R\right)$$

$$+ 2\beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right).$$

Then, using (5) and $W\hat{x}^* = 0_{M_d}$, we obtain $\langle v^k - v^k \rangle = -\beta (x^k - x^* - \hat{x}^k - v^k)$, and

$$E\left[\|z^k - z^{k+1}\|^2_Q\right] + 2\beta \left(\frac{x^k - x^*}{\lambda_0\eta_0} + \beta W - R\right)$$

$$+ 2\beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right).$$

By incorporating (21) into (20) and then combining the resulting equation with (19), we have

$$E\left[\|z^k - z^*\|^2_Q\right] - E\left[\|z^{k+1} - z^*\|^2_Q\right]$$

$$= 2\beta \left(\frac{x^k - x^*}{\lambda_0\eta_0} + \beta W - R\right) + 2\beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*ight)$$

$$+ 2\beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right).$$

Subsequently, we provide a lower bound for the first term on the right-hand side of (22). To do so, we utilize the AM-GM inequality and (11) to derive

$$E\left[\|z^k - z^*\|^2_Q\right] - E\left[\|z^{k+1} - z^*\|^2_Q\right]$$

$$\geq - (1 - \eta_0) E\left[\|x^k\|^2\right]$$

$$- \beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right)$$

$$- \beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right).$$

Due to the Lipschitz continuity of each $\nabla f_i$ and the unbiasedness of $g(x^k)$, we have $E[\|x^k - x^*\|^2, g(x^k) - \nabla f(x^k)] = E[\|x^k - x^*\|^2, \nabla f(x) - \nabla f(x^*)] \geq E[\|x^k - x^*\|^2, \nabla f(x^k) - \nabla f(x^*)].$ We multiply this inequality by $(1 - \eta_0)$ and then add it to (23), which leads to

$$E\left[\|x^k - x^*\|^2, g(x^k) - \nabla f(x^*)\right]$$

$$- \eta_0 E\left[\|x^k - x^*\|^2, \nabla f(x^k) - \nabla f(x^*)\right].$$

By incorporating (21) into (20) and then combining the resulting equation with (19), we have

$$E\left[\|z^k - z^*\|^2_Q\right] - E\left[\|z^{k+1} - z^*\|^2_Q\right]$$

$$= 2\beta \left(\frac{x^k - x^*}{\lambda_0\eta_0} + \beta W - R\right) + 2\beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right)$$

$$+ 2\beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right).$$

Subsequently, we provide a lower bound for the first term on the right-hand side of (22). To do so, we utilize the AM-GM inequality and (11) to derive

$$E\left[\|z^k - z^*\|^2_Q\right] - E\left[\|z^{k+1} - z^*\|^2_Q\right]$$

$$\geq - (1 - \eta_0) E\left[\|x^k\|^2\right]$$

$$- \beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right)$$

$$- \beta \left(\|x^k - x^{k+1}\|^2 + x^k - x^*\right).$$

Due to the Lipschitz continuity of each $\nabla f_i$ and the unbiasedness of $g(x^k)$, we have $E[\|x^k - x^*\|^2, g(x^k) - \nabla f(x^k)] = E[\|x^k - x^*\|^2, \nabla f(x) - \nabla f(x^*)] \geq E[\|x^k - x^*\|^2, \nabla f(x^k) - \nabla f(x^*)].$ We multiply this inequality by $(1 - \eta_0)$ and then add it to (23), which leads to

$$E\left[\|x^k - x^*\|^2, g(x^k) - \nabla f(x^*)\right]$$

$$- \eta_0 E\left[\|x^k - x^*\|^2, \nabla f(x^k) - \nabla f(x^*)\right].$$
\[
\begin{align*}
\geq & -\mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}\right)^{2}_{(\mathbf{A}_{M}(1-\eta_{s})+\beta_{M})/4}\right] - N\tau \sigma^{2}/\beta. \quad (24)
\end{align*}
\]

Because of (13) and \(\mathbf{W}^{s} = 0_{nM}\), we have \(\mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}\right)^{2}_{(\mathbf{A}_{M}(1-\eta_{s})+\beta_{M})/4}\right] \geq m_{\mathbf{E}}\mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}\right)^{2}\right] - \frac{2}{\mathbf{E}\left[\left(\mathbf{x}^{k}\right)^{2}_{W}\right]}\). This, along with (24), results in
\[
\begin{align*}
\mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}, g(\mathbf{x}^{k}) - \nabla f(\mathbf{x}^{k})\right)^{2}_{(\mathbf{A}_{M}(1-\eta_{s})+\beta_{M})/4}\right] - N\tau \sigma^{2}/\beta \\
+ \eta_{s}m_{\mathbf{E}}\mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}\right)^{2}\right] - \eta_{s}\mathbf{E}\left[\left(\mathbf{x}^{k}\right)^{2}_{W}\right]/2. \quad (25)
\end{align*}
\]

Next, we bound the third term on the right-hand side of (22). Because \(\mathbf{H}^{k} - \mathbf{R} = h(\mathbf{k}) - \frac{\mathbf{A}_{M}+\mathbf{A}_{m}}{2}\) and because of (9), we have \(\frac{\mathbf{A}_{m} - \mathbf{A}_{M}}{2} \leq \mathbf{H}^{k} - \mathbf{R} \leq \frac{\mathbf{A}_{m} - \mathbf{A}_{M}}{2}\). Let \(c_{0} > 0\). Then, similar to [13, eq. (30)], we obtain
\[
\begin{align*}
\mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}, (\mathbf{H}^{k} - \mathbf{R})(\mathbf{x}^{k+1} - \mathbf{x}^{k})\right)^{2}_{(\mathbf{A}_{m} - \mathbf{A}_{M})/2}\right] \\
\geq -c_{0}\mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}\right)^{2}\right]/2 - \mathbf{E}\left[\left(\mathbf{x}^{k+1} - \mathbf{x}^{k}\right)^{2}_{(\mathbf{A}_{m} - \mathbf{A}_{M})/2}\right]/(8c_{0}). \quad (26)
\end{align*}
\]

Combining (25) and (26) with (22) gives (18).

In addition to Lemma 1, below we provide an upper bound on \(\mathbf{E}[\|\mathbf{z}^{k} - \mathbf{z}^{k}_{Q}\|^{2}]\). From the definitions of \(\mathbf{v}^{k}\) and \(\mathbf{v}^{*}\), we have \(\mathbf{v}^{k}, \mathbf{v}^{*}, \mathbf{v}^{k} - \mathbf{v}^{*} \in \{\mathbf{x} \in \mathbb{R}^{nX}|x_{1} + \cdots + x_{n} = \eta_{s}\}\). Through (10), (11), (12), the AM-GM inequality and the above property, for any \(c_{1}, c_{2} > 0\),
\[
\begin{align*}
\mathbf{E}[\|\mathbf{v}^{k} - \mathbf{v}^{*}\|^{2}] &= \mathbf{E}[\|\left(\mathbf{W}^{k}\right)^{2}\mathbf{W}^{k}(\mathbf{v}^{k} - \mathbf{v}^{*})\|^{2}] \\
&= \mathbf{E}[\|\left(\mathbf{W}^{k}\right)^{2}\left(\mathbf{H}^{k}(\mathbf{x}^{k} - \mathbf{x}^{k+1}) - \beta_{M}\mathbf{W}^{k} - g(\mathbf{x}^{k}) + \nabla f(\mathbf{x}^{k})\right)\|^{2}] \\
&\leq (1 + c_{1})\mathbf{E}[\|\left(\mathbf{W}^{k}\right)^{2}\left(\mathbf{H}^{k}(\mathbf{x}^{k} - \mathbf{x}^{k+1}) - g(\mathbf{x}^{k}) + \nabla f(\mathbf{x}^{k})\right)\|^{2}] \\
+ (1 + 1/c_{1})\mathbf{E}[\|\left(\mathbf{W}^{k}\right)^{2}\left(\beta_{M}\mathbf{W}^{k} + \nabla f(\mathbf{x}^{k}) - \nabla f(\mathbf{x}^{k})\right)\|^{2}] \\
&\leq 2(1 + c_{1})\mathbf{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^{k}\|^{2}_{(\mathbf{A}_{m} + D_{2})/\mathbf{L}_{W}} + 2(1 + c_{1})N\tau \sigma^{2}/\mathbf{L}_{W} \\
+ \beta_{M}^{2}(1 + 1/c_{1})(1 + c_{2})\mathbf{E}[\|\mathbf{x}^{k}\|_{\mathbf{W}}^{2}] \\
+ (1 + 1/c_{1})(1 + 1/c_{2})\mathbf{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^{k}\|^{2}_{(\mathbf{A}_{m} + D_{2})/\mathbf{L}_{W}}/\mathbf{L}_{W}, \\
\end{align*}
\]

leading to
\[
\begin{align*}
\mathbf{E}[\|\mathbf{z}^{k} - \mathbf{z}^{k}_{Q}\|^{2}] &\leq 2(1 + c_{1})\mathbf{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^{k}\|^{2}_{(\mathbf{A}_{m} + D_{2})/\mathbf{L}_{W}} \\
+ 2(1 + c_{1})N\tau \sigma^{2}/\mathbf{L}_{W} + \beta_{M}^{2}(1 + 1/c_{1})(1 + c_{2})\mathbf{E}[\|\mathbf{x}^{k}\|_{\mathbf{W}}^{2}] \\
+ \mathbf{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^{k}\|^{2}_{(\mathbf{A}_{m} + D_{2})/\mathbf{L}_{W}}]. \\
\end{align*}
\]

Pick an arbitrary \(\delta_{s} \in (0, 1)\). By subtracting (27) multiplied by \(\delta_{s}\) from (18), we have
\[
\begin{align*}
(1 - \delta_{s})\mathbf{E}[\|\mathbf{z}^{k} - \mathbf{z}^{k}_{Q}\|^{2}] - \mathbf{E}[\|\mathbf{z}^{k+1} - \mathbf{z}^{k}_{Q}\|^{2}] \\
&\geq \beta_{s}\mathbf{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^{k}\|^{2}_{\mathbf{Q}_{2}}] + 2\delta_{s}\mathbf{E}[\|\mathbf{x}^{k}\|_{\mathbf{W}}^{2}] \\
&\geq \left(2(1 + c_{1})\delta_{s}/\mathbf{L}_{W} + 2N\tau \sigma^{2}, \\
\end{align*}
\]